Statistique appliquée pour une recherche pluridisciplinaire à l'interface entre statistique, informatique et biologie : étude des mécanismes de régulation des gènes

Sophie Lèbre

Statistique appliquée pour une recherche pluridisciplinaire à l'interface entre statistique, informatique et biologie : étude des mécanismes de régulation des gènes.

-

Intégrer des connaissances *métier* à la modélisation statistique

-

Sophie Lèbre

Mémoire pour l'obtention d'une habilitation à diriger des recherches
de l'Université de Montpellier

Ecole Doctorale "Ìnformation, Structure et Systèmes"
Spécialité : Mathématiques appliquées

Soutenance le 28/11/2023.

Anne-Laure Boulesteix	Professeure, LMU Munich	Rapportrice
Andrea Rau	Directrice de recherche, INRAE Jouy-en-Josas	Rapportrice
Nathalie Vialaneix	Directrice de recherche , INRAE Toulouse	Rapportrice
Sarah Djebali	Chargée de recherche, INSERM Toulouse	Examinateur
Pierre Neuvial	Directeur de recherche, CNRS Toulouse	Présidente du jury
Benoite de Saporta	Professeure, Université de Montpellier	

UNIVERSITÉ DE MONTPELLIER

"Je déclare avoir respecté, dans la conception et la rédaction de ce mémoire d'HDR, les valeurs et principes d'intégrité scientifique destinés à garantir le caractère honnête et scientifiquement rigoureux de tout travail de recherche, visés à l'article L.211-2 du Code de la recherche et énoncés par la Charte nationale de déontologie des métiers de la recherche et la Charte d'intégrité scientifique de l'Université de Montpellier. Je m'engage à les promouvoir dans le cadre de mes activités futures d'encadrement de recherche."

Table des matières

Première partie

Activité de recherche

Préambule

Cette partie donne un aperçu du **contexte scientifique** de mes activités de recherche pour l'inférence statistique de réseaux de régulation de gènes (mentionnées dans les sections 2.1.2, 2.2.3, 2.3.3 et 4.3) et d'extraction de caractéristiques de séquence ADN associées à la régulation de l'expression (mentionnées dans les sections 3.2.1, 3.2.2 et 4.1).

Ces travaux se placent dans la continuité de mes travaux de thèses pour l'inférence statistique des réseaux de gènes. Les plus anciens ont été menés en collaboration avec **Dirk Husmeier** (BIOSS, Edinburgh) et **Frank Dondelinger** (BIOSS, Edinburgh). Une grande partie a été réalisée grâce à une collaboration pluridisciplinaire avec **Laurent Bréhélin** (LIRMM, Montpellier) et **Charles Lecellier** (IGMM, LIRMM, Montpellier), et plus récemment avec **Joseph Salmon** (IMAG, Montpellier) ainsi qu'**Antoine Martin** (IPSIM, Montpellier). Ils sont en grande partie le résultat de la qualité et de l'investissement des jeunes docteur.e.s que nous avons eu le plaisir de co-encadrer : **May Taha**, **Raphaël Roméro**, **Florent Bascou** et **Océane Cassan**, ainsi que d'une partie des travaux de thèse de **Christophe Menichelli** et du projet post-doctoral de **Jimmy Vandel**.

Dans un autre domaine d'application, la thèse CIFRE de **Faustine Bousquet**, co-encadrée avec **Christian Lavergne** (IMAG, Montpellier) a permis une contribution (présentée en Annexe C) pour la prédiction du taux de clics à partir de modèles de mélange de données longitudinales et non-gaussiennes.

Chapitre 1

Introduction

L'inférence statistique pour la reconstruction de réseaux de régulation de gènes à grande échelle est un domaine de recherche qui est apparu avec l'acquisition de données quantitatives d'expression, d'abord avec les puces à ADN ou *microarray* (Schena et al., 1995) puis avec le séquençage à haut débit ou *RNA-Seq* (Chu and Corey, 2012). Ces données permettent d'avoir une mesure instantanée du niveau d'expression de l'ensemble des gènes d'un organisme, voire de tous les transcrits avec le RNA-Seq (gènes codants ou non, microARN, ...) et ce, dans une condition spécifique (traitement, type de cellule, ...) et/ou à un instant précis d'un processus biologique. C'est une information globale (à l'échelle du génome entier), qui suscite un vif intérêt en biologie et en statistique (Mochida et al., 2018; Erbe et al., 2022) et requiert une attention particulière en termes de modélisation (Angelin-Bonnet et al., 2018).

La complexité en jeu et le relativement faible nombre de réplicats biologiques (répétitions de la même expérience permettant de mesurer la variabilité) place généralement ce problème dans un contexte de grande dimension. Les réseaux biologiques sont connus pour être modulaires (Segal et al., 2001). De nombreux ensembles de gènes ont des comportements similaires. La sélection de variables dans ce contexte de grande dimension et de forte corrélation entre variables explicatives représente un réel défi statistique (Wasserman and Roeder, 2009; Hastie et al., 2020), en particulier dans un objectif explicatif (Murdoch et al., 2019; Belle and Papantonis, 2021). Bien que des résultats de prédiction soient utilisés pour évaluer la qualité des modèles, il s'agit davantage de comprendre le processus observé et d'identifier l'ensemble des facteurs impliqués, que de prédire le niveau d'expression des gènes. En pratique, l'analyse des mécanismes de régulation requiert une succession d'étapes (normalisation, pré-sélection, ...) à l'issue desquelles il est particulièrement difficile d'évaluer la significativité des résultats (Boulesteix and Hoffmann, 2022).

Par ailleurs, de nombreux niveaux d'observation des mécanismes de régulation des gènes sont accessibles (Djebali et al., 2012). On est capable de mesurer précisément l'ouverture de chromatine (*ATAC-seq*), la fixation de facteurs de transcription (*ChIP-seq*, *DNase footprinting*), les sites d'initiation de la transcription (*CAGE-seq*), ... La qualité et la quantité des données augmente (Mardis, 2017), et leur accès est facilité par le développement de vastes bases de données issues de projets tels que ENCODE, TCGA, FANTOM, GTEx. Aussi, il existe un besoin croissant d'approches multivariées efficaces à même d'intégrer des données hétérogènes et de grande dimension (Rohart et al., 2017; Mariette and Villa-Vialaneix, 2018; Mollandin et al., 2022; Vahabi and Michailidis, 2022).

Ce manuscrit traite essentiellement des questions de modélisation, d'intégration de données hétérogènes et de sélection de variables, dans le contexte de l'étude des mécanismes de régulation des gènes. Le chapitre 2 est dédié aux méthodes d'inférence statistique de réseaux de régulation de gènes à partir de données d'expression uniquement, et avec intégration de données Omics. Le chapitre 3 présente des développements récents pour l'extraction de caractéristiques de séquence ADN associées à la régulation des gènes. Le chapitre 4 aborde la question - essentielle dans un objectif explicatif - de la mesure de l'importance des variables, dans un contexte de grande dimension et de forte corrélation propre aux données Omics. Des perspectives de recherche sont proposées dans le chapitre 5.

Chapitre 2

Inférence de réseaux de régulation de l'expression des gènes

2.1 Modélisation

Cette section présente les grandes catégories de modèles utilisés pour l'inférénce de GRN à partir de données d'expression à grande échelle.

2.1.1 Modèles graphiques

Dépendances univariées (*Relevance network*) Les premières approches ont consisté à identifier les liens 2 à 2 entre les niveaux d'expression des gènes (*relevance network*) sur la base d'une distance telle que la corrélation (`WGCNA` Langfelder and Horvath (2008) ou l'information mutuelle Butte and Kohane (2000) (`ARACNE` (Margolin et al., 2006), `ARACNE-AP` (Lachmann et al., 2016), `CLR` (Faith et al., 2007), `MRNET` (Meyer et al., 2007), `minet` (Meyer et al., 2008)). L'ensemble des liens identifiés définissent les arêtes d'un réseau non-orienté. Ces approches identifient des groupes de gènes co-exprimés, potentiellement impliqués dans un même processus biologique.

Modèles graphiques gaussiens (GGM) Les GGM fournissent un cadre théorique pour modéliser les dépendances multivariées. On considère un vecteur aléatoire auquel on associe un graphe non dirigé, où chaque noeud correspond à une variable du vecteur (ici, le niveau d'expression d'un gène). Une arête n'est pas présente entre deux noeuds si les variables aléatoires correspondantes sont indépendantes conditionnellement aux variables restantes. Dans un cadre gaussien, une indépendance conditionnelle entre deux éléments du vecteur aléatoire correspond exactement à une valeur nulle dans la matrice de concentration (c'est-à-dire l'inverse de la matrice de covariance) (Lauritzen and Lauritzen, 1996; Pearl, 1988; Whittaker, 1990). Ainsi, il s'agit de détecter les éléments non nuls dans la matrice de concentration.

Le passage à l'échelle (considérer plusieurs centaines de gènes simltanément) avec un nombre limité d'observations place cette problématique dans le contexte de la grande dimension "small n, large p". Traditionnellement, dans le cadre des modèles graphiques, un algorithme glouton forward-backward est utilisé pour déterminer les coefficients nuls (Lauritzen and Lauritzen, 1996; Whittaker, 1990) mais cette méthode est computationnellement intractable, même pour un nombre modéré de variables (Banerjee et al., 2008). Cependant, les réseaux de régulation sont parcimonieux (ou *sparse*). On cherche un nombre d'arêtes relativement petit. D'un point de vue statistique, cette information a priori de parcimonie permet une réduction de dimension mise à contribution dans de nombreux développements en statistique. On retrouve notamment des approches reposant sur des heuristiques et tests statistiques (Wille and Bühlmann, 2006; Castelo and Roverato, 2006; Drton and Perlman, 2007, 2008) ; des méthodes d'inférence bayésienne (Dobra et al., 2004; Jones et al., 2005) qui restent limitées en terme de dimension.

L'inférence de GGM à partir d'une régularisation en norme l_1 via le Lasso (Tibshirani, 1996) a été particulièrement explorée. d'Aspremont et al. (2005) utilise l'approche *sparse PCA* (Zou et al.,

2006). Meinshausen and Bühlmann (2006) utilise un simple Lasso pour la régression linéaire et montrent que l'estimateur résultant est consistent, même pour les graphes en grande dimension. Cette approche est ensuite améliorée ensuite par Yuan and Lin (2007) à l'aide d'un critère de type BIC puis Banerjee et al. (2008) en améliorant les algorithmes d'optimisation de deux façons : (i) une descente de gradient par bloc ou (ii) un algorithme de premier ordre (Nesterov, 2005). Friedman et al. (2008) propose le *graphical Lasso*, une version modifiée du lasso dédiée à l'inférence d'une matrice de concentration parcimonieuse. D'autres approches utilisant une régularisation ont été proposées. Schäfer and Strimmer (2005) considèrent une combinaison d'estimateurs (dont des estimateurs pénalisés) implémentée dans le package R `GeneNet`. Ambroise et al. (2009) utilise les caractéristiques modulaire des réseaux de régulation dans le package R `SIMoNe`.

Réseaux bayésiens Parmi les modèles graphiques, les réseaux bayésiens (Pearl, 1988) sont également largement utilisés pour la modélisation de réseaux de régulation. Contrairement aux GGM, les réseaux bayésiens sont définis par un graphe orienté, qui correspond à une factorisation de la densité jointe de l'ensemble des variables. Ce graphe permet également d'identifier des indépendances conditionnelles entre deux variables. Cependant, ce graphe est soumis à la contrainte d'être acyclique (DAG pour *Directed Acyclic Graph*), ce qui complique à la fois la modélisation (les réseaux de régulation contiennent des boucles, notamment des boucles de rétroactions) et l'interprétation des arêtes (il faut construire le graphe moral associé au DAG pour identifier les indépendances conditionnelles) (Lauritzen and Lauritzen, 1996). L'estimation de la structure du graphe a succité quelques dévelopements statistiques (Friedman and Koller, 2003; Grzegorczyk and Husmeier, 2008; Ellis and Wong, 2008)

En pratique, les réseaux bayésiens sont souvent utilisés en présence de données temporelles, via les réseaux bayésiens dynamiques (Friedman et al., 1998; D'haeseleer et al., 1999) . Dès lors, chaque gène est représenté par autant de nœuds qu'il y a de points de temps. La recherche de structure du graphe orienté acyclique (DAG) est alors largement simplifiée car l'information temporelle permet d'orienter les arêtes. Les parents d'un nœud sont recherchés parmi les nœuds des temps précédents.

Un grand nombre de modélisations dynamiques (à temps discret) se ramènent à un DBN (Murphy and Mian, 1999), qu'ils soient définis par des modèles discrets, auto-régressifs, HMM ou non paramétrique (Kim et al., 2003).

Hypothèse non gaussienne (données de comptage) Les développements les plus récents permettent de modéliser la nature discrète des données de comptages acquise par RNA-Seq dans le cadre des GGM (Gallopin et al., 2013; Chiquet et al., 2019) et des Réseaux bayésiens dynamiques (Thorne, 2018).

2.1.2 Modèles dynamiques

Choisir un délai Avec la présence de données temporelles, les approches univariées ont été utilisées pour quantifier les liens entre les niveaux d'expression des gènes observés à 2 points de temps successifs (`Time-delay ARACNE` (Zoppoli et al., 2010), `Time-lagged MRNET` et `Time-lagged CLR` (Lopes et al., 2012). Comme pour les réseaux bayésiens, la difficulté reste à connaître et à être en mesure d'observer le 'bon' délai, qui n'est pas forcément constant et commun à l'ensemble des gènes (Lopes and Bontempi, 2013).

Equations différentielles Afin de modéliser plus finement la dynamique de régulation de l'expression, des équations différentielles ont été utilisées pour modéliser plus précisément la dynamique de régulation de l'expression, incluant notamment la dégradation de l'ARNm (Cao et al., 2012). Les différentes étapes de modélisation (et de génération de données) sont finement décrites par Angelin-Bonnet et al. (2018). Les équations différentielles sont actuellement utilisées dans les approches d'inférence à grande échelle telles que `dynGenie3` (Huynh-Thu and Geurts, 2018) ou `The Inferelator` (Skok Gibbs et al., 2022), décrites plus en détail en section 2.2.

Réseaux non homogènes L'inférence statistique suppose que l'on observe des réplicats d'une même expérience. Aussi, en l'absence de réplicat biologique (pour un même point de temps), les

réseaux dynamiques sont supposés homogènes au cours du temps : une arête est présente si le lien entre les 2 variables est observé tout au long du processus. C'est une hypothèse forte car à partir de données temporelles, on peut s'attendre à une cascade de régulation.

Cette hypothèse a été relâchée de façon à considérer des réseaux non homogènes au cours du temps, notamment par des approches d'inférence de régularisation en norme l_1 pour des graphes de corrélation sur variables discrétisées (`TESLA` (Ahmed and Xing, 2009). Plus récemment, l'approche `TDCor` (Time Delay Correlation) identifie des cascades de régulations en identifiant des corrélations sur une plage de temps délimitée à l'intérieur d'une série temporelle (Lavenus and Lucas, 2022).

Andrieu and Doucet (1999) ont proposé une approche MCMC à sauts réversibles pour la sélection de variables dans le modèle linéaire (avec intégration des coefficients de régression). Nous avons adapté ces résultats pour l'inférence de réseaux bayésiens dynamiques non homogènes (NHDBN)dans une méthode nommée `ARTIVA` (Lèbre et al., 2010). `ARTIVA` permet de reconstruire la structure non homogène de réseau sans information a priori sur les positions de rupture.

Nous avons ensuite proposé une amélioration significative de cette approche par l'introduction de partage d'information entre les structures successives du graphe (Dondelinger et al. (2013) et **disponible en Annexe** E). Le partage d'information est introduit en utilisant une loi a priori pour le choix des arêtes, non pas uniforme comme dans `ARTIVA`, mais qui dépend des arêtes présentes dans la structure du réseau de la plage temporelle précédente. Nous avons utilisé une loi a priori de type exponentiel (un paramètre), puis une loi a priori de type binomial (deux paramètres) qui permet de différencier le degré de partage d'information des arêtes présentes de celui des arêtes absentes. Ceci représente un réel avantage dans le contexte des GRN : les réseaux de régulation étant essentiellement creux, les arêtes absentes sont majoritairement très conservées au cours du temps. Afin d'obtenir un bon mélange de l'algorithme MCMC, nous avons encore amélioré cette méthode en ajoutant un mouvement 'global' qui propose le changement simultané de la structure des réseaux de plusieurs phases successives.

2.1.3 Causalité

Des modélisations ont été proposées pour inférer de la causalité. Parmi les plus connus, le `PC algorithm` (Spirtes et al., 2000) (package R `pcalg`) repose sur l'inférence d'un modèle graphique qui permet d'orienter certaines arêtes, quand les données le permettent. Il a ensuite été étendu au PC-stable algorithm (Colombo and Maathuis, 2014) afin de rendre l'algorithme indépendant de l'ordre des variables.

Dans le cadre des GGM, Opgen-Rhein and Strimmer (2007) (package R `GeneNet`) ont développé une approche pour pour inférer l'orientation des arêtes à partir d'un réseau de corrélations partielles. L'orientation, quand elle est possible, utilise un test statistique qui repose sur le ratio de variance partielle standardisé (*i.e.* la variance partielle d'une variable conditionnellement à l'ensemble des variables divisée par sa variance) entre 2 variables reliée par une arête. Si le ratio est différent de 1, l'arête est orientée dans le sens de la variable qui a le plus de variance partielle standardisée, vers celle qui en a le moins. L'idée est qu'on explique bien la variance de la variable cible (car on connait sa cause). En revanche, la variable parent a une variabilité qui n'est pas expliquée par les autres variables observées (c'est une variable exogène), son ratio de variance partielle standardisé reste proche de 1.

Des données interventionnelles, mesurées suite à des modifications génétiques (*knock out* ou *knock down*) sont particulièrement intéressante pour rétablir des liens de causalité, mais coûteuses. Aussi de nombreux développements ont été proposés pour l'inférence de causalité à partir de données 'observationnelles' (voir Glymour et al. (2019) pour une revue).

Une solution alternative, utilisée depuis les débuts de l'inférence de réseaux pour inférer de la causalité, consiste à intégrer différentes sources de données. Cette problématique, encore très actuelle avec le volume et la diversité des données pour l'observation des méchanismes de régulation, est développée en section 2.3.

2.2 Apprentissage statistique en pratique (Passer à l'échelle)

Parallèlement aux efforts de modélisation, le passage à l'échelle a motivé l'utilisation d'un simple modèle de régression pour l'inférence de réseaux en grande dimension.

2.2.1 Se ramener à p sous-problèmes de régression

Dans le cadre des modèles graphiques gaussiens (GGM), Meinshausen and Bühlmann (2006) exploitent le lien étroit entre les coefficients β_{ij} du modèle de régression linéaire et l'inverse $K = \Sigma^{-1}$ de la matrice de covariance d'un GGM :

$$\textit{Soit } X \sim \mathcal{N}(\mu, \Sigma),\ K = \Sigma^{-1},\ \ \forall 1 \leq i \leq p,\ X_i = \sum_{j \neq i} \beta_{ij} X_j + \varepsilon_i, \quad \textit{alors} \quad \forall i \neq j,\ \beta_{ij} = -\frac{K_{ij}}{K_{ii}}$$

et introduisent le *graphical Lasso*. Un modèle de régression par variable X_i est considéré. La sélection des coefficients non nuls réalisée par le Lasso (régression pénalisée en norme l_1 (Tibshirani, 1996)) permet d'identifier un voisinage de X_i. L'ensemble des arêtes (associées aux corrélations partielles non nulles) résulte alors de l'intersection ou de l'union de l'ensemble des p voisinages (avec les mêmes propriétés asymptotiques).

En pratique, afin de reconstruire des réseaux orientés (non acycliques dans le cadre statique), l'inférence de réseaux est souvent ramenée à p sous-problèmes de régression, sur lesquels repose la sélection de variables. Pour chaque variable sélectionnée, une arête orientée de la variable explicative à la variable cible est inclue au réseau inféré. Quelques exemples de modèles utilisés fréquemment, dans un contexte statique (Eq. (2.1)) ou dynamique (Eq. (2.2, 2.3,2.4)), où toute fonction f peut être considérée (linéaire ou non) sont listés ci-dessous.

1. *Steady State model* :
$$\forall 1 \leq i \leq p, \quad X_i = f(X_{-i}) + \varepsilon_i \tag{2.1}$$
avec X_i le niveau d'expression du gène i, X_{-i} le vecteur des niveaux d'expression des $p-1$ autres gènes et ε_i l'erreur du modèle.
2. *Time-step model* :
$$\forall 1 \leq i \leq p,\ \forall 1 \leq k < n,\ X_i(t_{k+1}) = f\left(X(t_k)\right) + \varepsilon_i \tag{2.2}$$
3. *ODE model* :
$$\forall 1 \leq i \leq p,\ \forall 1 \leq k < n,\ \frac{X_i(t_{k+1}) - X_i(t_k)}{t_{k+1} - t_k} + \alpha_i X_i(t_k) = f\left(X(t_k)\right) + \varepsilon_i \tag{2.3}$$
4. *ODE-log model* :
$$\forall 1 \leq i \leq p,\ \forall 1 \leq k < n,\ \frac{X_i(t_{k+1}) - X_i(t_k)}{log(t_{k+1} - t_k)} + \alpha_i X_i(t_k) = f\left(X(t_k)\right) + \varepsilon_i \tag{2.4}$$
avec $X_i(t_k)$ le niveau d'expression du gène i observé au temps t_k, $X(t_k)$ le vecteur des niveaux d'expression des p gènes, α_i le taux de dégradation de l'ARN messager (ou transcrit) et ε_i l'erreur du modèle.

Pour une équation différentielle, cela requiert (i) de construire la variable réponse à partir des niveaux d'expression du gène i et des temps d'observation t_k et (ii) de choisir les paramètres α_i traduisant les taux de dégradation des ARN. Ce paramètre est parfois commun à tous les gènes (Cirrone et al., 2020). Il est fixé à partir de connaissances biologiques, ou bien estimé sur les données d'expression (`JUMP3` (Huynh-Thu and Sanguinetti, 2015), `dynGENIE3` (Huynh-Thu and Geurts, 2018), `The Inferelator` (Skok Gibbs et al., 2022)). Ces différentes modélisations étendent naturellement les approches de régression au cadre dynamique. Dans tous les cas, il s'agit d'estimer la fonction f pour chaque gène i.

C'est dans le cadre de p sous-problèmes de régression que se place la suite de ce chapitre.

2.2.2 Un aperçu des méthodes d'inférence x logiciels

Les challenges DREAM (Dialogue on Reverse Engineering Assessment and Methods. ont été organisés pour faire face à une difficulté majeure pour l'inférence statistique de réseaux de régulation : on ne dispose que d'une représentation fragmentaire et partiellement correcte des interactions entre les gènes. Ces challenges reposent sur des jeux de données artificielles ou réelles avec éléments de validation.

Les modèles de régression se sont révélé particulièrement performants dans les challenges DREAM4 *Multifactorial Network challenge* et DREAM5 *Network Inference challenge*, et en particulier les méthodes `TIGRESS` (Haury et al., 2012) en régression linéaire et `GENIE3` (Huynh-Thu et al., 2010)) en régression non-linéaire comme cela a été analysé et mis en évidence par Marbach et al. (2012a). Les données collectées et générées pour ce challenge ont été largement ré-utilisées depuis pour améliorer les choix de modélisation et d'inférence de réseaux (*e.g.* Greenfield et al. (2013); Petralia et al. (2015); Huynh-Thu and Geurts (2018); Cirrone et al. (2020)).

Cette section donne un aperçu des grandes catégories de méthodes (et logiciels) pour l'inférence de réseaux de régulation par régression.

Modèles linéaires régularisés Les approches parmi les plus performantes dans le challenge DREAM5 (en particulier sur les données artificielles) reposent sur une régularisation en norme l_1 ou Lasso pour *Least Absolute Shrinkage and Selection operator* proposée par Tibshirani (1996). Grâce à la géométrie de la norme l_1, le Lasso réalise simultanément l'estimation des paramètres et la sélection de variables. Cela permet d'obtenir rapidement une prédiction. En revanche, le calcul de l'importance des variables nécessite le recours au sous-échantillonnage (Bach, 2008; Meinshausen and Bühlmann, 2010), en particulier dans un contexte de corrélations élevées (voir section 4.1).

Parmi les packages les plus cités, Haury et al. (2012) ont introduit l'approche nommée `TIGRESS` pour *Trustful Inference of Gene REgulation using Stability Selection* qui utilise l'algorithme LARS (*Least Angle Regression* (Efron et al., 2004)) suivie d'une étape de stabilité de sélection. L'algorithme LARS est proche de la sélection pas à pas et du Lasso. A chaque étape de l'algorithme, une nouvelle variable est ajoutée (celle dont la corrélation avec le résidu est la plus forte). Contrairement à une sélection pas à pas classique, LARS ne ré-estime pas entièrement le modèle à l'inclusion d'une nouvelle variable, mais l'affine partiellement. Greenfield et al. (2013) utilise l'*elasticNet* introduit par (Zou and Hastie, 2005) qui repose sur une combinaison des régularisations en norme l_1 et l_2, qui est ensuite intégrée à la suite `The Inferelator` (Skok Gibbs et al., 2022). Guo et al. (2016) proposent `PLSNET`, un algorithme de sélection de variables à partir de régression PLS et de sous-échantillonage.

Forêts aléatoires L'utilisation des forêts aléatoires (RF pour Random Forests) (Breiman, 2001) pour l'inférence de réseaux de régulation a été introduite par Huynh-Thu et al. (2013) avec l'approche nommée `GENIE3` pour *GEne Network Inference with Ensemble of trees*. Les RF reposent sur une agrégation d'arbres de régression. Un arbre de régression ou de classification (`CART` pour *Classification And Regression Tree* (Breiman et al., 1984)) est caractérisé par une succession de règles de décision (les nœuds de l'arbre) portant sur les valeurs des variables explicatives. Dans un cadre de régression, les prédictions sont données par la moyenne des observations classées dans chaque nœud terminal de l'arbre (feuille). Pour plus de robustesse, les forêts aléatoires considèrent une prédiction moyenne obtenue à partir d'une agrégation d'arbres, chacun arbre étant estimé sur une modification du jeu de données (sous-échantillon bootstrap et tirage aléatoire d'un sous-ensemble de prédicteurs).

Plus récemment, le modèle a été adapté à la modélisation dynamique ((Maduranga et al., 2013), `JUMP3` (Huynh-Thu and Sanguinetti, 2015), `dynGENIE3` (Huynh-Thu and Geurts, 2018)), à l'intégration de connaissance a priori sur la présence d'arêtes dans le réseau (`iRafNet` (Petralia et al., 2015)), et à la combinaison des deux (`OutPredict` (Cirrone et al., 2020)). L'estimation des forêts aléatoires a également été optimisée par un algorithme stochastic gradient boosting (Friedman, 2002), avec l'approche `GRNBOOST` (Aibar et al., 2017; Moerman et al., 2019) de façon à analyser des données cellule unique utilisée dans `SCENIC` (Aibar et al., 2017; González-Blas et al., 2022).

Scores bayésiens avec a priori parcimonieux Le cadre bayésien permet naturellement la sélection de variables au moyen d'une loi a priori en faveur d'un petit nombre de prédicteurs (Andrieu and Doucet, 1999). Certaines approches pour l'inférence de réseaux utilisent un score bayésien, par exemple avec un a priori parcimonieux (Zellner, 1983) dans l'approche `BBSR` (*Bayesian Best Subset Regression* (Greenfield et al., 2013)) ou bien une loi a priori de la forme d'une sigmoïde à valeur entre 0 et 1 (fonction logit) dans l'approche `MERLIN` (Roy et al., 2013). La sélection de modèle se fait par un algorithme glouton, avec un filtre en amont lorsque le nombre de prédicteurs potentiels est grand. `MERLIN` propose également de considérer un a priori modulaire du réseau (section 2.2.4).

Les formes de lois a priori utilisées dans ces deux approches permettent également l'intégration de données et sont utilisées dans des développements plus récents tel que `MERLIN+Prior` (Siahpirani and Roy, 2017) ou `The Inferelator` (Skok Gibbs et al., 2022). Ces deux approches sont présentées en section 2.3.3.

Interactions entre prédicteurs Les arbres de régression (et donc les forêts aléatoires) permettent non seulement une modélisation non linéaire, mais aussi reposent par construction sur des interactions entre prédicteurs (ce que ne fait pas le modèle linéaire par défaut). Bien sûr, dans un cadre linéaire on peut ajouter l'ensemble des interactions 2 à 2 à la liste des prédicteurs, et éventuellement considérer des opérateurs logiques AND, OR ou XOR entre prédicteurs comme dans la première version de `The Inferelator` (Bonneau et al., 2006), mais cela accroit fortement le problème de dimension. Dans le contexte de l'inférence de réseaux, le nombre d'observations disponibles limite dans tous les cas l'ordre des interactions que l'on peut raisonnablement identifier. De nombreux développements en statistique et optimisation ont été proposés pour la sélection d'interactions dans le modèle linéaire, et sont discutés dans la section 3.2.1.

Single cell Avec le développement des données cellule unique (*Single cell*), la reconstruction de réseau est plus récemment considérée à l'échelle de la cellule (Nguyen et al., 2021), en particulier les approches `SCENIC` (Aibar et al., 2017), `SCENIC+` (González-Blas et al., 2022) et `CellOracle` (Kamimoto et al., 2023).

2.2.3 Réduction de dimension, en amont de l'estimation

Quelle que soit la méthode choisie, l'apprentissage et/ou la modélisation des réseaux de régulation sont grandement facilités en réduisant la dimension de l'espace de recherche en amont de l'inférence. En pratique, de nombreux moyens sont utilisés à partir des données d'expression et de la connaissance du processus étudié (espèce, conditions, voies de régulation, présence de motifs, ...). Un aperçu des pratiques courantes est donné ci-dessous.

1. **(Données) Sélection des conditions expérimentales d'intérêt** Un même gène peut être régulé par différents mécanismes/TF sous différentes conditions. Aussi, pour la reconstruction d'un seul réseau, il est important de sélectionner un ensemble de conditions au sein desquelles les mêmes relations de régulations sont en jeu. En revanche, cela diminue le nombre de conditions exploitables pour l'inférence statistique. Si l'on souhaite conserver des conditions relativement hétérogènes, une option consiste à relâcher l'hypothèse d'homogénéité pour considérer une inférence conjointe avec partage d'information entre différentes plages temporelles (voir section 2.1.2) ou entre différents jeux de données d'expression (voir section 2.2.4).
2. **(Données) Sélectionner les gènes différentiellement exprimés.** Beaucoup de gènes ont un profil plat. dans les conditions d'intérêt. Ils sont généralement retirés de l'analyse. Différents critères sont utilisés en pratique (Expression différentielle (`DIANE`, Cassan et al. (2021), seuil de variance (Miraldi et al., 2019) , ...). En plus de limiter le nombre de régulateurs, cela réduit également le nombre de cibles et donc de modèles de régression à estimer. Dans le cas d'un grand nombre de conditions expérimentales disponibles, la méthode `cMonkey` (Reiss et al., 2006) également disponible dans la suite `The Inferelator` (Skok Gibbs et al., 2022) propose une approche par bi-clustering pour identifier simultané-

ment des sous-ensembles de gènes et de conditions dans lesquelles les gènes sont co-régulés (Tâches 1. et 2.).

3. **(Connaissances) Pré-sélectionner les gènes qui seront considérés comme régulateurs potentiels** C'est ici que se situe le gain le plus important. Utiliser la connaissance du rôle fonctionnel des gènes permet de réduire drastiquement l'espace de recherche des prédicteurs. Lorsque l'espèce a été largement étudiée (plante *A. thaliana*, Souris, Homme), la plupart des gènes codant pour un facteur de transcription sont déjà connus. On peut alors chercher les régulateurs parmi ce sous ensemble de gènes. Cela permet une réduction drastique de l'espace de recherche. Éventuellement, la connaissance de voies metaboliques (`KEGG`,Kanehisa et al. (2016)) ou la présence de motifs de fixation de TF dans la région promotrice du gène cible permet d'affiner encore la pré-sélection. Pour les espèces dont les TF ne sont pas connus, utiliser les annotations telles que les termes GO (The Gene Ontology Consortium, 2008) pour identifier les gènes annotés comme *DNA binding* ou *Transcription Factor*, ou bien une identification par orthologie à partir d'annotations pour une espèce proche.
4. **(Données) Grouper les prédicteurs corrélés** Pour améliorer l'apprentissage statistique, il est souhaitable de regrouper les gènes (parmi les régulateurs potentiels) dont les profils d'expression sont très similaires. quelle que soit la distance choisie (corrélation, information mutuelle, ...). Le modèle linéaire par exemple repose sur l'hypothèse que la matrice des observations des variables explicatives est de plein rang. Bien que des méthodes de régularisation soient utilisées (Lasso (Tibshirani, 2013), Elastic Net (Zou and Hastie, 2005)), il est préférable pour l'inférence et pour l'interprétation de regrouper les profils presque indistinguables. Les distances 2 à 2 ne permettent pas d'identifier directement des groupes de gènes. Des approches de classification non supervisée sont utilisées pour identifier ces profils très fortement similaires.

Ce protocole rassemble des pratiques couramment citées mais rarement disponibles dans une seule et même application, mettant à disposition des méthodes statistiques avancées pour l'inférence de réseaux. Afin de le rendre facilement accessible, et dans un objectif de reproductibilité, nous avons mis à disposition une application en ligne interactive nommée `DIANE` (Cassan et al. (2021) et **disponible en Annexe** B) pour l'inférence de réseaux à partir des données brutes. `DIANE` permet de reproduire ce protocole de réduction de dimension en amont de l'inférence de réseaux : exploration des données par ACP (package R `ade4`), analyse différentielle de l'expression (package R `EdgeR` (McCarthy et al., 2012)), pré-sélection des régulateurs annotés comme TF pour de nombreuses espèces, regroupement des régulateurs les plus corrélés par la méthode de Louvain (package R `igraph`). L'application inclut également des outils pour la normalisation des données (TMM (Robinson et al., 2010) ou DESeq2 (Love et al., 2014)) et pour le clustering de profils d'expression (package R `coseq` (Rau and Maugis-Rabusseau, 2018)).

L'inférence de réseaux est proposée à partir du package R `GENIE3` (Huynh-Thu et al., 2010). Nous avons ajouté la possibilité de choisir le critère d'importance MDA au lieu du critère MDI qui proposé par défaut (voir section 4.2, dédiée à la mesure de l'importance des variables pour les agrégations d'arbre) et d'associer une p-valeur à chaque arête selon un mode d'estimation non paramétrique (voir la section 4.3, dédiée à la sélection de variables dans le cadre de l'inférence de réseaux).

Au sein du réseau estimé, l'identification de modules de gènes avec des similarités de connexion est proposée à partir du modèle SBM (*Stochastic Block Model*, package R `sbm` (Leger, 2016)). Des enrichissements en termes GO au sein d'un ensemble de gènes peuvent être calculés (package R `clusterProfiler` (Yu et al., 2012)).

Depuis sa publication en 2021, l'application `DIANE` a été utilisée notamment dans 2 contributions pour l'inférence de réseaux, dans les racines du lupin blanc (Le Thanh et al., 2021) et le champignon Trichoderma reesei (Beier et al., 2022).

Ces développements nous ont également permis de reconstruire un réseau de régulation de la réponse à l'élévation du CO_2 dans les racines de la plante *A. thaliana* (Cassan et al., 2023) à partir d'un plan d'expérience impliquant la combinaison de deux de effets, apport en nitrate et niveau de CO_2. Le rôle dans la stimulation de la croissance par l'élévation du CO_2 a été validé

expérimentalement pour 3 TF (MYB15, WOX11, EDF3) parmi ceux de fort degré dans ce réseau.

2.2.4 Améliorer l'apprentissage statistique

Des stratégies ont été développées pour améliorer l'apprentissage statistique dans le contexte d'inférence de GRN à partir de données d'expression, notamment à partir des connaissances relatives à la topologie des réseaux, pour combiner plusieurs jeux de données d'expression ou méthodes d'inférence.

Connaissance de la topologie du réseau Les réseaux biologiques sont connus pour avoir une structure modulaire. Cette propriété a été utilisée pour l'identification de groupes de gènes co-régulés, en particulier à partir d'ensembles de données complémentaires à l'expression, telles que la présence de motifs de fixation de TF, de modifications génétiques ou d'autres types de données expérimentales (voir section 2.3.3). Segal et al. (2003) ont étendu ces approches pour identifier des groupes de gènes co-régulés par les mêmes TF à partir de l'expression uniquement. Dès sa création, la suite `The Inferelator` (Bonneau et al., 2006) regroupe les gènes corrélés avec l'approche pour le bi-clustering (gènes x conditions) implémentée dans `cMonkey` (Reiss et al., 2006). Cet objectif est également présent dans d'autres approches pour l'optimisation d'un score bayésien (Roy et al., 2013), en plus de l'intégration d'une connaissance a priori du réseau issus de données hétérogènes (Siahpirani and Roy (2017), section 2.3.3).

Mukherjee and Speed (2008) ont proposé un modèle de réseau bayésien exploitant des lois a priori décrivant la parcimonie et la distribution des arêtes dans le réseau dans un cadre d'apprentissage bayésien (MCMC). Dans le cadre des GGM, Ambroise et al. (2009) propose un algorithme *EM-like* permettant de reconstruire la topologie du réseau exploitant un modèle de mélange (Daudin et al., 2008) sur les nœuds, chaque classe étant définie par sa connectivité dans le réseau. Ce modèle est ensuite étendu par Charbonnier et al. (2010) avec une approche *weighted Lasso* permettant d'intégrer une connaissance a priori du réseau (voir section 2.3.3), ou d'inférer un réseau avec une pénalité Lasso non uniforme sur les arêtes favorisant une structure modulaire à partir de données d'expression uniquement.

Intégration de données homogènes (partage d'information) Dans le cadre des GGM, Chiquet et al. (2011) propose l'inférence jointe de plusieurs réseaux à partir de jeux de données différents à partir d'une pénalité lasso favorisant les arêtes présentes dans plusieurs réseaux.

Nous avons proposé un partage d'information dans un cadre bayésien, pour l'inférence conjointe de réseaux bayésiens dynamiques isuus de jeux de données différents (Dondelinger et al., 2012) ou entre les plages de temps successives d'un réseau bayésien dynamique non homogène au cours du temps (Dondelinger et al. (2013) et **disponible en Annexe** E).

La pénalité *group Lasso* (Yuan and Lin, 2006) a également été utilisée pour le partage d'information entre plusieurs jeux de données (Liu et al., 2014). Cette pénalité incite à sélectionner les mêmes variables dans les différents jeux de données. Omranian et al. (2016) utilise une pénalité *fused Lasso* permettant de favoriser la similarité 2 à 2 entre les jeux de données successifs, initialement ordonnés selon la similarité des données.

Steady State + Time series Avec `dynGENIE3`, Huynh-Thu and Geurts (2018) obtiennent de meilleurs résultats sur les données DREAM4 en intégrant des données stationnaires et des données temporelles dans un même modèle de régression, défini par une équation différentielle (Eq. (2.3)). Pour les données non temporelles, le terme $\frac{X_i(t_{k+1})-X_i(t_k)}{t_{k+1}-t_k}$ est considéré égal à 0 (approchant $\frac{\partial X_i(t)}{\partial t}$, il est considéré nul dans un état stationnaire). Le taux de dégradation α_i et la fonction de lien f sont communs aux 2 types de données. Cette approche a aussi été utilisée ensuite avec les méthodes `Outpredict` (Cirrone et al., 2020) et `The Inferelator` (Skok Gibbs et al., 2022).

Wisdom of crowds L'intégration de plusieurs méthodes d'inférence (à partir d'un même jeu de données) s'est également révélé très utile, en particulier dans l'analyse des résultats du challenge *DREAM5* réalisée par Marbach et al. (2012a). Les approches reposant sur l'ensemble des prédictions, intégrée en moyennant les rangs des arêtes proposées par les différentes méthodes surpassent

clairement les approches individuelles. D'autres approches ont été proposées pour combiner les prédictions de différentes méthodes d'inférence de réseaux (Ruyssinck et al., 2014; Schiffthaler et al., 2021).

2.3 Intégration de données hétérogènes

L'analyse statistique des données Omics évolue progressivement vers l'intégration de données provenant de différentes plateformes (Noor et al., 2019).

2.3.1 Motivations et aperçu des types de données

Avec le développement du séquençage à haut débit, de nombreux niveaux d'observation des mécanismes de régulation des gènes sont accessibles à l'échelle du génome (Djebali et al., 2012). On est capable de mesurer précisément l'ouverture de chromatine (*ATAC-seq*), la fixation de facteurs de transcription (*ChIP-seq*, *DNase footprinting*), les sites d'initiation de la transcription (*CAGE-seq*), ...

Motifs de fixation (matrices PWM) Les facteurs de transcription (protéines) se lient préférentiellement à des séquences d'ADN spécifiques, qui sont résumées dans des modèles statistiques connus sous le nom de 'matrices de poids de position' ou *PWM* pour *Position Weight Matrix* (Wasserman and Sandelin, 2004; Stormo, 2013). Le plus souvent, les matrices PWM sont obtenues à partir du log ratio des probabilités d'occurence de chacune des 4 lettres A, C, G, T pour chaque position (rassemblées dans une matrice *PPM* pour *Position Probability Matrix*). Ces matrices PWM sont disponibles pour de nombreux TF dans des bases de données telles que JASPAR (Sandelin et al., 2004; Castro-Mondragon et al., 2022) et HOCOMOCO (Kulakovskiy et al., 2016), et permettent de calculer un score pour chaque position d'une séquence ADN. Un score élevé indique un potentiel site génomique de fixation. Cependant, dans une condition spécifique, les TF ne s'associent qu'à un petit sous-ensemble de leurs sites génomiques potentiels Il y a beaucoup de faux positifs. En effet, d'une part, le nombre de familles de domaine de fixation des protéines à l'ADN est faible par rapport au nombre de TF. Plusieurs TF ont souvent des motifs de fixation très similaires, bien qu'ils présentent généralement des sites de fixation distincts *in vivo* (Jolma et al., 2013; Shen et al., 2018). De plus, une part importante des sites de fixation observés *in vivo* ne correspond pas au fixation de liaison connu du TF (Wang et al., 2012; Jolma et al., 2013), potentiellement en raison de fixation indirecte, via une autre protéine. Des données d'interaction de protéine peuvent être utiles pour étendre la liste des les TF susceptibles de se fixer. Par ailleurs, un grand nombre de paramètres influencent la fixation des TF, et ces paramètres varient entre les types de cellules, de tissus ou les conditions expérimentales (Slattery et al., 2014; Srivastava and Mahony, 2020).

Accessibilité de la chromatine Le degré de condensation de l'ADN (autour des nucléosomes d'histones et des protéines non histones) est variable le long des chromosomes, dans la chromatine. selon les conditions et est également associé à la transcription. Lorsque le degré de condensation est faible, la chromatine est 'ouverte' et accessible à la machinerie de transcription. Des analyses de la fixation des TFs ont montré que la plupart des sites de fixation sont situés dans des régions ouverte de la chromatine (Ernst and Kellis, 2013). On peut noter qu'il n'est pas évident d'établir si l'état de la chromatine est une cause ou une conséquence de la fixation de TF (Huminiecki and Horbańczuk, 2020).

Variations génétiques (SNP, eQTL) Les variations génétiques, notamment les sites de polymorphismes mono-nucléotidiques au sein du génome (*SNP* pour *single-nucleotide polymorphisms*), peuvent également réguler l'expression des gènes. Des SNP sont identifiés pour être associés à l'expression (*eQTL* pour *expression Quantitative Trait Loci*) : la variabilité génétique des ces sites génomiques est associée à la variation de l'abondance d'un transcrit particulier (Schadt et al., 2003; Gilad et al., 2008), principalement en raison de leur impact sur la fixation des TF (Gaffney, 2013; Albert and Kruglyak, 2015). Plus généralement, les études de génomique génétique (Gaffney,

2013) analysent la manière dont les polymorphismes conduisent à la variation des caractéristiques moléculaires, telles que les profils d'ARNm, de protéines ou de métabolites.

Ainsi la diversité des données mesurées à l'échelle du génome augmente, et leur accès est facilité par le développement de vastes bases de données issues de projets tels que ENCODE, TCGA, FANTOM, GTEx. Les données d'expression ne donne qu'une vision partielle des mécanismes de régulation et les relations de causalité (section 2.1.3) restent difficiles à établir dans ce contexte de grande dimension, et de forte corrélation. Des modèles définis par des réseaux très distincts peuvent représenter les données de façon comparable. Aussi, l'intégration de données pour l'inférence de réseaux de régulation est largement considérée (Wani and Raza, 2019), mettant à profit une source de données complémentaire à l'expression pour réduire l'espace de recherche et/ou se rapprocher de la causalité.

2.3.2 Différentes modélisations

Combinaison de réseaux issus de différents types de données Marbach et al. (2012b) exploitent des réseaux inférés indépendamment, à partir de type de données complémentaires et de nature différentes (présence de motifs conservés, fixation de TF, modification de la chromatine) en plus du niveau d'expression des gènes, chez la drosophile. Les poids associés aux arêtes de chacun de ces réseaux sont utilisés comme prédicteurs dans un modèle de régression logistique (prédire les arêtes présentes) à partir d'un jeu de données de validation avec un niveau de confiance élevé. De Clercq et al. (2021) reprennent cette démarche et comparent différents classifieurs (forêts aléatoires, régression logistique et *gradient boosting* pour des arbres de décision) chez la plante *Arabidopsis thaliana*. Les meilleures performances sont obtenues avec le *gradient boosting*.

Inférence de l'activité des facteurs de transcription Etant donné que le niveau d'expression d'un TF ne représente pas nécessairement son activité (Schacht et al., 2014), Arrieta-Ortiz et al. (2015) proposent d'inférer l'activité des facteurs de transcription à partir des niveaux d'expression des gènes cibles et d'une matrice indiquant les relations de régulation connues a priori. C'est la matrice inférée traduisant l'activité des TFs qui est ensuite utilisée comme prédicteur pour l'inférence de réseau. Cette approche a été intégré à la suite `The Inferelator` (Skok Gibbs et al., 2022).

Variables explicatives de différentes natures Certains modèles de réseaux considèrent des nœuds de différentes natures. Cai et al. (2013) intègrent des données de variants associés à l'expression de certains gènes ($eQTL$), et utilisent cette information supplémentaire pour orienter les arêtes en terme de causalité. Kim et al. (2014) intègrent des données de méthylation de l'ADN et des données de variation du nombre de copies d'un gènes (CNV). Cette approche est reprise par Zarayeneh et al. (2017) en n'incluant que les gènes dans le réseau, mais la structure du réseau est estimées conditionnellement à des données de CNV et de méthylation.

L'intégration de variables explicatives de différentes natures (méthylation, CNV, fixation de TF, expression de micro ARN) est également utilisée pour expliquer le niveau d'expression des gènes, par exemple dans les approches `RACER` (Li et al., 2014) et `TEPIC` (Schmidt et al., 2017) qui sont comparées aux modèles utilisant uniquement des données de séquence ADN dans la section 3.2.1. L'approche `RACER` utilise également l'inférence de l'activité des facteurs de transcription. La méthode d'inférence repose sur 2 modèles de régression successifs (le premier pour estimer l'activité des TF, le deuxième pour inférer le réseau).

Pré- ou post- sélection des arêtes D'un point de vue méthodologique, les approches les plus simples reposent sur une étape de pré-sélection des régulations possibles, en amont de l'inférence. On peut limiter les arêtes entrantes pour un gène cible donné aux TF dont le motif est situé dans la région promotrice du gène. Chaque arête inférée est alors soutenue par la présence de corrélation partielle entre l'expression du TF et du gène cible et la présence d'un motif de fixation. On a alors grande confiance en les arêtes inférées, mais cela limite le réseau aux facteurs de transcription dont le ou les motifs de fixation sont connus. La même chose est possible à partir de données

expérimentales de fixation de protéines (ChIP-chip, puis ChIP-seq) (Youn et al., 2010). Cela est particulièrement intéressant si ces données sont disponibles pour un grand nombre de facteurs de transcription, et dans les conditions étudiées. Ces approches reposant sur un filtre via des données hétérogènes sont utilisées dans des approches récentes pour l'inférence de réseau à partir de données cellule unique. Par exemple, CellOracle (Kamimoto et al., 2020) utilisent la présence de motifs de fixation dans les régions de chromatine ouvertes (ATAC-seq) pour pré-sélectionner les arêtes potentielles.

A l'inverse, d'autres approches utilisent la présence de motifs (`SCENIC` (Aibar et al., 2017)) ou des données de fixation de TFs (`GRACE`, Banf and Rhee (2017)) afin d'élaguer un réseau inféré en premier lieu uniquement à partir de données d'expression (en cellule unique). L'approche est étendue (`SCENIC+`, González-Blas et al. (2022)) pour intégrer de données cellule unique mesurant l'ouverture de chromatine (*single-cell ATAC-seq*), afin d'identifier des liens entre les triplets (accessibilité du site de fixation du TF, expression du TF et expression du gène cible) afin de proposer des associations entre régions de régulation distantes (*enhancer*) et gènes, spécifique d'un type cellulaire.

Plutôt que d'utiliser cette information partielle du réseau en amont ou en aval de l'inférence (à partir de données d'expression), les méthodes présentées dans la section suivante permettent d'utiliser ces différents types de données conjointement pour résoudre un problème d'optimisation sous contraintes.

2.3.3 Intégrer un a priori pour guider l'inférence

Les premières approches pour l'intégration de données hétérogènes ont été proposées dans un cadre de classification non supervisée. Afin de reconstruire des réseaux modulaires, des groupes sont identifiés sur la base de similitudes entre l'expression des gènes et la présence de motifs de fixation de TF connus dans les promoteurs, des annotations fonctionnelle des gènes ou des données de fixation (D'haeseleer et al., 2000; Segal et al., 2001; Bar-Joseph et al., 2003).

Dans un contexte de régression, une matrice de connaissance a priori du réseau construite à partir de données OMICS (même partielle et imparfaite), peut servir de guide à l'inférence statistique. Ceci a été considéré dans différents cadres statistiques.

Inférence bayésienne Utiliser une information a priori est le propre de l'inférence bayésienne, et c'est dans ce cadre que naturellement l'utilisation de connaissances a priori a été utilisée pour l'inférence de réseaux, à partir de données de fixation de facteurs de transcription (Hartemink et al., 2002; Bernard and Hartemink, 2005; Werhli and Husmeier, 2007), de la séquence promotrice (Tamada et al., 2003) de données génotypiques (Zhu et al., 2004, 2007), de données d'interactions entre protéines (Imoto et al., 2003; Nariai et al., 2004), ou d'une combinaison de plusieurs types de données (Zhu et al., 2008).

D'un point de vue méthodologique, Hartemink et al. (2002) ont proposé d'incorporer des données de localisation génomique pour guider l'inférence du modèle. Tamada et al. (2003) ont développé une méthode qui détecte itérativement les motifs de consensus sur la base de la structure du modèle de réseau estimé, puis évalue le réseau à l'aide du résultat de la détection des motifs, jusqu'à ce que le réseau inféré devienne stable. Imoto et al. (2003) ont introduit un cadre utilisant la distribution de Gibbs où une fonction d'énergie a été utilisée pour évaluer la probabilité d'une arête dans les réseaux inférés. Cette approche a ensuite été étendue pour intégrer de multiples sources de connaissances préalables dans l'apprentissage dynamique des réseaux bayésiens (DBN) via l'échantillonnage MCMC (Bernard and Hartemink, 2005; Werhli and Husmeier, 2007). Cependant ces approches restent limitées en termes d'échelle (quelques dizaines de TFs).

Critères bayésiens Plus récemment, afin de passer à l'échelle, des méthodes d'optimisation d'un critère bayésien intégrant des connaissances a priori ont été proposées. La contrainte de parcimonie modélisée par un score bayésien (section 2.2.2) est étendue de façon à exploiter une connaissance a priori de la topologie du réseau. Greenfield et al. (2013) utilisent une approche exhaustive (*Bayesian Best Subset Regression*) après considération des 2^k modèles de régression possibles à partir de k variables explicatives (avec un filtre en amont pour réduire k à 10). L'algorithme consiste à minimiser l'espérance du critère BIC pour un modèle de régression bayésien avec un

a priori sur le vecteur des coefficients qui repose sur une modification du g-prior (Zellner, 1983), c'est-à-dire une distribution similaire à une loi normale multivariée d'espérance un vecteur donné β^0 et de matrice de variance-covariance liée aux données via un scalaire positif. Le vecteur β^0 est choisi nul de façon à satisfaire les attentes de parcimonie. Le lien aux données dépend d'un vecteur de dimension égale au nombre de prédicteurs, et dont les éléments prennent la valeur g si l'arête est soutenue par les connaissances a priori, $\frac{1}{g}$ sinon. Cette approche a été ajoutée à la suite *The Inferelator* (Skok Gibbs et al., 2022). Siahpirani and Roy (2017) utilisent un algorithme glouton pour optimiser un score défini par la vraisemblance contrainte par une loi a priori de type logistique. C'est une extention d'un premier modèle prenant en compte de la structure modulaire (Roy et al., 2013) (voir section 2.2.4).

Régularisation l1 (*weighted Lasso*) Des approches pour l'intégration de données à partir d'un vecteur de poids reposant sur des connaissances a priori ont également été proposées via l'optimisation de modèles linéaires régularisés (*weighted Lasso*) ou de forêts aléatoires (*weighted RF*). Dans les 2 cas, ce vecteur de poids guide l'inférence statistique.

La régularisation en norme l_1 offre elle aussi un cadre naturel pour l'intégration de données via un poids de pénalité spécifique w_k pour chaque coefficient β_k (Eq. (2.5)). Les coefficients β_k associés aux prédicteurs identifiés comme régulateurs potentiels par une source de données complémentaire sont moins pénalisés que les autres, via poids w_k plus faible.

$$\hat{\beta}_{weighted\ Lasso} = argmin_\beta \quad ||y - \sum_{k=1}^{K} x_k\beta_k||^2 + \sum_{k=1}^{K} w_k|\beta_k| \tag{2.5}$$

Cette approche a été considérée pour l'inférence de GRN par Yong et al. (2008) avec des poids choisis dans l'intervalle $[0, +\infty[$, définis à partir de recherches bibliographiques ou de la présence de motifs de fixation de TFs dans les régions promotrices. Bergersen et al. (2011) proposent une analyse détaillée du *weighted Lasso* pour l'intégration de données, et considèrent différente façons de définir le vecteur des poids de pénalité w_k en fonction des connaissances a priori.

Greenfield et al. (2013) l'utilisent avec l'*elastic Net* (Zou and Hastie, 2005; Zou and Zhang, 2009) afin de combiner les avantages de la régularisation en norme l_2 (plus stable que le Lasso) et un vecteur de poids de pénalité en norme l_1 spécifique à chaque TF :

$$\hat{\beta}_{weighted\ eNet} = argmin_\beta \quad ||y - \sum_{k=1}^{K} x_k\beta_k||^2 + \lambda_2||\beta||_2^2 + \sum_{k=1}^{K} w_k|\beta_k| \tag{2.6}$$

L'intensité de l'a priori est fixée en fonction de performances évaluée sur des données de validation biologiques (solidement construites, mais partielles et imparfaites).

L'intégration de données via une régularisation en norme l_1 a été proposée également dans le cadre de l'inférence de GGM pour la reconstruction de GRN avec le *weighted graphical Lasso* (Li and Jackson, 2015; Zuo et al., 2017). Plus récemment, Lingjærde et al. (2021) proposent de définir les poids à partir d'une fonction de lien logistique dépendant d'un paramètre k qui permet de contrôler la force de l'a priori (écart entre les arêtes soutenues par le prior et les autres) et introduisent le *tailored graphical Lasso*. La valeur de k est choisie pour l'ensemble des gènes du réseau en optimisant le critère BIC étendu aux GGM ($eBIC$, Foygel and Drton (2010)), après avoir estimé un poids de régularisation global qui garantit la stabilité du modèle via l'approche `StARS` pour *Stability Approach to Regularization Selection* (Liu et al., 2010) (voir section 4.1).

L'intégration de données avec le Lasso pour l'inférence de réseaux a également été proposée de façon différente, en utilisant l'a priori initialiser l'algorithme itératif `ISTA` pour *Iterative Soft Thresholding Algorithm* (Qin et al., 2014).

Forêts aléatoires (*weighted RF*) Les forêts aléatoires permettent également d'intégrer une information a priori au moment de la pré-sélection aléatoire des variables, qui est effectuée à la création de chaque nouveau nœud d'un arbre (contrairement aux *gradient-boosted trees* (Friedman, 2001) pour lesquels l'ensemble des variables est considéré à chaque nœud). Au lieu de considérer une loi uniforme (par défaut dans les forêts aléatoires), les *weighted RF* consistent à fixer un poids

spécifique pour chaque variable explicative, à partir d'une information a priori venant appuyer ou non la sélection de cette variable. Petralia et al. (2015) ont exploité cette modélisation pour l'inférence de réseau dans une approche nommée `IRafNet`, et ont considéré l'intégration de différents types de données telles que des capacités de fixation à l'ADN ou d'interactions entre protéines. Plus tard, Cirrone et al. (2020) ont étendu cette approche d'intégration de données à une modélisation dynamique comme proposé par la méthode `dynGENIE3` (Huynh-Thu and Geurts, 2018). Cette approche nommée `OutPredict` inclue deux types de modélisation dynamique : *Time-step* (Eq. (2.2)) et *ODE* (Eq.(2.3)), présentées en section 2.2.1.

Optimisation de l'intensité d'intégration des connaissances a priori La connaissance a priori d'une relation de régulation entre un TF et un gène cible issue de données complémentaires à l'expression augmente la confiance que l'on peut avoir dans la sélection d'une arête. En revanche, quel que soit le type de données complémentaires, les faux positifs sont nombreux : la présence d'un motif ne signifie pas qu'il est fonctionnel, ni qu'il est fixé dans la condition étudiée ; la fixation d'un TF dans les conditions expérimentales requises pour l'analyse n'implique pas qu'il se fixe dans les conditions réelles, ... (voir section 2.3.1)

L'intensité d'intégration est contrôlée par la construction du vecteurs des poids. Les écarts entre les poids associés aux différentes variables peuvent être plus ou moins importants, permettant ainsi de contrôler l'avantage donné aux prédicteurs soutenus par l'a priori. Le choix de l'intensité repose souvent sur la maximisation de critères de validation (tels que le $F1 - score$ ou l'aire sous la courbe Précision-Rappel) en fonction d'une référence solidement construite (Greenfield et al., 2013; Petralia et al., 2015; Siahpirani and Roy, 2017). Cependant, ce *gold standard* reste imparfait et parfois non disponible, en particulier si l'on s'intéresse à un processus biologique encore peu étudié. Des approches reposent sur la minimisation de l'erreur de prédiction sur un jeu de données test (Cirrone et al., 2020) ou bien un critère tel que le BIC (Charbonnier et al., 2010; Lingjærde et al., 2021). A notre connaissance, l'intensité d'intégration pour l'inférence de réseaux est fixée de façon globale, pour tous les gènes cibles.

Par ailleurs, si le jeu de données complémentaire est informatif, il met souvent en avant les gènes les plus connus et les mieux étudiés (ceux pour lesquels on dispose de données de validation). On serait alors tenté d'utiliser une intensité d'intégration élevée. Cela permettrait probablement d'améliorer les critères de précision et de rappel calculés sur des données de validation, ces données étant souvent liées aux données utilisées a priori. En revanche, cela a pour effet de concentrer l'intérêt sur des TFs connus, alors que d'autres sont très peu étudiés. Ainsi, Weidemüller et al. (2021) mettent en avant l'intérêt de comprendre le rôle de ces '*dark TF*'. C'est pourquoi, quelle que soit la modélisation choisie, tout l'enjeu consiste à optimiser l'intensité d'intégration des données.

Nous avons récemment initié une analyse comparative de l'intégration de données de motifs de fixation (PWM) chez la plante *A. thaliana* par *weighted Lasso* et *weighted Random Forests* (Thèse d'O. Cassan (2022), section 2.3). Ainsi deux grandes catégories de modèles fréquemment utilisés pour l'inférence de GRN sont considérées : le *Lasso*, un modèle linéaire régularisé en norme l_1 (généralisé à un modèle de Poisson pour mieux tenir compte de la nature des données de comptages) et les forêts aléatoires, un modèle non-linéaire et non-paramétrique. Ces travaux ont montré que pour la plante *A. thaliana*, la connaissance des motifs de fixation (PWM) et des séquences promotrices des gènes permet d'améliorer l'inférence de réseaux à partir de l'expression. On peut cependant noter que cela est propre au génome d'*A. thaliana*, qui est de relativement faible complexité. Il y a très peu de régions d'ADN non codant, et l'on s'attend à ce que l'essentiel des régulations soient contrôlées par la région promotrice. Cependant, des données d'ouverture de chromatine (ATAC-seq) pourrait permettre de délimiter plus précisément les régions de l'ADN au sein desquelles rechercher la présence de motifs.

Dans la continuité de ces travaux (Manuscrit en cours de rédaction), nous proposons d'optimiser l'intensité d'intégration à partir d'une approche originale reposant sur la permutation des profils d'expression des TFs (le lien entre les profils d'expression et le vecteur de poids issus de l'a priori est alors rompu). Nous introduisons un nouveau critère permettant d'optimiser l'intégration à partir de la comparaison de l'erreur de prédiction sur données permutées ou non.

Nous faisons de plus l'hypothèse que l'intégration de l'a priori est parfois utile pour certains

gènes uniquement. Aussi, dans le cadre de p modèles de régression, nous proposons de choisir l'intensité d'intégration, non pas pour l'ensemble des gènes cibles (pour les p modèles) mais individuellement, pour chaque gène cible. Dans cette étude, seulement un tiers des gènes tirent avantage de l'intégration de données de motifs de fixation de TFs.

Chapitre 3

Extraction de caractéristiques de séquences génomiques impliquées dans la régulation de l'expression

La diversité des données disponibles amène à de nouvelles questions méthodologiques, en particulier en termes de modélisation et d'inférence statistique, pour l'analyse intégrative de divers ensembles de données. Des modèles statistiques pour la compréhension de la régulation de la transcription sont développés et largement utilisés en biologie computationnelle. Et cela mobilise tous les niveaux d'observation, incluant le niveau d'expression des gènes ou transcrits, la fixation des facteurs de transcription (TF) à l'ADN, l'accessibilité de la chromatine, la structure de la chromatine et ses modifications, la fixation des TF à l'ARN, ...

Ce chapitre traite des méthodes de modélisation et d'inférence statistique pour identifier l'information encodée dans l'ADN associée à ces mécanismes de régulation de l'expression des gènes.

3.1 Modéliser la régulation de l'expression à partir de la séquence ADN

Motivations biologiques Les cellules eucaryotes (possèdant un noyau contenant le génome) produisent de nombreux types d'ARN primaires et transformés. Chez l'homme, on estime que les trois quarts du génome peuvent être transcrits. La totalité des ARN n'est pas encore connue, mais les méthodes de mesure des quantités d'ARN dans les cellules se sont largement développé et diversifié. On est capable aujourd'hui d'observer l'expression, la localisation, les régions régulatrices ainsi que des modifications des ARN, qu'ils soient déjà annotés ou non (Djebali et al., 2012).

Une part importante des capacités de régulation d'une cellule dépend de la synthèse de l'ARN, de son traitement, son transport, sa modification et sa traduction. A partir d'une même séquence ADN, la transcription du génome (ARN) est spécifique d'un contexte tel que le type de cellule ou de traitement. Aussi, l'ARN représente une expression directe de l'information génétique encodée par les génomes, que l'on cherche à identifier.

Changement de paradigme Au délà de la seule présence d'un motif spécifique, la fixation d'un facteur de transcription dépend du contexte et notamment de l'accessibilité de la chromatine, de la configuration 3D de l'ADN, de la fixation de cofacteurs de transcription (voir section 2.3.1).

Des modèles ont été développés pour identifier les régulateurs de la transcription de l'ADN à partir de données expérimentales. C'est un changement de paradigme par rapport à l'inférence de réseaux de régulation. Il s'agit d'un modèle général, construit pour l'ensemble des niveaux d'expression (ARN) à l'échelle du génome. Ainsi pour un modèle du niveau d'expression des gènes, l'individu statistique est le gène (parmi n gènes du génome) et les variables explicatives sont des caractéristiques associées à chaque gène (fixation de TF dans la région promotrice, ouverture de

chromatine, ...). Dans un cadre linéaire, on peut considérer le modèle de régression suivant :

$$Y = X\beta + \varepsilon \tag{3.1}$$

où $Y_{[n\times 1]} = (y_1, \ldots, y_n)'$ est le vecteur des niveaux d'expression de l'ensemble des gènes ou transcrits dans une condition donnée, $X_{[n\times k]} = (x_{ij})$ la matrice des variables explicatives (x_{ij} décrit la caractéristique j pour le gène i), $\beta_{[k\times 1]} = (\beta_0, \beta_1, \ldots, \beta_k)'$ le vecteur des p coefficients de régression et $\varepsilon_{[n\times 1]} = (\varepsilon_1, \ldots, \varepsilon_n)'$ le vecteur des erreurs. L'ensemble de ces données est mesuré dans une condition d'intérêt, par exemple une tumeur d'un certain type de cancer issue de la base de données `TCGA`.

On se place alors dans un contexte statistique beaucoup plus favorable que dans le cadre de l'inférence de réseaux de régulation : on dispose ici d'autant de réplicats que de gènes ou transcrits, qui est élevé même pour une seule condition. Par exemple, le génome humain possède environ 20 000 gènes codants, et la plante *A. thaliana* environ 30 000. Ainsi le nombre de réplicats n est en général plus élevé que le nombre p de variables explicatives. On sort du contexte de grande dimension. Cependant, d'une part le nombre de variables peut rester élevé, voir très élevé si l'on considère les interactions entre variables (section 3.2.1), et d'autre part, ces variables sont fortement corrélées.

Utiliser uniquement la séquence ADN Des données expérimentales telles que des données de fixation de TF ou bien des données de conformation de la chromatine sont utilisées pour expliquer le niveau d'expression des gènes, notamment par des approches de réduction de dimension dans un cadre linéaire (régression sur composantes principales (Ouyang et al., 2009), Lasso (Li et al., 2014), ElasticNet (Schmidt et al., 2017)).

Parallèlement, l'information encodée dans la séquence ADN a été utilisée pour prédire ce type de données expérimentales. Par exemple, des approches de classification reposant sur un modèle de forêts aléatoires (`Epigram` (Whitaker et al., 2015)), ou sur des réseaux de neurones convolutionnels (CNN) (`deepSea` (Zhou and Troyanskaya, 2015), `Expecto` (Zhou et al., 2018)) permettent de prédire l'accessibilité de la chromatine ou la fixation de facteurs de transcription. Ainsi, une partie des déterminants de la régulation de la transcription est associé à la séquence ADN. Cela a motivé le développement de modèles de prédiction du niveau d'expression des gènes uniquement à partir de la séquence ADN, avec des réseaux de convolution (`Expecto` (Zhou et al., 2018), `Xpresso` (Agarwal and Shendure, 2020) ou des modèles linéaires (Bessière et al. (2018), **disponible en Annexe** D).

D'une manière générale, l'étude de la séquence ADN a motivé de nombreux développements à partir de méthodes d'apprentissage profond. Les réseaux de convolution, qui se sont montré très performants pour l'analyse d'images via la prise en compte du voisinage en 2 dimension sont particulièrement adaptés pour réaliser une prédiction à partir de séquence ADN (1D). En pratique, la séquence ADN est encodée via une matrice à 4 lignes (une pour chaque nucléotide A, C, G, T) dans laquelle chaque colonne indique le nucléotide lu à la position t (*One hot encoding matrix*). Les développements de CNN pour l'analyse de séquence ADN sont nombreux que ce soit dans un objectif de classification (`deepbind` (Alipanahi et al., 2015), `deepSEA` (Zhou and Troyanskaya, 2015), `Expecto` (Zhou et al., 2018), `DanQ` (Quang and Xie, 2016), `BPnet`(Avsec et al., 2021)), ou dans un objectif de régression (`Expecto` (Zhou et al., 2018), `Bassenji` (Kelley et al., 2018), `Xpresso` (Agarwal and Shendure, 2020), `deepSTR` (Grapotte et al., 2021), `Enformer` (Avsec et al., 2021)).

Deep vs. shallow learning L'utilisation de modèles simples (tels que des modèles linéaires) est souvent justifiée par la facilité d'interprétation et le faible coût en temps de calcul, même si la qualité de prédiction peut être inférieure à celle de modèles plus complexes. Cependant, la disponibilité de gros volumes de données favorise et généralise l'utilisation de méthodes d'apprentissage profond. Ce dilemme *deep* versus *shallow learning* (voir (Zrimec et al., 2021) pour une revue des différentes approches pour l'analyse de séquences génomiques) se place au cœur du compromis entre qualité de prédiction et interprétabilité des modèles.

Contrairement aux modèles d'apprentissage profond, les modèles *shallow* utilisent des caractéristiques construites à partir d'un jeu de données d'intérêt, et n'exécutent que quelques étapes

d'inférence (associées à un modèle statistique choisi, linéaire ou non). Il est a priori plus simple d'expliquer un modèle *shallow*, bien que cela soulève des questions non résolues dans un contexte de forte corrélation entre les variables explicatives, en particulier en termes de robustesse et de mesure de l'importance des variables (voir chapitre 4). Par ailleurs, l'explicabilité des modèles issus de l'apprentissage profond est un domaine très étudié et de nombreux développements ont été proposés pour la compréhension des mécanismes de régulation (voir Li et al. (2023) pour une revue récente). Certaines approches reposent sur l'utilisation d'exemples (`deepbind` (Alipanahi et al., 2015)), ou de perturbations des données en entrée du modèle (`deepbind` (Alipanahi et al., 2015), `BPnet` (Avsec et al., 2021)). D'autres approches permettent de calculer une importance pour chaque élément en entrée, par exemple à partir du calcul du gradient du modèle de prédiction par rapport aux variables d'entrée (Sundararajan et al., 2017), de la valeur de Shapley qui repose sur l'importance d'une variable au sein de tous les sous-ensembles de variables (Lundberg and Lee, 2017) ou bien d'algorithmes de *backpropagation* (`deepLift` (Shrikumar et al., 2017)).

La suite de ce chapitre est dédiée à la catégorie de modèles *shallow*, et plus particulièrement à des méthodes reposant sur des modèles linéaires pour la régression ou la classification, que nous avons proposés pour l'extraction de caractéristiques de séquences.

3.2 Construction de caractéristiques

Le choix des caractéristiques considérées dans un modèle statistique est cruxial. Construire des prédicteurs pertinents est tout l'enjeu des approches de *feature engineering* en amont de l'étape de *feature selection*. La qualité du modèle repose alors sur la construction de la variable réponse et des caractéristiques pertinentes à inclure au modèle. La connaissance du domaine d'application - ici la biologie moléculaire et les données Omics - est alors essentielle.

3.2.1 Expliquer la quantité de transcrits (Régression)

Motivations biologiques La présence de motifs de fixation de facteurs de transcription dans des régions impliquées dans la régulation de la transcription, en particulier le promoteur, conditionne certains modes de régulation. On sait également que la composition en nucléotidique joue un rôle essentiel (voir Boeva (2016) pour une revue). Tout d'abord, les séquences riches en `GC` (nucléotide `G` ou `C`, nucléotides complémentaires sur les 2 brins de l'ADN), ou en `AT` (nucléotide `A` ou `T`) sont souvent situées dans les promoteurs et participent à l'initiation de la transcription. En effet, environ un quart des promoteurs humains contiennent une séquence riche en `AT` dont 10% environ contiennent le motif `TATA` canonique, qui recrute des protéines capables d'ouvrir la chromatine. La force de liaison entre les nucléotides `A` et `T` sur les 2 brins complémentaires de l'ADN sont moins fortes (qu'entre `G` et `C`) et cela facilite également l'ouverture de la chromatine (Yang et al., 2007). Les trois quart restant des promoteurs humains sont riches en `GC` et peuvent contenir plusieurs occurrences d'un site de fixation d'un activateur de la transcription (SP1, famille des *zing finger*). Ces régions peuvent également contiennent des ilôts `CpG` (c'est à dire des séquences de faible complexité, riches à la fois en `GC` et en dinucléotides `CpG`, longues de 300 à 3000 nucléotides) dont le niveau élevé de méthylation est associé à la répression de la transcription.

Par ailleurs, la composition des séquences impacte naturellement les scores de motifs. Les motifs contenant plusieurs occurrences des nucléotides `G` ou `C` ont un score naturellement plus élevé dans une région riche en `GC`. Aussi, prendre en compte la composition en nucléotides permet d'identifier la part d'information provenant uniquement des motifs.

Construire des caractéristiques à partir d'annotations issues de bases de données Dans une première approche, un modèle linéaire avec sélection de variable via le lasso (Tibshirani, 1996) nous a permis de montrer que la composition en nucléotide et dinucléotide de certaines régions bien choisies dans la séquence ADN contient une information associé au niveau d'expression des gènes (Bessière et al. (2018), **disponible en Annexe** D). Ce modèle permet de prédire le niveau d'expression des gènes avec une qualité de prédiction comparable à celle de modèles utilisant des données épigénétiques tels que `RACER` (Li et al., 2014), ou `TEPIC` (Schmidt et al., 2017).

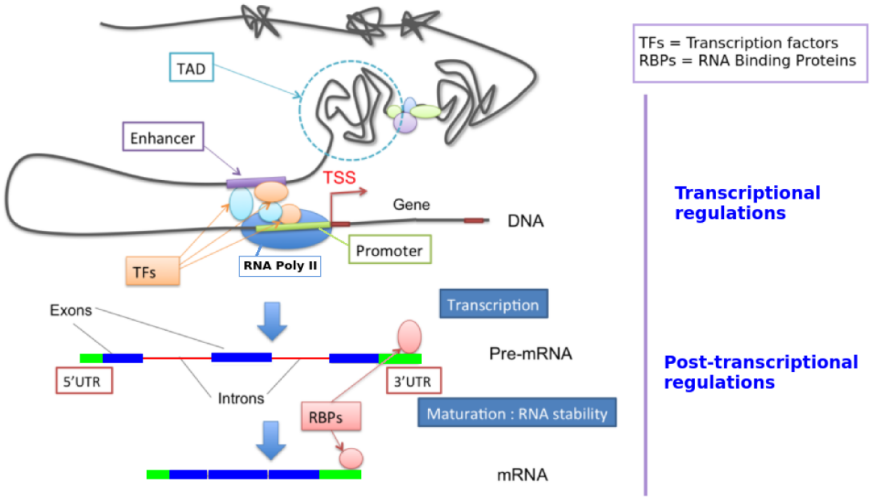


FIGURE 3.1 – **Aperçu de différents mécanismes de régulation de l'expression des gènes.** Ici la régulation de la transcription est contrôlée par la fixation de plusieurs TF dans deux régions promotrices : le promoteur du gène, situé à proximité du site d'initiation de la transcription (TSS), et une région promotrice distante (*enhancer*). À partir de la séquence ADN, un gène est d'abord transcrit en *pre-mRNA* par des ARN polymérases (RNA poly II). Le *pre-mRNA* subit ensuite un processus de maturation au niveau post-transcriptionnel,qui abouti à un ARN messager stable. Pendant la maturation de l'ARN, des introns sont éliminés de façon spécifique (épissage alternatif). Les extrémités 5'UTR et 3'UTR sont également impliquées dans la stabilité de l'ARN et peuvent être fixées par des protéines spécifiques (*RNA-binding proteins* ou (RBP)).

Il s'agit de régions spécifiques délimitées au moyen d'annotations issues des bases de données `GENCODE` (Harrow et al., 2012) et `Ensembl` (Aken et al., 2016), à partir de la connaissance actuelle des mécanismes de régulation de l'expression des gènes (Figure 3.1). En particulier, le cœur du promoteur (à proximité du TSS) mais également des régions particulières telles que les introns, les régions codantes (exons) ou les extrémités des gènes (5'UTR ou 3'UTR) se sont révélé informatives. Ceci va à l'encontre du point de vue selon lequel l'expression des gènes dépend en priorité des marques épigénétiques, qui ne sont pas nécessairement contrôlées par la séquence ADN (Zeitlinger, 2020). Cependant, l'information apportée par la présence de motifs n'apporte que peu à ce modèle général de la transcription, et dans nos analyses chez l'Homme, ce sont essentiellement les gènes dit de ménage (*housekeeping genes*) dont l'expression est expliquée.

Utiliser un modèle heuristique (*Model-based feature engineering*) Des régions de faible complexité, c'est-à-dire des régions enrichies en un mot de longueur k, comme par exemple les ilôts `CpG`, sont également associées à la régulation de la transcription et connues sous le nom de *Long Regulatory Elements (LRE)*. Dans le cadre de sa thèse, C. Menichelli a cherché à identifier, et à délimiter des régions situées à proximité du TSS et dont la fréquence d'un mot donné de longueur k ($k \geq 2$) est corrélée au niveau d'expression du gène par une approche nommée DExTER pour *Domain Exploration To Explain gene Regulation* (Menichelli et al., 2021). Un parcours selon un demi-treillis permet de considérer efficacement un ensemble de combinaisons (région x k-mer), avec une initialisation à $k = 2$. L'ensemble des caractéristiques pré-sélectionnées, ainsi que les scores de fixation de motifs (issus des PWM) sont ensuite confrontés dans un modèle linéaire avec sélection de variables via le Lasso (Tibshirani, 2013). Les modèles appris sont parfois spécifiques d'une condition, mais seulement pour certains organismes, notamment pour *Plasmodium*, un parasite associé à la transmission du paludisme. Bien que quelques TF spécifiques aient été identifiés, les principaux facteurs de la régulation de la transcription sont présents dans le génome de Plasmodium (Toenhake et al., 2018). La pertinence d'une des régions identifiées située en amont du TSS enrichie en le trinucléotide `ATA` a pu être évaluée de façon expérimentale *in vivo*.

Inclure les interactions Certaines caractéristiques de séquence ont un effet seulement en interaction avec d'autres ; par exemple lorsque la fixation de 2 TF est requise pour activer la transcription d'un gène, ou encore si un motif de fixation d'un TF n'est fonctionnel que dans une région riche en certains nucléotides ou dinucléotides. Aussi, un de nos objectifs consiste à identifier des interactions entre les caractéristiques considérées. Face à la dimension du problème, une approche classique est de considérer une hypothèse de structure hiérarchique entre les effets simples et les interactions (Nelder, 1977; Chipman, 1996). Une interaction entre deux variables est alors considérée seulement si ces deux variables ont chacune un effet propre significatif (hiérarchie forte), ou alors si c'est le cas pour au moins l'une des deux variables (hiérarchie faible). Différentes approches ont été proposées pour la sélection d'interactions via le LASSO à partir de ces hypothèses (Bien et al., 2013; Lim and Hastie, 2015). Nous avons proposé de relâcher cette hypothèse d'hérédité (forte et faible) via une adaptation de l'Elastic Net (Zou and Hastie, 2005) permettant de considérer toutes les interactions d'ordre 2 (Bascou, 2022). Afin d'éviter de construire la matrice des interactions, nous avons adapté un algorithme de descente par coordonnées (Tseng, 2001; Friedman et al., 2007), qui permet de mettre à jour les coefficients d'interactions un par un (les colonnes de la matrice des interactions à la volée, sans jamais avoir besoin de la stocker entièrement en mémoire). Dans objectif de parcimonie, des algorithmes d'ensembles actifs ou *active set* (Fan et al., 2008) ont été développés pour sélectionner efficacement un sous-ensemble de variables exploitant un tri des variables en amont. C'est le cas de l'algorithme `CELER` (Massias et al., 2018), que nous avons adapté pour la considération d'interactions d'ordre 2 en séparant les deux ensembles de variables (effets simples et effets d'interaction). Ces développements ont permis de considérer toutes les interactions d'ordre 2 dans le modèle proposé dans notre première approche pour la modélisation de l'expression des gènes à partir de la séquence (Bessière et al. (2018), **disponible en Annexe** D). Différents opérateurs d'interactions (produit, minimum, maximum) ont été comparés. La majorité des interactions sélectionnées croisent deux effets principaux issus de régions différentes, ce qui nous incite à étudier les interactions entre les régions (promoteur, introns, exons, extrémités du gène).

3.2.2 Caractériser la grammaire de fixation des TFs (Classification)

L'apport des motifs de fixation reste moindre dans les modèles de l'expression des gènes présentés dans la section précédente. Ceci peut s'expliquer par le fait qu'un modèle général de l'expression de l'ensemble des gènes est assez peu spécifique de la condition choisie (section 3.2.1). Afin d'étudier la grammaire de fixation des TF, nous avons cherché à modéliser la fixation à l'ADN (Chip-Seq) de façon spécifique à chaque TF.

Grammaire de fixation des TF Il existe de l'ordre de 2000 TFs dans les cellules humaines, chacun reconnaissant des éléments génomiques différents, et notamment un ou plusieurs motifs de fixation. Un motif annoté comme associé à un TF est généralement utilisé pour prédire sa fixation sur une séquence ADN. L'approche classique consiste à retenir le meilleur score de motif pour chaque séquence, et à prédire comme fixées les séquences au-dessus d'un certain seuil. Mais la fixation des TF ne dépend pas uniquement de la présence d'un motif de fixation. Il y a beaucoup de faux positifs. Un élément déterminant est la configuration de l'ADN, qui est associé à sa composition nucléotidique (voir `Epigram` (Whitaker et al., 2015), `deepSea` (Zhou and Troyanskaya, 2015), `Expecto` (Zhou et al., 2018), section 3.1). Ainsi le contexte nucléotidique de la région autour du motif peut contribuer largement à améliorer le modèle (Levo et al., 2015). Par ailleurs, la fixation des TF peut se faire de façon coopérative, entre plusieurs TF. Plusieurs modèles biologiques de fixation des TF ont été décrits, connus sous le nom de 'grammaire' de fixation des TF (voir Jindal and Farley (2021) pour une revue récente). De multiples mécanismes peuvent conduire à la coopération des TF (Morgunova and Taipale, 2017; Reiter et al., 2017; Ibarra et al., 2020). La coopération peut impliquer des interactions directes entre deux TF (interaction protéine-protéine), requises pour la fixation à l'ADN. Mais elle peut aussi impliquer la fixation de plusieurs TFs à l'ADN, selon différentes règles. Ainsi, le modèle *enhanceosome* est défini par un ordre, une distance, une affinité et/ou une orientation relative pour plusieurs sites de fixations. Il s'oppose au modèle *billboard* qui autorise une grammaire plus souple de l'arrangement des sites de fixation des TF. Aussi le modèle statistique pour la fixation de TF peut être largement amélioré en ajoutant la

contribution d'autres facteurs tels que la composition nucléotidique de la séquence ADN et la présence d'autres motifs de fixation que celui du TF cible.

Il a de plus été observé que les cofacteurs d'un TF peuvent varier d'un type cellulaire à un autre. L'enjeu consiste alors à identifier la combinatoire de cofacteurs spécifique d'un type cellulaire, pour chaque TF.

Identifier des cofacteurs spécifiques d'une condition Wang et al. (2018) ont proposé d'identifier les motifs de fixation associés à la fixation de 2 TFs (TCF7L2 et MAX) dans un type cellulaire donné via un modèle de forêts aléatoires. Les variables explicatives considérées en entrée sont le nombre d'occurrences de chaque motif. Le package R `inTrees` Deng (2014) leur permet d'extraire des règles de décision sur-représentées dans l'ensemble des arbres. Au total plus de 2000 motifs issus de la base de données mise à disposition par le projet `ENCODE` (Luo et al., 2020) sont utilisés dans cette étude.

Nous avons construit un classifieur linéaire (régression logistique avec régularisation lasso) exploitant les données de motifs de fixation ainsi que la composition nucléotidique des séquences ADN (Vandel et al., 2019). Afin de limiter les redondances et la dimension de l'espace de recherche, nous avons utilisé une sélection d'environ 600 motifs de fixation issus des bases de données `JASPAR` (Mathelier et al., 2016a), `CisBP` (Weirauch et al., 2014) et `HOCCOMOCCO` (Kulakovskiy et al., 2016). Avec ce modèle, nommé `TFcoop`, nous avons réalisé une analyse de la combinatoire des cofacteurs asociée à la fixation d'un TF cible, sur des séquences promotrices de différents types d'ARN (mRNA, lncRNA, miRNA), ainsi que sur des séquences dites *enhancers*, qui sont des régions impliquées dans la régulation de la transcription, bien qu'éloignée du site d'initiation de la transcription (voir Figure 3.1 et Andersson (2015) pour une revue). A travers une analyse sur un vaste jeu de données de fixation de TF (409 expériences pour 106 TF et 41 types cellulaires dstincts) issus de la base de données `ENCODE` (Luo et al., 2020), ce modèle montre une meilleure qualité de prédiction que d'autres approches exploitant des données *DNA shape* issues de modèles physiques (Rohs et al., 2009; Li et al., 2017) prenant en compte la structure de l'ADN (`TRAP` (Roider et al., 2007), `DNAshape` (Mathelier et al., 2016b). La qualité de prédiction est comparable à celle obtenue avec `deepSEA` (Zhou and Troyanskaya, 2015) à partir d'un modèle CNN. De plus, cette analyse permet d'étudier la fixation de promoteurs dans des processus biologiques spécifiques. Notamment, nous avons pu mettre en évidence que, pour un TF et un type de cellule donnés, les combinaisons de TF sont différentes entre les promoteurs et les enhancers (*i.e.* les modèles ne sont pas interchangeables). Par ailleurs, les combinaisons qui régissent la fixation des TF sur les enhancers sont plus spécifiques du type de cellule que dans les promoteurs. Enfin, l'analyse des TF coopérants (qui contribuent à la fixation d'un autre TF) montre une sur-représentation des TFs pionniers, dont la fixation est parfois requise pour ouvrir la chromatine condensée avant la fixation d'autres TFs (Zaret and Carroll, 2011; Sherwood et al., 2014).

Plus récemment, des modèles utilisant des données de conformation de l'ADN en plus de la séquence ADN par apprentissage profond (Wang et al., 2021; Zhang et al., 2021) ou forêts aléatoires (Barissi et al. 2022) (Barissi et al., 2022) ont montré de meilleures performances que `deepbind` (Alipanahi et al., 2015) un modèle CNN utilisant uniquement la séquence ADN. Ceci peut être considéré comme un challenge pour mieux capturer l'information de séquence pour prédire la fixation de TF.

Construire des caractéristiques pour discriminer 2 conditions Avec la diversité des types cellulaires et des conditions expérimentales, se sont développé naturellement des approches de classification multi-classe pour caractériser les facteurs de fixation d'un TF spécifique d'un type cellulaire ou d'une condition (Wang et al., 2018; Yuan et al., 2019).

Dans le cadre d'une étude des mécanismes de régulation de ces deux TFs dans les cancers du sein dits "triple négatifs" (Bejjani et al., 2021), le modèle `TFcoop` (Vandel et al., 2019) nous a permis de discriminer les séquences fixées *in vivo* par deux TFs (Fra-1 et Fra-2) de la même famille (Fos) et ayant des motifs de fixation très proches. Cependant, le motif associé à un TF, tel que représenté par une matrice PWM issus d'une base de données, n'est pas spécifique d'une de ces conditions.

De même, les TF paralogues (descendants d'un même TF ancestral) reconnaissent des motifs

similaires mais peuvent se fixer dans des régions génomiques très différentes. Nous avons proposé un classifieur linéaire, nommé `TFscope`, qui a pour enjeu d'identifier les caractéristiques de l'ADN qui différentient la fixation observée entre deux expériences de fixation de TF (ChIP-seq), ciblant soit le même TF dans deux conditions différentes (types cellulaires ou traitements), soit deux TF paralogues (Roméro et al. (2022), **disponible en Annexe** A). Afin de discriminer deux conditions, `TFscope` réalise une sélection (via une régularisation Lasso) parmi plusieurs caractéristiques construites en amont : (i) un nouveau motif discriminant, (ii) l'environnement nucléotidique autour du site de fixation et (iii) la présence de motifs de fixation associés à d'autres TF, dans une région délimitée spécifiquement pour chaque motif et pour les deux conditions considérées. Le motif discriminant est construit à partir d'un modèle logistique exploitant l'information du nucléotide présent parmi les quatre `A`, `C`, `G`,`T` (variable catégorielle) à chaque position t du motif général pour distinguer les deux conditions. L'environnement nucléotidique est décrit via la fréquence de k-mer dans des sous-régions identifiées via une adaptation de l'approche `DExTER` (Menichelli et al., 2021) pour la classification. La région associée à chaque autre motif est elle aussi déterminée par une approche exhaustive.

Cette approche identifie ainsi les principales caractéristiques de la séquence ADN utilisée pour la prédiction ainsi que la contribution de chacune de ces caractéristiques à l'explication des différences de liaison. A travers une étude sur plus de 350 comparaisons de contexte de fixation, nos expériences ont montré que les caractéristiques génomiques qui distinguent la fixation des TF dans deux contextes différents varient en fonction des TF considérés et/ou des conditions. Pour discriminer la fixation d'un même TF dans différents types de cellules ou sous différents traitements, la fixation de TF 'coopérants' et l'environnement nucléotidique expliquent la plupart des différences entre les sites de fixation. En revanche, pour les TF paralogues, des différences subtiles dans le motif du TF cible semblent être la principale raison des différences.

Chapitre 4

Mesure de l'importance des variables

La capacité à interpréter correctement les résultats d'un modèle de prédiction est extrêmement importante. D'une part, elle est source de confiance de la part de l'utilisateur, mais elle aide également (i) à savoir comment un modèle peut être amélioré et (ii) à comprendre le processus modélisé. Aussi l'explicabilité des modèles prend une part de plus en plus importante dans les développements méthodologiques en apprentissage statistique (Murdoch et al., 2019; Belle and Papantonis, 2021), y compris en dehors des méthodes d'apprentissage profond.

Dans un contexte de forte corrélation entre les variables prédictives, les modèles les plus simples manquent de robustesse et la mesure de l'importance relative des variables est une tâche difficile.

4.1 Modèles linéaires

Avec les modèles linéaires régularisés en norme l_1, on obtient naturellement un sous ensemble parcimonieux qui minimise l'erreur de prédiction sous contrainte (par validation croisée ou selon un critère tel que C_p, BIC ou AIC) (Hastie et al., 2020). En revanche, dans un contexte de grande dimension et de fortes corrélations entre les prédicteurs, ce sous-ensemble n'est pas robuste, et a tendance à inclure un nombre important de faux positifs (Knight and Fu, 2000; Zou, 2006).

Sous-échantillonnage Des méthodes de sous-échantillonnage sont utilisées pour identifier un sous-ensemble stable de variables explicatives, notamment par une simple approche bootstrap (*Bolasso* (Bach, 2008), *StARS-GGM* (Liu et al., 2010)), par l'introduction d'un aléa sur le poids de la pénalité lasso (*Randomized Lasso* (Meinshausen and Bühlmann, 2010)), ou encore par une division aléatoire du jeu de données pour ne retenir que les variables sélectionnées conjointement sur les 2 sous-échantillons ($CPSS$ pour *Complementary Pairs Stability Selection* (Shah et al., 2013)). Pour l'inférence de GRN (Chapitre 2), la méthode `TIGRESS` (Haury et al., 2012) utilise une déclinaison du *Randomized Lasso*, `The inferelator` (Skok Gibbs et al., 2022) l'approche $StARS - GGM$ adaptée pour le Lasso. La sélection de variables se fait alors en fonction de la fréquence de sélection sur l'ensemble des sous-échantillons. Cela nécessite de fixer a minima 2 paramètres : le poids de la pénalité λ pour minimiser l'erreur de prédiction et un seuil π de fréquence de sélection des arêtes.

Récemment, Bodinier et al. (2023) ont mis en évidence que le choix du seuil de fréquence de sélection π peut dépendre de nombreux paramètres, et notamment de la dimension et de la topologie du réseau. Ils proposent une nouvelle approche pour optimiser conjointement ces 2 paramètres (λ, π) via une approche originals qui consiste à minimiser la vraisemblance jointe $L_{\lambda,\pi}$ de la classification des variables explicatives (entre trois catégories : sélectionnée de façon stable, non sélectionnée de façon stable ou instable) sous l'hypothèse d'équi-probabilité de sélection, avec l'idée qu'un bon classifieur ne donne pas la même probabilité de sélection à toutes les variables.

Apprentissage multi-tâches Dans le cas de plusieurs jeux de données relativement proches (succession de points de temps, réplicats biologiques, ...), l'apprentissage simultané des modèles

associés à chaque condition avec un partage d'information favorisant la sélection des mêmes variables dans les différentes conditions permet de stabiliser la sélection de variables. Le critère de pénalisation *group lasso* (Yuan and Lin, 2006) permet cela en définissant comme groupe de variables, la même variable au sein de chacun des modèles associés à une condition : Si une variable est sélectionnée, elle l'est pour toutes les conditions. Cette pénalisation incite à choisir un petit nombre de variables qui permettent d'expliquer l'ensemble des réponses observées dans les différents contextes. Elle a été utilisée par exemple par (Liu et al., 2014) pour l'inférence de GRN. Cette approche s'est également révélé particulièrement utile pour identifier un sous-ensemble robuste de variables dans le contexte de forte corrélation généré par les caractéristiques construites à partir d'une recherche exhaustive de composition en mots de k lettres dans une sous-région (Menichelli et al., 2021). Selon le contexte et les objectifs, différentes déclinaisons de pénalité en norme l_1 peuvent être envisagées. En particulier, l'approche *sparse group* Lasso (Rao et al., 2013) encourage de plus à la sélection de variables à l'intérieur des groupes de variables sélectionnées. A l'inverse, l'approche *exclusive* Lasso (Zhou et al., 2010) permet d'identifier des prédicteurs utilisés exclusivement dans une des tâches ou condition, et permet ainsi d'identifier des variables spécifiques.

Importance relative (Approches par permutations centrées sur l'individu) Dans les travaux menés pour la construction d'un modèle général de l'expression des gènes (Bessière et al. (2018) et **disponible en Annexe** D), nous avons utilisé une approche par permutation afin de vérifier la confiance que l'on peut avoir dans le sous-ensemble de variables sélectionnées. Pour chaque individu statistique (ici le gène i), les valeurs des variables explicatives qui lui sont associées sont permutées : les variables explicatives (colonne de la matrice X, Eq. (3.1)) n'ont alors plus de sens. La permutation sur les données de composition de séquence fait chuter les performances du modèles comme attendu, mais ce n'est pas le cas pour les données expérimentales (données de fixation de TF ou de méthylation) utilisées par `RACER` (Li et al., 2014) et `TEPIC` (Schmidt et al., 2017). Cela met en évidence de fortes redondances dans ces jeux de données expérimentales : le modèle utilise essentiellement l'information commune portée par ces variables, traduisant ici l'accessibilité de la chromatine. Schmidt and Schulz (2019) ont ensuite proposé une normalisation du score de fixation de facteur de transcription à partir de le l'ensemble des TFs fixés dans une région donnée, ce qui permet de capturer l'importance d'un TF donné, en dehors de l'accessibilité de la chromatine.

4.2 Aggrégations d'arbres

Critères empiriques Un arbre de régression est construit en sélectionnant à chaque noeud la variable explicative (et le seuil associé) qui minimise l'erreur de prédiction sur l'échantillon d'apprentissage. Les forêts aléatoires fournissent une prédiction à partir de la moyenne des prédictions sur l'ensemble des arbres, mais ne permettent pas directement de sélection ou de tri des variables. Cela requiert une analyse *post-hoc*.

Deux critères ont été initialement proposés pour mesurer l'importance relative des variables explicatives des RFs dans un objectif de régression : la diminution de la précision (MDA pour *Mean Decrease in Accuracy*) (Breiman, 2001) ou la diminution de l'impureté des noeuds (MDI pour *Mean Decrease in Impurity*) (Breiman, 2002) apportée par une variable explicative donnée. Le critère MDA pour une variable X_k mesure l'augmentation de l'erreur de prédiction sur la part de l'échantillon non utilisée pour construire un arbre (*out-of-bag*) lorsque les valeurs observées de la variable X_k sont permutées (ce qui rompt sa relation avec la variable réponse, mais aussi avec les autres variables explicatives). On peut noter cependant qu'il n'y a pas de consensus dans la formulation exacte du critère MDA (Bénard et al., 2022), seules les implémentations proposées par `randomForest` (Liaw and Wiener, 2002), `ranger` (Wright and Ziegler, 2017) et `randomForestSRC` (Ishwaran and Kogalur, 2023) suivent la définition originale :

$$MDA(X_k) = \frac{1}{M}\sum_{l=1}^{M} MSE_{l,(X_k\ permuted)} - MSE_l \tag{4.1}$$

avec $MSE_l = \frac{1}{n_l}\sum_{m=1}^{n_l}(y_m - \hat{y}_m)^2$ la moyenne des erreurs de prédiction sur l'out-of-bag pour l'arbre l, n_l le nombre d'observations non utilisées pour construire l'arbre l (taille de l'out-of-bag) et $MSE_{l,(X_k\ permuted)}$ le même calcul réalisé à partir des prédictions obtenues lorsque les valeurs de la variable X_k sont permutées aléatoirement.

Pour une mesure d'impureté donnée (ici la variance, Eq. (4.2)), le critère MDI pour une variable X_k est obtenu à partir une somme pondérée des diminutions d'impureté sur chaque nœud t associé à la variable X_k (Eq. (4.3,4.4)), moyennée sur tous les arbres T_m de la forêt aléatoire.

Pour ces deux critères, une valeur élevée signifie que la variable explicative porte une part importante dans la prédiction de la forêt.

$$I(t) = \frac{1}{n(t)}\sum_{j=1}^{n(t)}(y_j - \overline{y}_t)^2 \tag{4.2}$$

$$\Delta_t = I(t) - \left(\frac{n(t_l)}{n(t)}I(t_l) + \frac{n(t_r)}{n(t)}I(t_r)\right) \tag{4.3}$$

$$MDI(X_k, T_m) = \sum_{t\in T_m, v(t)=X_k}\frac{n_t}{n}\Delta_t \tag{4.4}$$

$$MDI(X_k) = \frac{1}{M}\sum_{m=1}^{M}MDI(X_k, T_m) \tag{4.5}$$

Identification de biais et améliorations Différents biais on été mis en évidence dans les méthodes empiriques de mesure de l'importance des variables. Tout d'abord dans un contexte de classification, des biais ont été mis en évidence pour des variables catégorielles (favorisant les variables avec beaucoup de catégories (Strobl et al., 2007; Nicodemus, 2011), et les variables avec des catégories sur-représentées (Nicodemus, 2011; Boulesteix et al., 2012)) et en présence de corrélations entre les variables explicatives (Nicodemus and Malley, 2009). Ces biais ont été partiellement corrigés récemment (Li et al., 2019; Zhou and Hooker, 2021; Loecher, 2022).

Par ailleurs, Bénard et al. (2022) ont caractérisé des biais présents les deux critères empiriques couramment utilisés (MDI et MDA). L'importance mesurée par le critère MDI contient une partie de la variance résiduelle, répartie sur l'ensemble des variables explicatives. Le critère MDA mesure lui la somme de deux indices de Sobol (*Total Sobol index* (Sobol, 1993) et *Full total Sobol index* (Mara et al., 2015; Benoumechiara, 2023)), multipliée à la variance de la réponse, et additionnée d'un terme difficile à interpréter. Une exception est faite dans le cas de variables explicatives indépendantes, le MDA mesure bien l'importance relative des variables (multipliée à la variance de la réponse), mais cette situation se présente rarement. Ils définissent un nouveau critère appelé $Sobol - MDA$ pour lesquels des résultats de convergence asymptotique vers l'indice de Sobol (Sobol, 1993). Le $Sobol - MDA$ permet de mesurer l'importance relative d'une variable en retirant l'effet de cette variable dans le modèle (sans appliquer une nouvelle fois l'algorithme sur les $p-1$ autres variables explicatives). Ce critère est également utilisé pour l'identification de règles de décision stables (interactions d'ordre 2) à partir d'une aggrégation d'arbres, de profondeur égale à 2 avec la méthode `SIRUS` (Bénard et al., 2021a,b).

4.3 Construction d'un réseau

La sélection des arêtes au sein d'un graphe est un problème de classification : on chercher à identifier les arêtes présentes. Dans le cadre de p sous-problème de régression (chacun associé à un gène cible) présenté dans la section 2.2, on recherche un sous ensemble robuste de variables explicatives (facteurs de transcription) qui minimise l'erreur de prédiction. Dès lors, que le cadre soit linéaire ou non, la construction d'un réseau global nécessite de sélectionner (ou d'interclasser) des arêtes issues de p modèles de régression. On cherche alors à mesurer une importance relative à travers p modèles de régression.

Pour les forêts aléatoires, les critères MDI et MDA évaluent l'importance relative des variables à l'intérieur d'un modèle de régression. En revanche, les ordres de grandeur des valeurs prises par

	ML régularisé + Stabilité	
The inferelator 3.0	Skok Gibbs et al. (2022)	Rang moyen de la variable X_k
(AMuSR - multi task)	Castro et al. (2019)	en terme de variance expliquée S_{ik} :
(TF activity inference)	Arrieta-Ortiz et al. (2015)	$S_{ik} = 1 - \frac{\hat{\sigma}^2_{\text{modèle sélectionné}}}{\hat{\sigma}^2_{\text{modèle sans } X_k}}$
(Structure prior)	Greenfield et al. (2013)	
(StARS-lasso)	Miraldi et al. (2019)	Rang de la variable k (fréq. de sélection) + $\lvert pcorr(Y_i, X_k)\rvert$
TIGRESS	Haury et al. (2012)	Fréq. de sélection moyenne de la variable X_k parmi les $l = 1, ..., L$ premières.
	Forêts aléatoires	
DIANE	Cassan et al. (2021)	MDA + p-valeur non-paramétrique
dynGENIE3	Huynh-Thu and Geurts (2018)	MDI normalisé
iRafNet	Petralia et al. (2015)	MDI
GENIE3	Huynh-Thu et al. (2010)	MDI

TABLE 4.1 – Exemples de critère de tri des arêtes (régulateur X_k, gène cible i) dans les approches pour l'inférence de réseaux présentées dans ce chapitre : Modèles linéaires régularisés en norme l1 + stabilité de sélection et forêts aléatoires.

ces critères peuvent varier en fonction de la variable réponse. Pour l'inférence de GRN à partir de forêts aléatoires, le critère le plus couramment utilisé est le MDI, qui a été implémenté initialement dans l'approche `GENIE3` (Huynh-Thu et al., 2010). La méthode `dynGENIE3` (Huynh-Thu and Geurts, 2018) développée plus récemment (ainsi que la version actuelle de `GENIE3`) utilise un score MDI normalisé par la somme des poids de tous les variables explicatives (régulateurs potentiels). Outre les biais associés à ces critères évoqués dans la section précédente, ces critères permettent d'obtenir un tri des variables explicatives, mais il est difficile de placer un seuil de sélection. Une pratique courante avec les forêts aléatoires repose sur l'introduction de variables de bruit (ne portant aucune information) qui permet de définir un critère d'arrêt dans la sélection des variables. Dans un contexte de faible ratio signal-sur-bruit, cela peut mener à sélectionner un nombre trop élevé de variables, en particulier si l'on cherche à retenir un ensemble très parcimonieux comme c'est le cas pour les réseaux de régulation.

Dans l'application interactive `DIANE` pour l'inférence de réseaux (Cassan et al. (2021) **disponible en Annexe** B, voir aussi section 2.2.3), nous avons proposé de fixer un premier seuil selon la densité maximale attendue du réseau (à partir de la connaissance des réseaux en biologie), puis de calculer une p-valeur non-paramétrique pour les arêtes retenues (obtenue par une méthode de permutation des arêtes implémentée dans le package R `rfpermute` (Archer, 2022)).

Avec les modèles linéaires régularisés, de nombreux critères empiriques ont été utilisés pour estimer l'importance relative des arêtes issues de différents modèles (Table 4.1). Ajouter la valeur absolue de la corrélation partielle ($pcorr(Y_i, X_k) = cov(Y_i, X_k|X_{-k})/\sqrt{V(Y_i|X_{-k})V(X_k|X_{-k})}$) à la fréquence de sélection (Miraldi et al., 2019) permet de tenir compte de la variance de la réponse (à la différence du coefficient de régression $\beta_{i,k} = cov(Y_i, X_k|X_{-k})/V(X_k|X_{-k})$) pour interclasser les arêtes entre les différents gènes cibles, mais limite le poids de la régression régularisée dans la reconstruction du réseau. La moyenne des fréquences de sélection entre les L premières variables ($1/L\sum_{l=1}^{L} F(k,l)$ où $F(k,l)$ est la fréquence des sélection de la variable X_k parmi les l premières variables) proposée dans la méthode `TIGRESS` (Haury et al., 2012) permet de donner plus de poids aux variables fréquemment sélectionnées en premier, mais dépend fortement du nombre maximum de variables considérées L.

Greenfield et al. (2013) utilisent le rang moyen de la variable X_k en termes de variance expliquée S_{ik} dans le modèle sélectionné pour le gène i. Ce terme dépend, non pas de la différence de MSE avec ou sous la variable k comme le MDA (Equation (4.1)), mais de leur ratio :

$$S_{ik} \quad = \quad 1 - \frac{\hat{\sigma}^2_{\text{modèle sélectionné}}}{\hat{\sigma}^2_{\text{modèle sans } X_k}} = 1 - \frac{MSE_{\text{modèle sélectionné}}}{MSE_{\text{modèle sans } X_k}} \tag{4.6}$$

Considérer l'augmentation relative de l'erreur associée au retrait d'une variable (et non une différence comme le critère MDA) permet d'interclasser des arêtes issues de modèle de régression différents. Avec un modèle linéaire, la moyenne des carrés des erreurs sans X_k peut être obtenue en estimant le modèle sans la variable X_k comme proposé dans `The Inferelator` (Skok Gibbs et al., 2022), ou bien estimée en attribuant la valeur moyenne des observations à la variable X_k pour calculer une prédiction. Ce critère dépend uniquement du critère de sélection du poids de pénalité choisi pour la régularisation (CV, BIC, AIC) et du nombre de sous-échantillons. Ce nouveau critère proposé par Greenfield et al. (2013), et utilisé depuis dans la suite `The Inferelator` (Skok Gibbs et al., 2022), offre de plus l'avantage d'être borné entre 0 et 1, l'erreur de prédiction étant (sauf hasard de l'échantillonnage ou mauvais choix de modèle) généralement plus élevée après avoir retiré une variable d'un modèle sélectionné par régularisation. C'est celui que nous avons utilisé pour évaluer l'importance des variables dans une approche récemment développée pour optimiser l'intensité d'intégration pour l'inférence de GRN (Manuscrit en cours de rédaction) dans la continuité des travaux de thèse d'O. Cassan (2022).

Chapitre 5

Discussion et perspectives de recherche

Appliquer et développer l'intégration pour l'inférence de réseaux L'intégration de données pour l'inférence de réseaux a été développée dès les premiers développements de l'inférence statistique à partir de données d'expression, notamment avec les travaux de Segal et al. (2001) et Bar-Joseph et al. (2003). Avec le séquençage à haut débit, on a maintenant accès à une grande diversité de données pour observer la régulation de l'expression, et pour de plus en plus de conditions et/ou d'espèce différentes. Il s'agit aujourd'hui d'identifier et de modéliser finement les liens entre ces différents types de données qui offrent des point de vue complémentaires sur la machinerie de la régulation, et ce idéalement à grande échelle.

La complémentarité effective des jeux de données dépend de nombreux paramètres, des type de données considérées, mais aussi des conditions expérimentales et de la question étudiée. Les volumes et sources de données sont vastes, un des enjeux consiste à arbitrer l'intégration. Quand l'intégration est-elle valable ? Comment choisir l'intensité d'intégration ? Il y a un réel enjeu à être en mesure de choisir l'intensité à partir des données. Nous avons proposé une méthode pour optimiser la force d'intégration pour chaque sous-problème, en contrôlant l'intensité en fonction de l'erreur de prédiction par une approche par permutation (Preprint en cours de rédaction) qui rompt le lien entre les 2 jeux de données dans la continuité de la thèse de Cassan (2022). Cette approche est utilisable dans un cadre général de régression avec a priori, pour tout type de données continues. Cette approche permet également d'étudier la complémentarité entre 2 jeux de données, et de l'exploiter de façon optimale pour le partage d'information. Il serait particulièrement intéressant de comparer l'intérêt de différentes sources de données complémentaires à l'expression pour l'inférence de réseaux (données de fixation de TFs, de configuration de chromatine, d'interaction entre protéines, ...) voir éventuellement d'étendre cette approche pour l'intégration de plusieurs types de données omics.

L'intégration de plusieurs types de données peut également se faire par la construction du prior. En particulier l'intégration de données ATAC-seq (traduisant l'ouverture de chromatine) permet de délimiter plus finement la région ouverte autour du site d'initiation de la transcription. Un des enjeux actuels en biologie moléculaire est de comprendre le fonctionnement des modes de régulation distante, par exemple les enhancers sont des régions distantes du TSS, mais impliquées dans la régulation de la transcription via la fixation de protéines également. Il a été constaté que la fixation des TFs sur les régions enhancer sont particulièrement spécifiques du type cellulaire (Reiter et al., 2023), et nous l'avons également observé dans l'analyse réalisée avec l'approche `TFcoop` (Vandel et al., 2019) pour l'étude de la coopérativité pour la fixation de TF. Aussi, il serait intéressant d'étendre la région au sein de laquelle les motifs de fixation sont recherchés à ces régions distantes. L'état de l'art en biologie ne permet pas encore d'identifier les enhancers associés à un gène. En revanche les données d'ouverture de chromatine (ATAC-seq), les données d'expression qui permettent de séquencer le début de l'ARN (extrémité 5') CAGE pour *Cap Analysis of Gene Expression* ou encore les données de configuration 3D de l'ADN (Hi-C) sont autant de sources d'information à exploiter pour identifier des enhancer potentiels, associés à un

gène dans une condition spécifique. Inclure ces régions candidates dans nos modèles de l'expression à partir de données de séquence pourrait permettre d'identifier de nouveaux TFs impliqués dans la régulation (distante) de l'expression des gènes, et de valider certaines régions candidates.

La combinatoire de structures de réseaux est très élevée si l'on considère l'ensemble des régulations possibles au sein d'une cellule. Un moyen d'identifier une information robuste pourrait être de poser une question plus précise. Ainsi, l'inférence 'différentielle' de réseaux met en évidence, non pas l'ensemble des régulations en cours dans la cellule, mais celles qui différencient au mieux les 2 jeux de données (Thorne, 2016; Kim et al., 2018; Berest et al., 2019). L'extension pour l'inférence différentielle de réseaux (classification) des méthodes d'intégration de données via *weighted* Lasso et les *weighted* RF est également une perspective à considérer.

Affiner la modélisation de séquence pour l'expression La présence d'information encodée dans la séquence associée à la régulation de l'expression des gènes (au delà de la présence de motifs de fixation) est aujourd'hui établie. Les méthodes pour l'extraction de caractéristiques de séquence se développent et se précisent, avec des approches d'apprentissage profond ou non. A partir de caractéristiques construites à partir de connaissances biologiques et de modélisations linéaires, nous avons développé des méthodes qui permettent l'identification de caractéristiques de séquence associées au niveau d'expression des gènes (Bessière et al., 2018; Menichelli et al., 2021) et à la fixation de TF (Vandel et al., 2019; Roméro et al., 2022). Ces modèles sont des outils pour explorer de façon plus large les différents niveaux de régulation de l'expression des gènes (configuration de la chromatine, grammaire de fixation des TFs, identification et caractérisation des régions régulatrices, spécificité cellulaire, ...). Par exemple, le modèle `TFscope` pourrait être utilisé pour caractériser les contextes de fixation fonctionnelle d'un TF. En effet, la fixation d'un TF dans une région promotrice n'assure pas l'activation de la transcription (ou inhibition) du gène. Des données d'expression collectées dans les mêmes conditions que les données de fixation d'un TF permettent de distinguer les fixations fonctionnelles des non fonctionnelles. Cela pose un nouveau problème de classification supervisée à explorer avec cet outil. Par ailleurs, les bases de données proposent essentiellement des motifs de fixation consensus. Il serait intéressant de contribuer à la construction d'une base de données de motifs discriminants, tel que ceux construits dans l'approche `TFscope`, afin de faciliter et de généraliser leur utilisation.

Plusieurs directions d'amélioration de ces méthodes sont à considérer. Par exemple, un enjeu essentiel consiste à caractériser le contexte de fixation des TF, en plus d'un motif discriminant et de la combinaison de cofacteurs. Le modèle de fixation des TF à l'ADN pourrait être amélioré dans ce sens en exploitant des données de structure locale de l'ADN construites à partir de modèles physiques (*DNAshape* ou *DNA Affinity*) qui sont particulièrement informatives (Bentsen et al., 2022). Chercher à prédire ces données est en moyen d'identifier de nouveaux prédicteurs de la fixation des TF, à intégrer au modèle global.

Associer l'information de la présence de motifs de fixation à l'expression des gènes reste une question largement ouverte, et requiert des efforts de modélisation. Une démarche naturelle consiste à affiner la question, comme nous l'avons fait pour modéliser la fixation de TF avec une approche discriminante. En outre, il serait intéressant de considérer l'intégration de données d'expression des TFs associés aux motifs de fixation afin de guider la sélection de motifs. Cette fois, à l'inverse de l'intégration de motifs pour l'expression (approche initiée dans la thèse d'O. Cassan (2022)), ce serait l'expression qui viendrait supplémenter la présence de motif de fixation.

Enfin une perspective consiste à poursuivre les recherches que nous avons initiées pour capturer les interactions entre variables, en grande dimension. Les développements réalisés via une accélération de descente de gradient coordonnées par coordonnées Bascou (2022) sont destinés à plusieurs applications, afin de de considérer l'apport de différents types d'interaction entre fréquences de mots de k lettres issus de différents régions et/ou présences de motifs. Etant donné la combinatoire des interactions et la relativement faible taille des jeux de données considérés, l'identification d'interactions robustes est un problème difficile. Les interactions étant de plus fortement corrélées (entre elles et avec les variables simples), cela soulève de plus le problème de la mesure de l'importance des variables. La prise en compte des interactions est cependant clé dans la modélisation des processus de régulation de l'expression. Aussi il semble essentiel de mener cette direction de recherche de front avec celle de la mesure de l'importance des variables.

Étendre les mesures d'importance des variables dans un contexte de forte corrélation
Beaucoup s'accordent à dire qu'un bon modèle est un modèle précis, robuste et interprétable. Dans un objectif explicatif, la précision du modèle n'est pas un objectif premier. Cependant, en biologie moléculaire, et en particulier pour la compréhension de la régulation de l'expression des gènes, on se situe dans une démarche exploratoire. Il s'agit de générer des hypothèses, qui pourront idéalement être validées expérimentalement.

On manque encore actuellement de données de validation par manque de connaissance du phénomène étudié. Aussi la précision du modèle est très souvent utilisée comme un indicateur de qualité du modèle. En revanche, l'interprétabilité est au cœur de cette démarche explicative. Et dans un contexte où les variables explicatives sont fortement corrélées, l'évaluation de l'importance des variables est une question difficile. Il n'y a pas de consensus même pour des modèles classiques tels que les modèles linéaires ou les aggrégations d'arbres. Dans les travaux que nous avons menés, nous avons utilisé différentes heuristiques empiriques pour identifier des variables importantes pour la reconstruction de réseau de régulation et pour l'extraction de caractéristiques de séquence. Cela nous a permis de premières analyses, et d'identifier certaines variables importantes, mais un des enjeux consiste maintenant à approfondir la mesure de l'importance des variables.

De récentes contributions sont à considérer, en particulier des modèles de régularisation en norme l_1 adaptés à un contexte de forte corrélation (*Precision Lasso* (Wang et al., 2019), *Whitening Lasso* (Zhu et al., 2021) ; une approche pour la sélection de variables à travers la stabilité de sélection dans un modèle linéaire régularisé en optimisant conjointement les paramètres de régularisation et de stabilité (Bodinier et al., 2023) ; ou encore une nouvelle définition du calcul de l'importance des variables dans les forêts aléatoires (Bénard et al., 2022) et des interactions d'ordre 2 (Bénard et al., 2021a,b).

On peut cependant questionner l'objectif premier. Chercher une variable sélectionnée de façon robuste au sein d'un groupe de variables corrélées, a fortiori dans un contexte de grande dimension n'est peut-être pas un objectif accessible, ni même souhaitable. Dans un objectif explicatif, il serait par exemple intéressant d'identifier les ensembles de variables dont les contributions sont 'indistinguables' d'après les données, dans un contexte donné.

Deuxième partie

Liste des publications et communications

Articles en révision

1. Cassan, O., Lecellier, C.-H. , Bréhélin. , L., Martin, A. , Lèbre, S. *Optimizing data integration improves Gene Regulatory Network inference in Arabidopsis thaliana*
   https://www.biorxiv.org/content/10.1101/2023.09.29.558791v1
2. Roméro, R. , Menichelli, C., Marin, J.-M. , Lèbre, S., Lecellier, C.-H. , Bréhélin., L. *Systematic analysis of the genomic features involved in the binding preferences of transcription factors.*
   https://www.biorxiv.org/content/10.1101/2022.08.16.504098v3

Articles en biologie publiés en revues d'audience internationale

1. Cassan, O. , Pimparé, L.-L. , Dubos, C., Gojon, A., Bach, L., Lèbre, S., Martin, A.. (2023) *A gene regulatory network reveals features and regulators of the root response to eCO2 in Arabidopsis.* New Physiologist
2. Bejjani, F. Tolza, C., Boulanger, M., Downes, D. , Roméro, R., Maqbool, M. A., Andrau, J.-C., Lèbre, S., Bréhélin, L., Parinello, H., Rohmer, M., Kaoma, T., Vallar, L., Hughes, J. R., Zibara, K., Lecellier, C.-H., Piechaczyk, M. and Jariel-Encontre, I. (2021) *Fra-1 regulates its target genes via binding to remote enhancers without exerting major control on chromatin architecture in triple negative breast cancers.* Nucleic Acids Res. 2021, Mar 18 ;49(5) :2488-2508. https://pubmed.ncbi.nlm.nih.gov/33533919/

Articles publiés en revues d'audience internationale

1. Cassan, O., Lèbre, S., Martin.,A (2021) *Inferring and analyzing gene regulatory networks from multi-factorial expression data : a complete and interactive suite.* BMC genomics, 22:387. doi.org/10.1186/s12864-021-07659-2
2. Menichelli, C., Guitard, V., Lèbre, S., Lopez-Rubio, J.-J., Lecellier, C.-H., Bréhélin, L. (2021) *Identification of long regulatory elements in the genome of Plasmodium falciparum and other eukaryotes.* PLOS Computational Biology 17(4).
3. Vandel, J., Cassan, O., Lèbre, S., Lecellier, C.-H., Bréhélin, L. (2019) *Probing transcription factor combinatorics in different promoter classes and in enhancers.*
   BMC Genomics, 20 :103. https://hal.archives-ouvertes.fr/hal-02070201
4. Bessière, C., Taha, M., Petitprez, F., Vandel, J., Marin, J.-M., Bréhélin, L., Lèbre, S., Lecellier, C.-H. (2018) *Probing instructions for expression regulation in gene nucleotide compositions.* PLoS Computational Biology, 14 (1).
5. Lèbre, S. , O. Gascuel (2017). *The combinatorics of overlapping genes.* Journal of Theoretical Biology, Vol 415, pages 90-101.
6. E. Benard, Lèbre, S. , C. J. Michel (2015). *Genome Evolution by Transformation, Expansion and Contraction (GETEC).* Biosystems, Vol 135, pages 15-34.
7. F. Dondelinger, Lèbre, S. and D. Husmeier (2013). *Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure*, Machine Learning, Vol 90, Iss. 2, pages 191-230.
8. Lèbre, S., C. J. Michel (2013). *A new molecular evolution model for limited insertion independent of substitution.* Mathematical Biosciences 245, pages 137-147.
9. F. Dondelinger, D. Husmeier and Lèbre, S. (2012). *Dynamic Bayesian networks in molecular plant science : Inferring gene regulatory networks from multiple gene expression time series*, Euphytica, 183(3), 361-377.
10. Lèbre, S. , C. J. Michel (2012). *An evolution model for sequence length based on residue insertion-deletion independent of substitution : an application to the GC content in bacterial genomes.* Bulletin of Mathematical Biology, 74, 1764-1788.
11. D. Marbach et al. (2012) *Wisdom of crowds for robust gene network inference*, Nature Methods 9, 796-804 (58 auteurs). Étude globale des résultats obtenus par les équipes participant au challenge DREAM5.

12. Lèbre, S. , C. J. Michel. (2010) *A stochastic evolution model for residue Insertion-Deletion Independent from Substitution (IDIS)*. Computational Biology and Chemistry, Vol. 34, Iss. 5-6, pages 259-267, 2010.
13. Lèbre, S. , J. Becq, F. Devaux, M. P. H. Stumpf, G. Lelandais. (2010) *Statistical inference of the time-varying structure of gene-regulation networks*. BMC Systems Biology, Vol. 4, Iss. 130, pages 1-16, 2010.
14. Lèbre, S. (2009) *Inferring dynamic bayesian network with low order independencies*, Statistical Applications in Genetics and Molecular Biology, Vol. 8, Iss. 1, Article 9, pages 1-40, 2009.
15. Lèbre, S. , P-Y. Bourguignon (2008) *An EM algorithm for estimation in the Mixture Transition Distribution Model*, Journal of Statistics Computation and Simulation, Vol. 78, Iss. 8, 713, pages 713-729, 2008.

Articles publiés dans des actes de congrès avec comité de lecture (rang A)

1. Bousquet, F. and Lèbre, S. and Lavergne, C. (2020) *From mixture of longitudinal and non-gaussian advertising data to Click-Through-Rate prediction.* 24th European Conference on Artificial Intelligence (ECAI) ECAI 2020 (Taux de sélection 26.8%)
2. D. Husmeier, F. Dondelinger, Lèbre, S. (2010) *Inter-time segment information sharing for non-homogeneous dynamic Bayesian networks*. Proceedings of the The Neural Information Processing Systems (NIPS), pages 1-9, 2010 (Taux de selection : 293/1219 = 24%).
3. F. Dondelinger, Lèbre, S., D. Husmeier. (2010) *Heterogeneous Continuous Dynamic Bayesian Networks with Flexible Structure and Inter-Time Segment Information Sharing*. Proceedings of the International Conference on Machine Learning, ICML, pages 1-8, 2010 (Taux de selection : 152/594 = 25%).

Livre

1. R. Nagarajan, M. Scutari, Lèbre, S. (2013) *Bayesian Networks in R with Applications in Systems Biology*. Use R ! : Vol. 48. Springer.

Chapitres de livre

1. Lèbre, S. , F. Dondelinger, D. Husmeier. Nonhomogeneous Dynamic Bayesian Networks in Systems Biology. In : Next Generation Microarray Bioinformatics (Junbai Wang, Aik Choon Tan, Tianhai Tian eds.) Humana Press, Springer Verlag, New York, Series : Method in Molecular Biology, Vol. 802, chapter 13, pages 199-214, 2012.
2. G. Lelandais and Lèbre, S.. Recovering Genetic Network from Continuous Dynamic Bayesian Networks In : Handbook of Statistical Systems Biology (M. P. H. Stumpf, D. J. Balding and M. Girolami eds), John Wiley & Sons, Ltd, Chichester, UK, chapter 12, pages 255-269, 2011.
   `http://onlinelibrary.wiley.com/doi/10.1002/9781119970606.ch12/summary`
3. Lèbre, S. and G. Lelandais. Modeling a Regulatory Network Using Temporal Gene Expression Data : Why and How ?, In : Automation in Genomics and Proteomics : An Engineering Case-Based Approach : MIT and Harvard interdisciplinary special studies courses (R. Benson, G. Alterovitz and M. Ramoni, eds), John Wiley & Sons, Ltd, Chichester, UK, chapter 4, pages 69-96, 2009.
   `http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470727233.html`

Mémoire de thèse

S. Lèbre Analyse de processus stochastiques pour la génomique : étude du modèle MTD et inférence de réseaux bayésiens dynamiques.

Doctorat en Sciences, spécialité Mathématiques appliquées, de l'Université d'Évry-Val-d'Essonne.
http://tel.archives-ouvertes.fr/docs/00/31/14/04/PDF/slebrePhDThesis.pdf

Packages récents

1. **Application R et Rshiny** `DIANE` (Cassan, O., Lèbre, S. and Martin, A.)
   Motivés par des analyses statistiques reproductibles, nous avons partagé notre pipeline pour l'inférence GRN, (Cassan, O., Lèbre, S. and Martin, A., BMC Genomics 2021) via une interface utilisateur graphique déployée en ligne, qui se présente également sous la forme d'un package R : Dashboard for the Inference and Analysis of Network from Expression data (DIANE).
   https://diane.bpmp.inrae.fr/
2. **Package R** `BinomialMix` (Bousquet, F., Lèbre, S. and Lavergne, C.)
   Algorithme EM developpé sous R pour l'estimation d'un mélange de modèles linéaires généralisés (GLM) pour des données longitudinales. Chaque composante du mélange est un GLM dans le cas binomial, caractérisé par des covariables (continues et catégorielles) spécifiques du contexte et de la périodicité (jour, plage horaire, ...). Cette approche de classification non supervisée permet d'identifier des profils temporels similaires décrits par une combinaison de variables (Bousquet, F., Lèbre, S. and Lavergne, C., ECAI, 2020).
   https://cran.r-project.org/web/packages/binomialMix/
3. **Package R** `DCODE` (Lèbre, S.)
   Algorithme de parcours de graphe developpé sous R permettant de lister les contraintes induites sur les protéines (chaine d'acides aminés) codées par deux gènes chevauchants, en termes d'acides aminés et polypeptides, en fonction du cadre de lecture. Nous avons montré que des contraintes linéaires simples lient la composition en acides aminés de ces deux protéines. Ces contraintes sont symétriques et permettent ainsi la caractérisation de la composition conjointe de protéines chevauchantes (S. Lèbre, O. Gascuel, Journal of Theoretical Biology, 2017).
   https://cran.r-project.org/web/packages/DCODE/

Conférence internationale (invitée)

1. **Computational and Methodological Statistics (CMStatistics), Londres, 2019.**
   http://cmstatistics.org/CMStatistics2019/
   Titre de l'exposé : "Non-homogeneous dynamic Bayesian networks with Bayesian regularization for gene regulatory network inference".

2. **Workshop on Learning and Inference in Computational and Systems Biology (LICSB), Londres, 2009.**
   Titre de l'exposé : "Time-varying genetic network inference using informative priors".

Workshops (invitée)

1. Atelier IGEN (Integrating GENomic prediction with GENe regulatory networks), Cirad, Montpellier, 2022. Titre de l'exposé : "Un aperçu de méthodes statistiques pour l'inférence de réseaux de régulation"
2. Journée Statistique et Sciences de la Santé, Lille, 2022. Titre de l'exposé : "Statistical modeling and inference to identify DNA sequence elements involved in transcription regulation"
3. Journées Netbio, Paris, 2017. Titre de l'exposé : "Modélisation de l'expression des gènes à partir de données de séquence ADN".
4. Biomathematics and Statistics Scotland Seminar,Edinburgh, UK,2013
   Titre de l'exposé : "An evolution model for Limited Insertion Independent from Substitution (LIIS)"

5. Journées "Inférence de réseaux" MIA, INRA, Paris, Février 2012. Titre de l'exposé : "Réseaux bayésiens dynamiques à structure variable dans le temps".
6. Colloque du GDR de Bioinformatique Moléculaire, Paris, Novembre 2009. Titre de l'exposé : "Statistical inference of time-varying structure of gene-regulation networks".

Conférences nationales récentes (présentation par le doctorant premier auteur)

1. Roméro, R. , Marin, J.-M, Lèbre, S., Lecellier,C.-H. , Bréhélin, L. Prediction of transcription factor binding sites between cell types JOBIM 2021, Journées Ouvertes en Biologie, Informatique et Mathématiques, Paris.
2. Bascou, F., Lèbre, S. , Salmon, J. Elastic Net avec gestion des interactions et débiaisage. EGC 2021, Extraction et Gestion des Connaissances, Montpellier.
3. Bascou, F., Lèbre, S. , Salmon, J. Debiasing the Elastic Net for models with interactions. JdS 2020, 53es Journées de Statistique de la SFdS, Nice.
4. Bousquet, F., Lèbre, S. and Lavergne, C. Classification de campagnes de publicité mobile : Modèle de mélange pour données longitudinales et non gaussiennes. JdS 2019, 51es Journées de Statistique de la SFdS, Nancy.
5. Bessière, C., Taha, M., PetitPrez, F., Vandel, J., Marin, J.-M., Bréhélin, L., Lèbre, S. and Lecellier, C.-H. Modélisation de l'expression des gènes à partir de données de séquence ADN. JdS 2017, 49es Journées de Statistique de la SFdS, 2017, Avignon, France.

Troisième partie

Bibliographie

Bibliographie

Vikram Agarwal and Jay Shendure. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. Cell Reports, 31(7) :107663, May 2020. ISSN 2211-1247. doi : 10.1016/j.celrep.2020.107663. URL https://www.sciencedirect.com/science/article/pii/S2211124720306161. 26

Amr Ahmed and Eric P. Xing. Recovering time-varying networks of dependencies in social and biological studies. Proceedings of the National Academy of Sciences of the United States of America, 106(29) :11878–11883, July 2009. ISSN 0027-8424. doi : 10.1073/pnas.0901910106. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2704856/. 13

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. SCENIC : single-cell regulatory network inference and clustering. Nature Methods, 14(11) :1083–1086, November 2017. ISSN 1548-7105. doi : 10.1038/nmeth.4463. URL https://www.nature.com/articles/nmeth.4463. Number : 11 Publisher : Nature Publishing Group. 15, 16, 21

Bronwen L. Aken, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, Kevin Howe, Andreas Kähäri, Felix Kokocinski, Fergal J. Martin, Daniel N. Murphy, Rishi Nag, Magali Ruffier, Michael Schuster, Y. Amy Tang, Jan-Hinnerk Vogel, Simon White, Amonida Zadissa, Paul Flicek, and Stephen M. J. Searle. The Ensembl gene annotation system. Database, 2016 :baw093, January 2016. ISSN 1758-0463. doi : 10.1093/database/baw093. URL https://doi.org/10.1093/database/baw093. 28

Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. Nature Reviews Genetics, 16(4) :197–212, April 2015. ISSN 1471-0064. doi : 10.1038/nrg3891. URL https://www.nature.com/articles/nrg3891. Number : 4 Publisher : Nature Publishing Group. 19

Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 33(8) :831–838, August 2015. ISSN 1546-1696. doi : 10.1038/nbt.3300. URL https://www.nature.com/articles/nbt.3300. Number : 8 Publisher : Nature Publishing Group. 26, 27, 30

Christophe Ambroise, Julien Chiquet, and Catherine Matias. Inferring sparse Gaussian graphical models with latent structure. Electronic Journal of Statistics, 3(none) : 205–238, January 2009. ISSN 1935-7524, 1935-7524. doi : 10.1214/08-EJS314. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-3/issue-none/Inferring-sparse-Gaussian-graphical-models-with-latent-structure/10.1214/08-EJS314.full. Publisher : Institute of Mathematical Statistics and Bernoulli Society. 12, 18

Robin Andersson. Promoter or enhancer, what's the difference ? Deconstruction of established distinctions and presentation of a unifying model. BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology, 37(3) :314–323, March 2015. ISSN 1521-1878. doi : 10.1002/bies.201400162. 30

C. Andrieu and A. Doucet. Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. IEEE Transactions on Signal Processing, 47(10) :2667–2676, October 1999. ISSN 1941-0476. doi : 10.1109/78.790649. Conference Name : IEEE Transactions on Signal Processing. 13, 16

Olivia Angelin-Bonnet, Patrick J. Biggs, and Matthieu Vignes. Gene regulatory networks : a primer in biological processes and statistical modelling. arXiv :1805.01098 [q-bio, stat], May 2018. URL http://arxiv.org/abs/1805.01098. arXiv : 1805.01098. 9, 12

Eric Archer. rfPermute : Estimate Permutation p-Values for Random Forest Importance Metrics, March 2022. URL https://CRAN.R-project.org/package=rfPermute. 36

Mario L. Arrieta-Ortiz, Christoph Hafemeister, Ashley Rose Bate, Timothy Chu, Alex Greenfield, Bentley Shuster, Samantha N. Barry, Matthew Gallitto, Brian Liu, Thadeous Kacmarczyk, Francis Santoriello, Jie Chen, Christopher D. A. Rodrigues, Tsutomu Sato, David Z. Rudner, Adam Driks, Richard Bonneau, and Patrick Eichenberger. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. Molecular Systems Biology, 11(11) :839, November 2015. ISSN 1744-4292. doi : 10.15252/msb.20156236. 20, 36

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18(10) :1–8, October 2021. ISSN 1548-7105. doi : 10.1038/s41592-021-01252-x. URL https://www.nature.com/articles/s41592-021-01252-x. Bandiera_abtest : a Cc_license_type : cc_by Cg_type : Nature Research Journals Primary_atype : Research Publisher : Nature Publishing Group Subject_term : Gene expression ;Machine learning ;Software ;Transcriptomics Subject_term_id : gene-expression ;machine-learning ;software ;transcriptomics. 26, 27

Francis R. Bach. Bolasso : model consistent Lasso estimation through the bootstrap. In Proceedings of the 25th international conference on Machine learning, ICML '08, pages 33–40, New York, NY, USA, July 2008. Association for Computing Machinery. ISBN 978-1-60558-205-4. doi : 10.1145/1390156.1390161. URL https://doi.org/10.1145/1390156.1390161. 15, 33

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. The Journal of Machine Learning Research, 9 :485–516, 2008. Publisher : JMLR. org. 11, 12

Michael Banf and Seung Y. Rhee. Enhancing gene regulatory network inference through data integration with markov random fields. Scientific Reports, 7(1) :41174, February 2017. ISSN 2045-2322. doi : 10.1038/srep41174. URL https://www.nature.com/articles/srep41174. Number : 1 Publisher : Nature Publishing Group. 21

Ziv Bar-Joseph, Georg K. Gerber, Tong Ihn Lee, Nicola J. Rinaldi, Jane Y. Yoo, François Robert, D. Benjamin Gordon, Ernest Fraenkel, Tommi S. Jaakkola, Richard A. Young, and David K. Gifford. Computational discovery of gene modules and regulatory networks. Nature Biotechnology, 21(11) :1337–1342, November 2003. ISSN 1087-0156. doi : 10.1038/nbt890. 21, 39

Sandro Barissi, Alba Sala, Miłosz Wieczór, Federica Battistini, and Modesto Orozco. DNAffinity : a machine-learning approach to predict DNA binding affinities of transcription factors. Nucleic Acids Research, 50(16) :9105–9114, September 2022. ISSN 0305-1048. doi : 10.1093/nar/gkac708. URL https://doi.org/10.1093/nar/gkac708. 30

Florent Bascou. Sparse linear model with quadratic interactions. phdthesis, Université de Montpellier, September 2022. URL https://theses.hal.science/tel-04058087. 29, 40

Sabrina Beier, Marlene Stiegler, Eva Hitzenhammer, and Monika Schmoll. Screening for genes involved in cellulase regulation by expression under the control of a novel constitutive promoter

in Trichoderma reesei. Current Research in Biotechnology, 4 :238–246, January 2022. ISSN 2590-2628. doi : 10.1016/j.crbiot.2022.04.001. URL https://www.sciencedirect.com/science/article/pii/S259026282200017X. 17

Fabienne Bejjani, Claire Tolza, Mathias Boulanger, Damien Downes, Raphaël Romero, Muhammad Ahmad Maqbool, Amal Zine El Aabidine, Jean-Christophe Andrau, Sophie Lebre, Laurent Brehelin, Hughes Parrinello, Marine Rohmer, Tony Kaoma, Laurent Vallar, Jim R. Hughes, Kazem Zibara, Charles-Henri Lecellier, Marc Piechaczyk, and Isabelle Jariel-Encontre. Fra-1 regulates its target genes via binding to remote enhancers without exerting major control on chromatin architecture in triple negative breast cancers. Nucleic Acids Research, 49(5) : 2488–2508, March 2021. ISSN 1362-4962. doi : 10.1093/nar/gkab053. 30

Vaishak Belle and Ioannis Papantonis. Principles and Practice of Explainable Machine Learning. Frontiers in Big Data, 4, 2021. ISSN 2624-909X. URL https://www.frontiersin.org/article/10.3389/fdata.2021.688969. 9, 33

Nazih Benoumechiara. Treatment of dependency in sensitivity analysis for industrial reliability - Archive ouverte HAL, 2023. URL https://theses.hal.science/tel-02936431/. 35

Mette Bentsen, Vanessa Heger, Hendrik Schultheis, Carsten Kuenne, and Mario Looso. TF-COMB – Discovering grammar of transcription factor binding sites. Computational and Structural Biotechnology Journal, 20 :4040–4051, July 2022. ISSN 2001-0370. doi : 10.1016/j.csbj.2022.07.025. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9358416/. 40

Ivan Berest, Christian Arnold, Armando Reyes-Palomares, Giovanni Palla, Kasper Dindler Rasmussen, Holly Giles, Peter-Martin Bruch, Wolfgang Huber, Sascha Dietrich, Kristian Helin, and Judith B. Zaugg. Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors : diffTF. Cell Reports, 29(10) : 3147–3159.e12, December 2019. ISSN 2211-1247. doi : 10.1016/j.celrep.2019.10.106. URL https://www.sciencedirect.com/science/article/pii/S2211124719314391. 40

Linn Cecilie Bergersen, Ingrid K. Glad, and Heidi Lyng. Weighted lasso with data integration. Statistical Applications in Genetics and Molecular Biology, 10(1) :/j/sagmb.2011.10.issue–1/sagmb.2011.10.1.1703/sagmb.2011.10.1.1703.xml, August 2011. ISSN 1544-6115. doi : 10.2202/1544-6115.1703. 22

Allister Bernard and Alexander J. Hartemink. Informative structure priors : joint learning of dynamic regulatory networks from multiple types of data. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pages 459–470, 2005. ISSN 2335-6928. 21

Chloé Bessière, May Taha, Florent Petitprez, Jimmy Vandel, Jean-Michel Marin, Laurent Bréhélin, Sophie Lèbre, and Charles-Henri Lecellier. Probing instructions for expression regulation in gene nucleotide compositions. PLOS Computational Biology, 14(1) :e1005921, January 2018. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1005921. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005921. 26, 27, 29, 34, 40

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A LASSO FOR HIERARCHICAL INTERACTIONS. Annals of statistics, 41(3) :1111–1141, June 2013. ISSN 0090-5364. doi : 10.1214/13-AOS1096. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4527358/. 29

Barbara Bodinier, Sarah Filippi, Therese Haugdahl Nost, Julien Chiquet, and Marc Chadeau-Hyam. Automated calibration for stability selection in penalised regression and graphical models, February 2023. URL http://arxiv.org/abs/2106.02521. arXiv :2106.02521 [stat]. 33, 41

Valentina Boeva. Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. Frontiers in Genetics, 7, February 2016. ISSN 1664-8021. doi : 10.3389/fgene.2016.00024. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4763482/. 27

Richard Bonneau, David J. Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S. Baliga, and Vesteinn Thorsson. The Inferelator : an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biology, 7(5) :R36, May 2006. ISSN 1474-760X. doi : 10.1186/gb-2006-7-5-r36. URL https://doi.org/10.1186/gb-2006-7-5-r36. 16, 18

Anne-Laure Boulesteix and Sabine Hoffmann. To adjust or not to adjust : It is not the tests you perform that count, but how you report them, July 2022. URL https://osf.io/preprints/metaarxiv/j986q/. 9

Anne-Laure Boulesteix, Andreas Bender, Justo Lorenzo Bermejo, and Carolin Strobl. Random forest Gini importance favours SNPs with large minor allele frequency : impact, sources and recommendations. Briefings in Bioinformatics, 13(3) :292–304, May 2012. ISSN 1467-5463. doi : 10.1093/bib/bbr053. URL https://doi.org/10.1093/bib/bbr053. 35

Faustine Bousquet. Modélisation de la réponse utilisateur à une campagne de publicité mobile. These de doctorat, Montpellier, December 2020. URL https://www.theses.fr/2020MONTS095. 117

Leo Breiman. Random Forests. Machine Learning, 45(1) :5–32, October 2001. ISSN 1573-0565. doi : 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324. 15, 34

Leo Breiman. Setting up, using, and understanding random forests breiman technical report v3.1. Technical report, UC Berkeley, Department of Statistics., 2002. 34

Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. Cart. Classification and regression trees, 1984. Publisher : Wadsworth and Brooks/Cole Monterey, CA, USA. 15

A. J. Butte and I. S. Kohane. Mutual information relevance networks : functional genomic clustering using pairwise entropy measurements. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pages 418–429, 2000. ISSN 2335-6928. doi : 10.1142/9789814447331_0040. 11

Clément Bénard, Gérard Biau, Sébastien da Veiga, and Erwan Scornet. Interpretable Random Forests via Rule Extraction. Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, 130 :937, April 2021a. URL https://hal.sorbonne-universite.fr/hal-02557113. 35, 41

Clément Bénard, Gérard Biau, Sébastien da Veiga, and Erwan Scornet. SIRUS : Stable and Interpretable RUle Set for Classification. Electronic Journal of Statistics, 15 :427, January 2021b. URL https://hal.science/hal-02190689. 35, 41

Clément Bénard, Sébastien Da Veiga, and Erwan Scornet. Mean decrease accuracy for random forests : inconsistency, and a practical solution via the Sobol-MDA. Biometrika, 109(4) :881–900, December 2022. ISSN 1464-3510. doi : 10.1093/biomet/asac017. URL https://doi.org/10.1093/biomet/asac017. 34, 35, 41

Xiaodong Cai, Juan Andrés Bazerque, and Georgios B. Giannakis. Inference of Gene Regulatory Networks with Sparse Structural Equation Models Exploiting Genetic Perturbations. PLOS Computational Biology, 9(5) :e1003068, May 2013. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1003068. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003068. Publisher : Public Library of Science. 20

Jiguo Cao, Xin Qi, and Hongyu Zhao. Modeling Gene Regulation Networks Using Ordinary Differential Equations. In Junbai Wang, Aik Choon Tan, and Tianhai Tian, editors, Next Generation Microarray Bioinformatics : Methods and Protocols, Methods in Molecular Biology, pages 185–197. Humana Press, Totowa, NJ, 2012. ISBN 978-1-61779-400-1. doi : 10.1007/978-1-61779-400-1_12. URL https://doi.org/10.1007/978-1-61779-400-1_12. 12

Océane Cassan. Inférence statistique des réseaux de régulation de gènes chez Arabidopsis thaliana en réponse à l'élévation des teneurs en CO2 atmosphérique, December 2022. URL http://www.theses.fr/s226735. 23, 37, 39, 40

Océane Cassan, Sophie Lèbre, and Antoine Martin. Inferring and analyzing gene regulatory networks from multi-factorial expression data : a complete and interactive suite. BMC Genomics, 22(1) :387, May 2021. ISSN 1471-2164. doi : 10.1186/s12864-021-07659-2. URL https://doi.org/10.1186/s12864-021-07659-2. 16, 17, 36

Océane Cassan, Léa-Lou Pimparé, Christian Dubos, Alain Gojon, Liên Bach, Sophie Lèbre, and Antoine Martin. A gene regulatory network in Arabidopsis roots reveals features and regulators of the plant response to elevated CO2. New Phytologist, 2023. ISSN 1469-8137. doi : 10.1111/nph.18788. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.18788. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/nph.18788. 17

R. Castelo and A. Roverato. Graphical model search procedure in the large p and small n paradigm with applications to microarray data. Journal of Machine Learning Research, 7 :2621–2650, 2006. 11

Dayanne M. Castro, Nicholas R. de Veaux, Emily R. Miraldi, and Richard Bonneau. Multi-study inference of regulatory networks for more accurate models of gene regulation, January 2019. URL https://www.biorxiv.org/content/10.1101/279224v3. Pages : 279224 Section : New Results. 36

Jaime A Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, Oriol Fornes, Tiffany Y Leung, Alejandro Aguirre, Fayrouz Hammal, Daniel Schmelter, Damir Baranasic, Benoit Ballester, Albin Sandelin, Boris Lenhard, Klaas Vandepoele, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2022 : the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Research, 50(D1) :D165–D173, January 2022. ISSN 0305-1048. doi : 10.1093/nar/gkab1113. URL https://doi.org/10.1093/nar/gkab1113. 19

Camille Charbonnier, Julien Chiquet, and Christophe Ambroise. Weighted-Lasso for Structured Network Inference from Time Course Data. Statistical Applications in Genetics and Molecular Biology, 9(1), January 2010. ISSN 1544-6115. doi : 10.2202/1544-6115.1519. URL http://arxiv.org/abs/0910.1723. arXiv :0910.1723 [stat]. 18, 23

Hugh Chipman. Bayesian variable selection with related predictors. Canadian Journal of Statistics, 24(1) :17–36, 1996. ISSN 1708-945X. doi : 10.2307/3315687. URL https://onlinelibrary.wiley.com/doi/abs/10.2307/3315687. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.2307/3315687. 29

Julien Chiquet, Yves Grandvalet, and Christophe Ambroise. Inferring multiple graphical structures. Statistics and Computing, 21(4) :537–553, October 2011. ISSN 1573-1375. doi : 10.1007/s11222-010-9191-2. URL https://doi.org/10.1007/s11222-010-9191-2. 18

Julien Chiquet, Stephane Robin, and Mahendra Mariadassou. Variational Inference for sparse network reconstruction from count data. In International Conference on Machine Learning, pages 1162–1171, May 2019. URL http://proceedings.mlr.press/v97/chiquet19a.html. 12

Yongjun Chu and David R. Corey. RNA Sequencing : Platform Selection, Experimental Design, and Data Interpretation. Nucleic Acid Therapeutics, 22(4) :271–274, August 2012. ISSN 2159-3337. doi : 10.1089/nat.2012.0367. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3426205/. 9

Jacopo Cirrone, Matthew D. Brooks, Richard Bonneau, Gloria M. Coruzzi, and Dennis E. Shasha. OutPredict : multiple datasets can improve prediction of expression and inference of causality. Scientific Reports, 10(1) :6804, April 2020. ISSN 2045-2322. doi : 10.1038/

s41598-020-63347-3. URL https://www.nature.com/articles/s41598-020-63347-3. Bandiera_abtest : a Cc_license_type : cc_by Cg_type : Nature Research Journals Number : 1 Primary_atype : Research Publisher : Nature Publishing Group Subject_term : Computer science ;Machine learning Subject_term_id : computer-science ;machine-learning. 14, 15, 18, 23

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. J. Mach. Learn. Res., 15(1) :3741–3782, 2014. 13

J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. Statistics and Computing, 18(2) :173–183, June 2008. ISSN 1573-1375. doi : 10.1007/s11222-007-9046-7. URL https://doi.org/10.1007/s11222-007-9046-7. 18

Inge De Clercq, Jan Van de Velde, Xiaopeng Luo, Li Liu, Veronique Storme, Michiel Van Bel, Robin Pottie, Dries Vaneechoutte, Frank Van Breusegem, and Klaas Vandepoele. Integrative inference of transcriptional networks in Arabidopsis yields novel ROS signalling regulators. Nature Plants, 7(4) :500–513, April 2021. ISSN 2055-0278. doi : 10.1038/s41477-021-00894-1. URL https://www.nature.com/articles/s41477-021-00894-1. Number : 4 Publisher : Nature Publishing Group. 20

Houtao Deng. Interpreting Tree Ensembles with inTrees, August 2014. URL http://arxiv.org/abs/1408.5456. arXiv :1408.5456 [cs, stat]. 30

P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mRNA expression levels during CNS development and injury. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pages 41–52, 1999. ISSN 2335-6928. doi : 10.1142/9789814447300_0005. 12

Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Nadav S. Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J. Luo, Eddie Park, Kimberly Persaud, Jonathan B. Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, and Thomas R. Gingeras. Landscape of transcription in human cells. Nature, 489(7414) :101–108, September 2012. ISSN 1476-4687. doi : 10.1038/nature11233. URL https://www.nature.com/articles/nature11233. Number : 7414 Publisher : Nature Publishing Group. 9, 19, 25

Adrian Dobra, Chris Hans, Beatrix Jones, Joseph R. Nevins, Guang Yao, and Mike West. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis, 90(1) : 196–212, July 2004. ISSN 0047-259X. doi : 10.1016/j.jmva.2004.02.009. URL https://www.sciencedirect.com/science/article/pii/S0047259X04000259. 11

Frank Dondelinger, Dirk Husmeier, and Sophie Lèbre. Dynamic Bayesian networks in molecular plant science : inferring gene regulatory networks from multiple gene expression time series. Euphytica, 183(3) :361–377, February 2012. ISSN 1573-5060. doi : 10.1007/s10681-011-0538-3. URL https://doi.org/10.1007/s10681-011-0538-3. 18

Frank Dondelinger, Sophie Lèbre, and Dirk Husmeier. Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. Machine Learning, 90(2) :191–230, February 2013. ISSN 1573-0565. doi : 10.1007/s10994-012-5311-x. URL https://doi.org/10.1007/s10994-012-5311-x. 13, 18

Mathias Drton and Michael D. Perlman. Multiple testing and error control in Gaussian graphical model selection. 2007. 11

Mathias Drton and Michael D. Perlman. A SINful approach to Gaussian graphical model selection. Journal of Statistical Planning and Inference, 138(4) :1179–1200, 2008. Publisher : Elsevier. 11

A. d'Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. Lanckriet. A direct formulation for sparse pca using semidefinite programming, in 'Advances in Neural Information Processing Systems', 2005. 11

Patrik D'haeseleer, Shoudan Liang, and Roland Somogyi. Genetic network inference : from co-expression clustering to reverse engineering. Bioinformatics, 16(8) :707–726, August 2000. ISSN 1367-4803. doi : 10.1093/bioinformatics/16.8.707. URL https://doi.org/10.1093/bioinformatics/16.8.707. 21

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least Angle Regression. The Annals of Statistics, 32(2) :407–451, 2004. ISSN 0090-5364. URL https://www.jstor.org/stable/3448465. Publisher : Institute of Mathematical Statistics. 15

Byron Ellis and Wing Hung Wong. Learning Causal Bayesian Network Structures from Experimental Data. Journal of the American Statistical Association, 103(482) :778–789, 2008. ISSN 0162-1459. URL https://www.jstor.org/stable/27640100. Publisher : [American Statistical Association, Taylor & Francis, Ltd.]. 12

Rossin Erbe, Jessica Gore, Kelly Gemmill, Daria A. Gaykalova, and Elana J. Fertig. The use of machine learning to discover regulatory networks controlling biological systems. Molecular Cell, 82(2) :260–273, January 2022. ISSN 1097-2765. doi : 10.1016/j.molcel.2021.12.011. URL https://www.sciencedirect.com/science/article/pii/S109727652101073X. 9

Jason Ernst and Manolis Kellis. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. Genome Research, 23(7) :1142–1154, July 2013. ISSN 1088-9051. doi : 10.1101/gr.144840.112. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3698507/. 19

Jeremiah J. Faith, Boris Hayete, Joshua T. Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J. Collins, and Timothy S. Gardner. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biology, 5(1) :e8, January 2007. ISSN 1545-7885. doi : 10.1371/journal.pbio.0050008. 11

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR : A library for large linear classification. the Journal of machine Learning research, 9 :1871–1874, 2008. Publisher : JMLR. org. 29

Rina Foygel and Mathias Drton. Extended Bayesian Information Criteria for Gaussian Graphical Models. In Advances in Neural Information Processing Systems, volume 23. Curran Associates, Inc., 2010. URL https://papers.nips.cc/paper/2010/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html. 22

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. The Annals of Applied Statistics, 1(2), December 2007. ISSN 1932-6157. doi : 10.1214/07-AOAS131. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-1/issue-2/Pathwise-coordinate-optimization/10.1214/07-AOAS131.full. 29

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. Biostatistics (Oxford, England), 9(3) :432–441, July 2008. ISSN 1468-4357. doi : 10.1093/biostatistics/kxm045. 12

Jerome H. Friedman. Greedy function approximation : A gradient boosting machine. Ann. Statist., 29(2) :1189–1232, 2001. URL http://dml.mathdoc.fr/item/1013203451/. 22

Jerome H. Friedman. Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4) :367–378, February 2002. ISSN 0167-9473. doi : 10.1016/S0167-9473(01)00065-2. URL https://www.sciencedirect.com/science/article/pii/S0167947301000652. 15

Nir Friedman and Daphne Koller. Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. Machine Learning, 50(1) :95–125, January 2003. ISSN 1573-0565. doi : 10.1023/A:1020249912095. URL https://doi.org/10.1023/A:1020249912095. 12

Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, UAI'98, pages 139–147, San Francisco, CA, USA, July 1998. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-555-8. 12

Daniel J. Gaffney. Global Properties and Functional Complexity of Human Gene Regulatory Variation. PLOS Genetics, 9(5) :e1003501, May 2013. ISSN 1553-7404. doi : 10.1371/journal.pgen.1003501. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003501. Publisher : Public Library of Science. 19

Mélina Gallopin, Andrea Rau, and Florence Jaffrézic. A Hierarchical Poisson Log-Normal Model for Network Inference from RNA Sequencing Data. PLOS ONE, 8(10) :e77503, October 2013. ISSN 1932-6203. doi : 10.1371/journal.pone.0077503. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077503. Publisher : Public Library of Science. 12

Yoav Gilad, Scott A. Rifkin, and Jonathan K. Pritchard. Revealing the architecture of gene regulation : the promise of eQTL studies. Trends in genetics : TIG, 24(8) :408–415, August 2008. ISSN 0168-9525. doi : 10.1016/j.tig.2008.06.001. 19

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. Frontiers in Genetics, 10 :524, 2019. ISSN 1664-8021. doi : 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/article/10.3389/fgene.2019.00524. 13

Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar, and Stein Aerts. SCENIC+ : single-cell multiomic inference of enhancers and gene regulatory networks, August 2022. URL https://www.biorxiv.org/content/10.1101/2022.08.19.504505v1. Pages : 2022.08.19.504505 Section : New Results. 15, 16, 21

Mathys Grapotte, Manu Saraswat, Chloé Bessière, Christophe Menichelli, Jordan A. Ramilowski, Jessica Severin, Yoshihide Hayashizaki, Masayoshi Itoh, Michihira Tagami, Mitsuyoshi Murata, Miki Kojima-Ishiyama, Shohei Noma, Shuhei Noguchi, Takeya Kasukawa, Akira Hasegawa, Harukazu Suzuki, Hiromi Nishiyori-Sueki, Martin C. Frith, Clément Chatelain, Piero Carninci, Michiel J. L. de Hoon, Wyeth W. Wasserman, Laurent Bréhélin, and Charles-Henri Lecellier. Discovery of widespread transcription initiation at microsatellites predictable by sequence-based deep neural network. Nature Communications, 12(1) :3297, June 2021. ISSN 2041-1723. doi : 10.1038/s41467-021-23143-7. URL https://www.nature.com/articles/s41467-021-23143-7. Number : 1 Publisher : Nature Publishing Group. 26

Alex Greenfield, Christoph Hafemeister, and Richard Bonneau. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics (Oxford, England), 29(8) :1060–1067, April 2013. ISSN 1367-4811. doi : 10.1093/bioinformatics/btt099. 15, 16, 21, 22, 23, 36, 37

Marco Grzegorczyk and Dirk Husmeier. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. Machine Learning, 71(2) :265–305, June 2008. ISSN 1573-0565. doi : 10.1007/s10994-008-5057-7. URL https://doi.org/10.1007/s10994-008-5057-7. 12

Shun Guo, Qingshan Jiang, Lifei Chen, and Donghui Guo. Gene regulatory network inference using PLS-based methods. BMC Bioinformatics, 17(1) :545, December 2016. ISSN 1471-2105. doi : 10.1186/s12859-016-1398-6. URL https://doi.org/10.1186/s12859-016-1398-6. 15

Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE : The reference human genome annotation for The ENCODE Project. Genome Research, 22(9) : 1760–1774, January 2012. ISSN 1088-9051, 1549-5469. doi : 10.1101/gr.135350.111. URL https://genome.cshlp.org/content/22/9/1760. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab. 28

Alexander J. Hartemink, David K. Gifford, Tommi S. Jaakkola, and Richard A. Young. Combining location and expression data for principled discovery of genetic regulatory network models. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pages 437–449, 2002. ISSN 2335-6928. 21

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best Subset, Forward Stepwise or Lasso ? Analysis and Recommendations Based on Extensive Comparisons. Statistical Science, 35(4) :579–592, November 2020. ISSN 0883-4237, 2168-8745. doi : 10.1214/19-STS733. URL https://projecteuclid.org/journals/statistical-science/volume-35/issue-4/Best-Subset-Forward-Stepwise-or-Lasso-Analysis-and-Recommendations-Based/10.1214/19-STS733.full. Publisher : Institute of Mathematical Statistics. 9, 33

Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. TIGRESS : Trustful Inference of Gene REgulation using Stability Selection. BMC Systems Biology, 6(1) : 145, November 2012. ISSN 1752-0509. doi : 10.1186/1752-0509-6-145. URL https://doi.org/10.1186/1752-0509-6-145. 15, 33, 36

Łukasz Huminiecki and Jarosław Horbańczuk. Can We Predict Gene Expression by Understanding Proximal Promoter Architecture ? Trends in Biotechnology, 38(4) :463, April 2020. ISSN 0167-7799, 1879-3096. doi : 10.1016/j.tibtech.2019.12.003. URL https://www.cell.com/trends/biotechnology/abstract/S0167-7799(19)30298-7. Publisher : Elsevier. 19

Vân Anh Huynh-Thu and Pierre Geurts. dynGENIE3 : dynamical GENIE3 for the inference of gene networks from time series expression data. Scientific Reports, 8(1) :3384, February 2018. ISSN 2045-2322. doi : 10.1038/s41598-018-21715-0. URL https://www.nature.com/articles/s41598-018-21715-0. Number : 1 Publisher : Nature Publishing Group. 12, 14, 15, 18, 23, 36

Vân Anh Huynh-Thu and Guido Sanguinetti. Combining tree-based and dynamical systems for the inference of gene regulatory networks. Bioinformatics (Oxford, England), 31(10) :1614–1622, May 2015. ISSN 1367-4811. doi : 10.1093/bioinformatics/btu863. 14, 15

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. PloS One, 5(9) :e12776, September 2010. ISSN 1932-6203. doi : 10.1371/journal.pone.0012776. 15, 17, 36

Vân Anh Huynh-Thu, Louis Wehenkel, and Pierre Geurts. Gene Regulatory Network Inference from Systems Genetics Data Using Tree-Based Methods. In Alberto de la Fuente, editor, Gene Network Inference : Verification of Methods for Systems Genetics Data, pages 63–85. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-45161-4. doi : 10.1007/978-3-642-45161-4_5. URL https://doi.org/10.1007/978-3-642-45161-4_5. 15

Ignacio L. Ibarra, Nele M. Hollmann, Bernd Klaus, Sandra Augsten, Britta Velten, Janosch Hennig, and Judith B. Zaugg. Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. Nature Communications, 11(1) :124, January 2020. ISSN 2041-1723. doi : 10.1038/s41467-019-13888-7. URL https://www.nature.com/articles/s41467-019-13888-7. Number : 1 Publisher : Nature Publishing Group. 29

S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. In Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003, pages 104–113, August 2003. doi : 10.1109/CSB.2003.1227309. 21

H. Ishwaran and U.B. Kogalur. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), 2023. URL https://cran.r-project.org/package=randomForestSRC. R package version 3.2.0. 34

Granton A. Jindal and Emma K. Farley. Enhancer grammar in development, evolution, and disease : dependencies and interplay. Developmental Cell, 56(5) :575–587, March 2021. ISSN 1878-1551. doi : 10.1016/j.devcel.2021.02.016. 29

Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-binding specificities of human transcription factors. Cell, 152(1-2) :327–339, January 2013. ISSN 1097-4172. doi : 10.1016/j.cell.2012.12.009. 19

Beatrix Jones, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West. Experiments in Stochastic Computation for High-Dimensional Graphical Models. Statistical Science, 20(4) :388–400, November 2005. ISSN 0883-4237, 2168-8745. doi : 10.1214/088342305000000304. URL https://projecteuclid.org/journals/statistical-science/volume-20/issue-4/Experiments-in-Stochastic-Computation-for-High-Dimensional-Graphical-Models/10.1214/088342305000000304.full. Publisher : Institute of Mathematical Statistics. 11

Kenji Kamimoto, Christy M. Hoffmann, and Samantha A. Morris. CellOracle : Dissecting cell identity via network inference and in silico gene perturbation, April 2020. URL https://www.biorxiv.org/content/10.1101/2020.02.17.947416v3. Pages : 2020.02.17.947416 Section : New Results. 21

Kenji Kamimoto, Blerta Stringa, Christy M. Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A. Morris. Dissecting cell identity via network inference and in silico gene perturbation. Nature, 614(7949) :742–751, February 2023. ISSN 1476-4687. doi : 10.1038/s41586-022-05688-9. 16

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research, 44(D1) :D457–D462, January 2016. ISSN 0305-1048. doi : 10.1093/nar/gkv1070. URL https://doi.org/10.1093/nar/gkv1070. 17

David R. Kelley, Yakir Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Research, page gr.227819.117, March 2018. ISSN 1088-9051, 1549-5469. doi : 10.1101/gr.227819.117. URL http://genome.cshlp.org/content/early/2018/03/27/gr.227819.117. 26

Dong-Chul Kim, Mingon Kang, Baoju Zhang, Xiaoyong Wu, Chunyu Liu, and Jean Gao. Integration of DNA Methylation, Copy Number Variation, and Gene Expression for Gene Regulatory Network Inference and Application to Psychiatric Disorders. In 2014 IEEE International Conference on Bioinformatics and Bioengineering, pages 238–242, November 2014. doi : 10.1109/BIBE.2014.71. 20

Sun Yong Kim, Seiya Imoto, and Satoru Miyano. Inferring gene networks from time series microarray data using dynamic Bayesian networks. Briefings in Bioinformatics, 4(3) :228–235, September 2003. ISSN 1467-5463. doi : 10.1093/bib/4.3.228. 12

Youngsoon Kim, Jie Hao, Yadu Gautam, Tesfaye B. Mersha, and Mingon Kang. DiffGRN : differential gene regulatory network analysis. International journal of data mining and bioinformatics, 20(4) :362–379, 2018. ISSN 1748-5673. doi : 10.1504/IJDMB.2018.094891. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6526019/. 40

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. The Annals of Statistics, 28(5) :1356–1378, October 2000. ISSN 0090-5364, 2168-8966. doi : 10.1214/aos/1015957397. URL https://projecteuclid.org/euclid.aos/1015957397. 33

Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan S. Yevshin, Anastasiia V. Soboleva, Artem S. Kasianov, Haitham Ashoor, Wail Ba-alawi, Vladimir B. Bajic, Yulia A. Medvedeva, Fedor A. Kolpakov, and Vsevolod J. Makeev. HOCOMOCO : expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Research, 44(D1) :D116–D125, January 2016. ISSN 0305-1048. doi : 10.1093/nar/gkv1249. URL https://doi.org/10.1093/nar/gkv1249. 19, 30

Alexander Lachmann, Federico M. Giorgi, Gonzalo Lopez, and Andrea Califano. ARACNe-AP : gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics, 32(14) :2233–2235, July 2016. ISSN 1367-4803. doi : 10.1093/bioinformatics/btw216. URL https://doi.org/10.1093/bioinformatics/btw216. 11

Peter Langfelder and Steve Horvath. WGCNA : an R package for weighted correlation network analysis. BMC Bioinformatics, 9(1) :559, December 2008. ISSN 1471-2105. doi : 10.1186/1471-2105-9-559. URL https://doi.org/10.1186/1471-2105-9-559. 11

Steffen L. Lauritzen and Steffen L. Lauritzen. Graphical Models. Oxford Statistical Science Series. Oxford University Press, Oxford, New York, May 1996. ISBN 978-0-19-852219-5. 11, 12

Julien Lavenus and Mikaël Lucas. How to Use the TDCor Algorithm to Infer Gene Regulatory Networks from Time Series Transcriptomic Data. Methods in Molecular Biology (Clifton, N.J.), 2395 :13–31, 2022. ISSN 1940-6029. doi : 10.1007/978-1-0716-1816-5_2. 13

Tamara Le Thanh, Bárbara Hufnagel, Alexandre Soriano, Fanchon Divol, Laurent Brottier, Célia Casset, Benjamin Péret, Patrick Doumas, and Laurence Marquès. Dynamic Development of White Lupin Rootlets Along a Cluster Root. Frontiers in Plant Science, 12, 2021. ISSN 1664-462X. URL https://www.frontiersin.org/articles/10.3389/fpls.2021.738172. 17

Jean-Benoist Leger. Blockmodels : A R-package for estimating in Latent Block Model and Stochastic Block Model, with various probability functions, with or without covariates, February 2016. URL http://arxiv.org/abs/1602.07587. arXiv :1602.07587 [stat]. 17

Michal Levo, Einat Zalckvar, Eilon Sharon, Ana Carolina Dantas Machado, Yael Kalma, Maya Lotam-Pompan, Adina Weinberger, Zohar Yakhini, Remo Rohs, and Eran Segal. Unraveling determinants of transcription factor binding outside the core binding site. Genome Research, 25(7) :1018–1029, July 2015. ISSN 1549-5469. doi : 10.1101/gr.185033.114. 29

Jinsen Li, Jared M. Sagendorf, Tsu-Pei Chiu, Marco Pasi, Alberto Perez, and Remo Rohs. Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. Nucleic Acids Research, 45(22) :12877–12887, December 2017. ISSN 1362-4962. doi : 10.1093/nar/gkx1145. 30

Xiao Li, Yu Wang, Sumanta Basu, Karl Kumbier, and Bin Yu. A Debiased MDI Feature Importance Measure for Random Forests. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/702cafa3bb4c9c86e4a3b6834b45aedd-Abstract.html. 35

Yue Li, Minggao Liang, and Zhaolei Zhang. Regression Analysis of Combined Gene Expression Regulation in Acute Myeloid Leukemia. PLOS Computational Biology, 10(10) :e1003908, October 2014. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1003908. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003908. Publisher : Public Library of Science. 20, 26, 27, 34

Yupeng Li and Scott A. Jackson. Gene Network Reconstruction by Integration of Prior Biological Knowledge. G3 (Bethesda, Md.), 5(6) :1075–1079, March 2015. ISSN 2160-1836. doi : 10.1534/g3.115.018127. 22

Zhongxiao Li, Elva Gao, Juexiao Zhou, Wenkai Han, Xiaopeng Xu, and Xin Gao. Applications of deep learning in understanding gene regulation. Cell Reports Methods, 3(1) :100384, January 2023. ISSN 2667-2375. doi : 10.1016/j.crmeth.2022.100384. URL https://www.sciencedirect.com/science/article/pii/S2667237522002892. 27

Andy Liaw and Matthew Wiener. Classification and regression by randomForest. R news, 2(3) : 18–22, 2002. 34

Michael Lim and Trevor Hastie. Learning Interactions via Hierarchical Group-Lasso Regularization. Journal of Computational and Graphical Statistics, 24(3) :627–654, July 2015. ISSN 1061-8600. doi : 10.1080/10618600.2014.938812. URL https://doi.org/10.1080/10618600.2014.938812. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/10618600.2014.938812. 29

Camilla Lingjærde, Tonje G. Lien, Ørnulf Borgan, Helga Bergholtz, and Ingrid K. Glad. Tailored graphical lasso for data integration in gene network reconstruction. BMC Bioinformatics, 22 (1) :498, October 2021. ISSN 1471-2105. doi : 10.1186/s12859-021-04413-z. URL https://doi.org/10.1186/s12859-021-04413-z. 22, 23

Han Liu, Kathryn Roeder, and Larry Wasserman. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models, June 2010. URL http://arxiv.org/abs/1006.3316. arXiv :1006.3316 [stat]. 22, 33

Li-Zhi Liu, Fang-Xiang Wu, and Wen-Jun Zhang. A group LASSO-based method for robustly inferring gene regulatory networks from multiple time-course datasets. BMC systems biology, 8 Suppl 3(Suppl 3) :S1, 2014. ISSN 1752-0509. doi : 10.1186/1752-0509-8-S3-S1. 18, 34

Markus Loecher. Unbiased variable importance for random forests. Communications in Statistics - Theory and Methods, 51(5) :1413–1425, March 2022. ISSN 0361-0926. doi : 10.1080/03610926.2020.1764042. URL https://doi.org/10.1080/03610926.2020.1764042. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/03610926.2020.1764042. 35

Miguel Lopes and Gianluca Bontempi. Experimental assessment of static and dynamic algorithms for gene regulation inference from time series expression data. Frontiers in Genetics, 4 :303, December 2013. ISSN 1664-8021. doi : 10.3389/fgene.2013.00303. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3872039/. 12

Miguel Lopes, Patrick Meyer, and Gianluca Bontempi. Estimation of temporal lags for the inference of gene regulatory networks from time series (inproceedings) Author. In BENELEARN'12, 2012. 12

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12) :550, December 2014. ISSN 1474-760X. doi : 10.1186/s13059-014-0550-8. URL https://doi.org/10.1186/s13059-014-0550-8. 17

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html. 27

Yunhai Luo, Benjamin C. Hitz, Idan Gabdank, Jason A. Hilton, Meenakshi S. Kagda, Bonita Lam, Zachary Myers, Paul Sud, Jennifer Jou, Khine Lin, Ulugbek K. Baymuradov, Keenan Graham, Casey Litton, Stuart R. Miyasato, J. Seth Strattan, Otto Jolanki, Jin-Wook Lee, Forrest Y. Tanaka, Philip Adenekan, Emma O'Neill, and J. Michael Cherry. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Research, 48(D1) : D882–D889, January 2020. ISSN 1362-4962. doi : 10.1093/nar/gkz1062. 30

Sophie Lèbre, Jennifer Becq, Frédéric Devaux, Michael PH Stumpf, and Gaëlle Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. BMC Systems Biology, 4(1) :130, September 2010. ISSN 1752-0509. doi : 10.1186/1752-0509-4-130. URL https://doi.org/10.1186/1752-0509-4-130. 13

D. A. K. Maduranga, Jie Zheng, Piyushkumar A. Mundra, and Jagath C. Rajapakse. Inferring Gene Regulatory Networks from Time-Series Expressions Using Random Forests Ensemble. In Alioune Ngom, Enrico Formenti, Jin-Kao Hao, Xing-Ming Zhao, and Twan van Laarhoven, editors, Pattern Recognition in Bioinformatics, Lecture Notes in Computer Science, pages 13–22, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-39159-0. doi : 10.1007/978-3-642-39159-0_2. 15

Thierry A. Mara, Stefano Tarantola, and Paola Annoni. Non-parametric methods for global sensitivity analysis of model output with dependent inputs. Environmental Modelling & Software, 72 :173–183, October 2015. ISSN 13648152. doi : 10.1016/j.envsoft.2015.07.010. URL https://linkinghub.elsevier.com/retrieve/pii/S1364815215300153. 35

Daniel Marbach, James C. Costello, Robert Küffner, Nicole M. Vega, Robert J. Prill, Diogo M. Camacho, Kyle R. Allison, Manolis Kellis, James J. Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. Nature Methods, 9(8) :796–804, August 2012a. ISSN 1548-7105. doi : 10.1038/nmeth.2016. URL https://www.nature.com/articles/nmeth.2016. Number : 8 Publisher : Nature Publishing Group. 15, 18

Daniel Marbach, Sushmita Roy, Ferhat Ay, Patrick E. Meyer, Rogerio Candeias, Tamer Kahveci, Christopher A. Bristow, and Manolis Kellis. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. Genome Research, 22(7) :1334–1349, July 2012b. ISSN 1549-5469. doi : 10.1101/gr.127191.111. 20

Elaine R. Mardis. DNA sequencing technologies : 2006–2016. Nature Protocols, 12(2) :213–218, February 2017. ISSN 1750-2799. doi : 10.1038/nprot.2016.182. URL https://www.nature.com/articles/nprot.2016.182. Number : 2 Publisher : Nature Publishing Group. 9

Adam A. Margolin, Kai Wang, Wei Keat Lim, Manjunath Kustagi, Ilya Nemenman, and Andrea Califano. Reverse engineering cellular networks. Nature Protocols, 1(2) :662–671, August 2006. ISSN 1750-2799. doi : 10.1038/nprot.2006.106. URL https://www.nature.com/articles/nprot.2006.106. Number : 2 Publisher : Nature Publishing Group. 11

Jérôme Mariette and Nathalie Villa-Vialaneix. Unsupervised multiple kernel learning for heterogeneous data integration. Bioinformatics, 34(6) :1009–1015, March 2018. ISSN 1367-4803. doi : 10.1093/bioinformatics/btx682. URL https://doi.org/10.1093/bioinformatics/btx682. 9

Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. Celer : a fast solver for the lasso with dual extrapolation. In International Conference on Machine Learning, pages 3315–3324. PMLR, 2018. 29

Anthony Mathelier, Oriol Fornes, David J. Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, Allen W. Zhang, François Parcy, Boris Lenhard, Albin Sandelin, and Wyeth W. Wasserman. JASPAR 2016 : a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Research, 44(D1) :D110–D115, January 2016a. ISSN 0305-1048. doi : 10.1093/nar/gkv1176. URL https://doi.org/10.1093/nar/gkv1176. 30

Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W. Wasserman. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. Cell Systems, 3(3) :278–286.e4, September 2016b. ISSN 2405-4712. doi : 10.1016/j.cels.2016.07.001. 30

Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research, 40(10) :4288–4297, May 2012. ISSN 0305-1048. doi : 10.1093/nar/gks042. URL https://doi.org/10.1093/nar/gks042. 17

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. The Annals of Statistics, 34(3) :1436–1462, June 2006. ISSN 0090-5364, 2168-8966. doi : 10.1214/009053606000000281. URL https://projecteuclid.org/journals/annals-of-statistics/volume-34/issue-3/High-dimensional-graphs-and-variable-selection-with-the-Lasso/10.1214/009053606000000281.full. Publisher : Institute of Mathematical Statistics. 12, 14

Nicolai Meinshausen and Peter Bühlmann. Stability selection. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 72(4) :417–473, 2010. ISSN 1467-9868. doi : 10.1111/j.1467-9868.2010.00740.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00740.x. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2010.00740.x. 15, 33

Christophe Menichelli, Vincent Guitard, Rafael M. Martins, Sophie Lèbre, Jose-Juan Lopez-Rubio, Charles-Henri Lecellier, and Laurent Bréhélin. Identification of long regulatory elements in the genome of Plasmodium falciparum and other eukaryotes. PLOS Computational Biology, 17 (4) :e1008909, April 2021. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1008909. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008909. Publisher : Public Library of Science. 28, 31, 34, 40

Patrick E. Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. Information-theoretic inference of large transcriptional regulatory networks. EURASIP journal on bioinformatics & systems biology, 2007(1) :79879, 2007. ISSN 1687-4145. doi : 10.1155/2007/79879. 11

Patrick E. Meyer, Frédéric Lafitte, and Gianluca Bontempi. minet : A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. BMC Bioinformatics, 9(1) :461, October 2008. ISSN 1471-2105. doi : 10.1186/1471-2105-9-461. URL https://doi.org/10.1186/1471-2105-9-461. 11

Emily R. Miraldi, Maria Pokrovskii, Aaron Watters, Dayanne M. Castro, Nicholas De Veaux, Jason A. Hall, June-Yong Lee, Maria Ciofani, Aviv Madar, Nick Carriero, Dan R. Littman, and Richard Bonneau. Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. Genome Research, 29(3) :449–463, January 2019. ISSN 1088-9051, 1549-5469. doi : 10.1101/gr.238253.118. URL https://genome.cshlp.org/content/29/3/449. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab. 16, 36

Keiichi Mochida, Satoru Koda, Komaki Inoue, and Ryuei Nishii. Statistical and Machine Learning Approaches to Predict Gene Regulatory Networks From Transcriptome Datasets. Frontiers in Plant Science, 9, 2018. ISSN 1664-462X. doi : 10.3389/fpls.2018.01770. URL https://www.frontiersin.org/articles/10.3389/fpls.2018.01770/full. Publisher : Frontiers. 9

Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. GRNBoost2 and Arboreto : efficient and scalable inference of gene regulatory networks. Bioinformatics, 35(12) :2159–2161, June 2019. ISSN 1367-4803. doi : 10.1093/bioinformatics/bty916. URL https://doi.org/10.1093/bioinformatics/bty916. 15

Fanny Mollandin, Hélène Gilbert, Pascal Croiseau, and Andrea Rau. Accounting for overlapping annotations in genomic prediction models of complex traits. BMC Bioinformatics, 23(1) :365, September 2022. ISSN 1471-2105. doi : 10.1186/s12859-022-04914-5. URL https://doi.org/10.1186/s12859-022-04914-5. 9

Ekaterina Morgunova and Jussi Taipale. Structural perspective of cooperative transcription factor binding. Current Opinion in Structural Biology, 47 :1–8, December 2017. ISSN 0959-440X. doi : 10.1016/j.sbi.2017.03.006. URL http://www.sciencedirect.com/science/article/pii/S0959440X17300088. 29

Sach Mukherjee and Terence P. Speed. Network inference using informative priors. Proceedings of the National Academy of Sciences of the United States of America, 105(38) :14313–14318, September 2008. ISSN 0027-8424. doi : 10.1073/pnas.0802272105. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2567188/. 18

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44) :22071–22080, October 2019. doi : 10.1073/pnas.1900654116. URL https://www.pnas.org/doi/full/10.1073/pnas.1900654116. Publisher : Proceedings of the National Academy of Sciences. 9, 33

Kevin Murphy and Saira Mian. Modelling gene expression data using dynamic Bayesian networks. Technical report, Citeseer, 1999. 12

N. Nariai, S. Kim, S. Imoto, and S. Miyano. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pages 336–347, 2004. ISSN 2335-6928. doi : 10.1142/9789812704856_0032. 21

J. A. Nelder. A Reformulation of Linear Models. Journal of the Royal Statistical Society. Series A (General), 140(1) :48–77, 1977. ISSN 0035-9238. doi : 10.2307/2344517. URL https://www.jstor.org/stable/2344517. Publisher : [Royal Statistical Society, Wiley]. 29

Yu. Nesterov. Smooth minimization of non-smooth functions. Mathematical Programming, 103 (1) :127–152, May 2005. ISSN 1436-4646. doi : 10.1007/s10107-004-0552-5. URL https://doi.org/10.1007/s10107-004-0552-5. 12

Hung Nguyen, Duc Tran, Bang Tran, Bahadir Pehlivan, and Tin Nguyen. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. Briefings in Bioinformatics, 22(3) :bbaa190, May 2021. ISSN 1477-4054. doi : 10.1093/bib/bbaa190. 16

Kristin K. Nicodemus. Letter to the Editor : On the stability and ranking of predictors from random forest variable importance measures. Briefings in Bioinformatics, 12(4) :369–373, July 2011. ISSN 1467-5463. doi : 10.1093/bib/bbr016. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137934/. 35

Kristin K. Nicodemus and James D. Malley. Predictor correlation impacts machine learning algorithms : implications for genomic studies. Bioinformatics, 25(15) :1884–1890, August 2009. ISSN 1367-4803. doi : 10.1093/bioinformatics/btp331. URL https://doi.org/10.1093/bioinformatics/btp331. 35

Elad Noor, Sarah Cherkaoui, and Uwe Sauer. Biological insights through omics data integration. Current Opinion in Systems Biology, 15 :39–47, June 2019. ISSN 2452-3100. doi : 10.1016/j.coisb.2019.03.007. URL https://www.sciencedirect.com/science/article/pii/S2452310019300125. 19

Nooshin Omranian, Jeanne M. O. Eloundou-Mbebi, Bernd Mueller-Roeber, and Zoran Nikoloski. Gene regulatory network inference using fused LASSO on multiple data sets. Scientific Reports, 6 :20533, February 2016. ISSN 2045-2322. doi : 10.1038/srep20533. 18

Rainer Opgen-Rhein and Korbinian Strimmer. From correlation to causation networks : a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Systems Biology, 1 :37, August 2007. ISSN 1752-0509. doi : 10.1186/1752-0509-1-37. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1995222/. 13

Zhengqing Ouyang, Qing Zhou, and Wing Hung Wong. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America, 106(51) :21521–21526, December 2009. ISSN 1091-6490. doi : 10.1073/pnas.0904863106. 26

Judea Pearl. Probabilistic reasoning in intelligent systems : networks of plausible inference. Morgan kaufmann, 1988. 11, 12

Francesca Petralia, Pei Wang, Jialiang Yang, and Zhidong Tu. Integrative random forest for gene regulatory network inference. Bioinformatics, 31(12) :i197–i205, June 2015. ISSN 1367-4803. doi : 10.1093/bioinformatics/btv268. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4542785/. 15, 23, 36

Jing Qin, Yaohua Hu, Feng xu, Hari Yalamanchili, and Junwen Wang. Inferring Gene Regulatory Networks by Integrating ChIP-seq/chip and Transcriptome Data via LASSO-type Regularization Methods. Methods, June 2014. doi : 10.1016/j.ymeth.2014.03.006. 22

Daniel Quang and Xiaohui Xie. DanQ : a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Research, 44(11) :e107, June 2016. ISSN 1362-4962. doi : 10.1093/nar/gkw226. 26

Nikhil Rao, Christopher Cox, Rob Nowak, and Timothy T Rogers. Sparse Overlapping Sets Lasso for Multitask Learning and its Application to fMRI Analysis. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, pages 2202–2210. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/4891-sparse-overlapping-sets-lasso-for-multitask-learning-and-its-application-to-fmri-analysis.pdf. 34

Andrea Rau and Cathy Maugis-Rabusseau. Transformation and model choice for RNA-seq co-expression analysis. Briefings in Bioinformatics, 19(3) :425–436, May 2018. ISSN 1477-4054. doi : 10.1093/bib/bbw128. URL https://doi.org/10.1093/bib/bbw128. 17

David J. Reiss, Nitin S. Baliga, and Richard Bonneau. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics, 7 (1) :280, June 2006. ISSN 1471-2105. doi : 10.1186/1471-2105-7-280. URL https://doi.org/10.1186/1471-2105-7-280. 16, 18

Franziska Reiter, Sebastian Wienerroither, and Alexander Stark. Combinatorial function of transcription factors and cofactors. Current Opinion in Genetics & Development, 43 :73–81, April 2017. ISSN 1879-0380. doi : 10.1016/j.gde.2016.12.007. 29

Franziska Reiter, Bernardo P. de Almeida, and Alexander Stark. Enhancers display constrained sequence flexibility and context-specific modulation of motif function. Genome Research, March 2023. ISSN 1088-9051, 1549-5469. doi : 10.1101/gr.277246.122. URL https://genome.cshlp.org/content/early/2023/03/20/gr.277246.122. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab. 39

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1) :139–140, January 2010. ISSN 1367-4803. doi : 10.1093/bioinformatics/btp616. URL https://doi.org/10.1093/bioinformatics/btp616. 17

Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixOmics : An R package for 'omics feature selection and multiple data integration. PLOS Computational Biology, 13(11) : e1005752, November 2017. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1005752. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005752. Publisher : Public Library of Science. 9

Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of DNA shape in protein–DNA recognition. Nature, 461(7268) :1248–1253, October 2009. ISSN 1476-4687. doi : 10.1038/nature08473. URL https://www.nature.com/articles/nature08473. Number : 7268 Publisher : Nature Publishing Group. 30

Helge G. Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics (Oxford, England), 23(2) : 134–141, January 2007. ISSN 1367-4811. doi : 10.1093/bioinformatics/btl565. 30

Raphaël Roméro, Christophe Menichelli, Jean-Michel Marin, Sophie Lèbre, Charles-Henri Lecellier, and Laurent Bréhélin. Systematic analysis of the genomic features involved in the binding preferences of transcription factors, August 2022. URL https://www.biorxiv.org/content/10.1101/2022.08.16.504098v3. Pages : 2022.08.16.504098 Section : New Results. 31, 40

Sushmita Roy, Stephen Lagree, Zhonggang Hou, James A. Thomson, Ron Stewart, and Audrey P. Gasch. Integrated Module and Gene-Specific Regulatory Inference Implicates Upstream Signaling Networks. PLOS Computational Biology, 9(10) :e1003252, October 2013. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1003252. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003252. Publisher : Public Library of Science. 16, 18, 22

Joeri Ruyssinck, Vân Anh Huynh-Thu, Pierre Geurts, Tom Dhaene, Piet Demeester, and Yvan Saeys. NIMEFI : Gene Regulatory Network Inference using Multiple Ensemble Feature Importance Algorithms. PLOS ONE, 9(3) :e92709, March 2014. ISSN 1932-6203. doi : 10.1371/journal.pone.0092709. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0092709. Publisher : Public Library of Science. 19

Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W. Wasserman, and Boris Lenhard. JASPAR : an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research, 32(suppl_1) :D91–D94, January 2004. ISSN 0305-1048. doi : 10.1093/nar/gkh012. URL https://doi.org/10.1093/nar/gkh012. 19

Theresa Schacht, Marcus Oswald, Roland Eils, Stefan B. Eichmüller, and Rainer König. Estimating the activity of transcription factors by the effect on their target genes. Bioinformatics, 30 (17) :i401–i407, September 2014. ISSN 1367-4803. doi : 10.1093/bioinformatics/btu446. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4147899/. 20

Eric E. Schadt, Stephanie A. Monks, Thomas A. Drake, Aldons J. Lusis, Nam Che, Veronica Colinayo, Thomas G. Ruff, Stephen B. Milligan, John R. Lamb, Guy Cavet, Peter S. Linsley, Mao Mao, Roland B. Stoughton, and Stephen H. Friend. Genetics of gene expression surveyed in maize, mouse and man. Nature, 422(6929) :297–302, March 2003. ISSN 0028-0836. doi : 10.1038/nature01434. 19

Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science, 270(5235) : 467–470, October 1995. doi : 10.1126/science.270.5235.467. URL https://www.science.org/doi/10.1126/science.270.5235.467. Publisher : American Association for the Advancement of Science. 9

Bastian Schiffthaler, Elena van Zalen, Alonso R. Serrano, Nathaniel R. Street, and Nicolas Delhomme. Seiðr : Efficient Calculation of Robust Ensemble Gene Networks, March 2021. URL https://www.biorxiv.org/content/10.1101/250696v3. Pages : 250696 Section : New Results. 19

Florian Schmidt and Marcel H Schulz. On the problem of confounders in modeling gene expression. Bioinformatics, 35(4) :711–719, February 2019. ISSN 1367-4803. doi : 10.1093/bioinformatics/bty674. URL https://doi.org/10.1093/bioinformatics/bty674. 34

Florian Schmidt, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, Julia K. Polansky, Peter Ebert, Karl Nordström, Matthias Barann, Anupam Sinha, Sebastian Fröhler, Jieyi Xiong, Azim Dehghani Amirabad, Fatemeh Behjati Ardakani, Barbara Hutter, Gideon Zipprich, Bärbel Felder, Jürgen Eils, Benedikt Brors, Wei Chen, Jan G. Hengstler, Alf Hamann, Thomas Lengauer, Philip Rosenstiel, Jörn Walter, and Marcel H. Schulz. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. Nucleic Acids Research, 45(1) :54–66, January 2017. ISSN 0305-1048. doi : 10.1093/nar/gkw1061. URL https://academic.oup.com/nar/article/45/1/54/2605711. 20, 26, 27, 34

Juliane Schäfer and Korbinian Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics, 21(6) :754–764, March 2005. ISSN 1367-4803. doi : 10.1093/bioinformatics/bti062. URL https://academic.oup.com/bioinformatics/article/21/6/754/199211. Publisher : Oxford Academic. 12

E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. Bioinformatics (Oxford, England), 17 Suppl 1 :S243–252, 2001. ISSN 1367-4803. doi : 10.1093/bioinformatics/17.suppl_1.s243. 9, 21, 39

Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks : identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics, 34(2) :166–176, June 2003. ISSN 1546-1718. doi : 10.1038/ng1165. URL https://www.nature.com/articles/ng1165. Number : 2 Publisher : Nature Publishing Group. 18

Rajen D. Shah, Richard J. Samworth, and R. J. Samsworth. Variable selection with error control : another look at stability selection. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 75(1) :55–80, 2013. ISSN 1369-7412. URL https://www.jstor.org/stable/23361014. Publisher : Wiley. 33

Ning Shen, Jingkang Zhao, Joshua L. Schipper, Yuning Zhang, Tristan Bepler, Dan Leehr, John Bradley, John Horton, Hilmar Lapp, and Raluca Gordan. Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding. Cell Systems, 6(4) :470–483.e8, April 2018. ISSN 2405-4712. doi : 10.1016/j.cels.2018.02.009. 19

Richard I. Sherwood, Tatsunori Hashimoto, Charles W. O'Donnell, Sophia Lewis, Amira A. Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K. Gifford. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nature Biotechnology, 32(2) :171–178, February 2014. ISSN 1546-1696. doi : 10.1038/nbt.2798. URL https://www.nature.com/articles/nbt.2798. 30

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pages 3145–3153, Sydney, NSW, Australia, August 2017. JMLR.org. 27

Alireza F. Siahpirani and Sushmita Roy. A prior-based integrative framework for functional transcriptional regulatory network inference. Nucleic Acids Research, 45(4) :e21, February 2017. ISSN 0305-1048. doi : 10.1093/nar/gkw963. URL https://doi.org/10.1093/nar/gkw963. 16, 18, 22, 23

Claudia Skok Gibbs, Christopher A. Jackson, Giuseppe-Antonio Saldi, Andreas Tjärnberg, Aashna Shah, Aaron Watters, Nicholas De Veaux, Konstantine Tchourine, Ren Yi, Tymor Hamamsy, Dayanne M. Castro, Nicholas Carriero, Bram L. Gorissen, David Gresham, Emily R. Miraldi,

and Richard Bonneau. High-performance single-cell gene regulatory network inference at scale : the Inferelator 3.0. Bioinformatics (Oxford, England), 38(9) :2519–2528, April 2022. ISSN 1367-4811. doi : 10.1093/bioinformatics/btac117. 12, 14, 15, 16, 18, 20, 22, 33, 36, 37

Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code : how transcription factors read the genome. Trends in Biochemical Sciences, 39(9) :381–399, September 2014. ISSN 0968-0004. doi : 10.1016/j.tibs.2014.07.002. 19

I. Sobol. Sensitivity Estimates for Nonlinear Mathematical Models. 1993. URL https://www.semanticscholar.org/paper/Sensitivity-Estimates-for-Nonlinear-Mathematical-Sobol/3e0b415213a580254226fdbcbfc9980b70dd0468. 35

Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. Causation, prediction, and search. MIT press, 2000. 13

Divyanshi Srivastava and Shaun Mahony. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms, 1863(6) :194443, June 2020. ISSN 1876-4320. doi : 10.1016/j.bbagrm.2019.194443. 19

Gary D. Stormo. Modeling the specificity of protein-DNA interactions. Quantitative biology, 1(2) :115–130, June 2013. ISSN 2095-4689. doi : 10.1007/s40484-013-0012-4. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4101922/. 19

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures : Illustrations, sources and a solution. BMC Bioinformatics, 8 (1) :25, January 2007. ISSN 1471-2105. doi : 10.1186/1471-2105-8-25. URL https://doi.org/10.1186/1471-2105-8-25. 35

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, pages 3319–3328. PMLR, July 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html. ISSN : 2640-3498. 27

Yoshinori Tamada, SunYong Kim, Hideo Bannai, Seiya Imoto, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. Bioinformatics (Oxford, England), 19 Suppl 2 :ii227–236, October 2003. ISSN 1367-4811. doi : 10.1093/bioinformatics/btg1082. 21

The Gene Ontology Consortium. The Gene Ontology project in 2008. Nucleic Acids Research, 36(suppl_1) :D440–D444, January 2008. ISSN 0305-1048. doi : 10.1093/nar/gkm883. URL https://doi.org/10.1093/nar/gkm883. 17

Thomas Thorne. NetDiff – Bayesian model selection for differential gene regulatory network inference. Scientific Reports, 6(1) :39224, December 2016. ISSN 2045-2322. doi : 10.1038/srep39224. URL https://www.nature.com/articles/srep39224. Number : 1 Publisher : Nature Publishing Group. 40

Thomas Thorne. Approximate inference of gene regulatory network models from RNA-Seq time series data. BMC Bioinformatics, 19, April 2018. ISSN 1471-2105. doi : 10.1186/s12859-018-2125-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5896118/. 12

Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society : Series B (Methodological), 58(1) :267–288, 1996. ISSN 2517-6161. doi : 10.1111/j.2517-6161.1996.tb02080.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x. 11, 14, 15, 27

Ryan J. Tibshirani. The lasso problem and uniqueness. Electronic Journal of Statistics, 7 (none) :1456–1490, January 2013. ISSN 1935-7524, 1935-7524. doi : 10.1214/13-EJS815. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-7/issue-none/The-lasso-problem-and-uniqueness/10.1214/13-EJS815.full. Publisher : Institute of Mathematical Statistics and Bernoulli Society. 17, 28

Christa Geeke Toenhake, Sabine Anne-Kristin Fraschka, Mahalingam Shanmugiah Vijayabaskar, David Robert Westhead, Simon Jan van Heeringen, and Richárd Bártfai. Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying Plasmodium falciparum Blood-Stage Development. Cell Host & Microbe, 23(4) :557–569.e9, April 2018. ISSN 1931-3128. doi : 10.1016/j.chom.2018.03.007. URL https://www.sciencedirect.com/science/article/pii/S1931312818301367. 28

P. Tseng. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. Journal of Optimization Theory and Applications, 109(3) :475–494, June 2001. ISSN 1573-2878. doi : 10.1023/A:1017501703105. URL https://doi.org/10.1023/A:1017501703105. 29

Nasim Vahabi and George Michailidis. Unsupervised Multi-Omics Data Integration Methods : A Comprehensive Review. Frontiers in Genetics, 13, 2022. ISSN 1664-8021. URL https://www.frontiersin.org/articles/10.3389/fgene.2022.854752. 9

Jimmy Vandel, Océane Cassan, Sophie Lèbre, Charles-Henri Lecellier, and Laurent Bréhélin. Probing transcription factor combinatorics in different promoter classes and in enhancers. BMC Genomics, 20(1) :103, February 2019. ISSN 1471-2164. doi : 10.1186/s12864-018-5408-0. URL https://doi.org/10.1186/s12864-018-5408-0. 30, 39, 40

Haohan Wang, Benjamin J. Lengerich, Bryon Aragam, and Eric P. Xing. Precision Lasso : accounting for correlations and linear dependencies in high-dimensional genomic data. Bioinformatics (Oxford, England), 35(7) :1181–1187, April 2019. ISSN 1367-4811. doi : 10.1093/bioinformatics/bty750. 41

Jie Wang, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, Oliver J. Rando, Ewan Birney, Richard M. Myers, William S. Noble, Michael Snyder, and Zhiping Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research, 22(9) :1798–1812, September 2012. ISSN 1549-5469. doi : 10.1101/gr.139105.112. 19

Siguo Wang, Qinhu Zhang, Zhen Shen, Ying He, Zhen-Heng Chen, Jianqiang Li, and De-Shuang Huang. Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture. Molecular Therapy. Nucleic Acids, 24 :154–163, February 2021. ISSN 2162-2531. doi : 10.1016/j.omtn.2021.02.014. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7972936/. 30

Xin Wang, Peijie Lin, and Joshua W. K. Ho. Discovery of cell-type specific DNA motif grammar in cis-regulatory elements using random Forest. BMC Genomics, 19(1) :929, January 2018. ISSN 1471-2164. doi : 10.1186/s12864-017-4340-z. URL https://doi.org/10.1186/s12864-017-4340-z. 30

Nisar Wani and Khalid Raza. Integrative approaches to reconstruct regulatory networks from multi-omics data : A review of state-of-the-art methods. Computational Biology and Chemistry, 83 :107120, December 2019. ISSN 1476-9271. doi : 10.1016/j.compbiolchem.2019.107120. URL https://www.sciencedirect.com/science/article/pii/S1476927118305577. 20

Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. The Annals of Statistics, 37(5A) :2178–2201, October 2009. ISSN 0090-5364, 2168-8966. doi : 10.1214/08-AOS646. URL https://projecteuclid.org/journals/annals-of-statistics/volume-37/issue-5A/High-dimensional-variable-selection/10.1214/08-AOS646.full. Publisher : Institute of Mathematical Statistics. 9

Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics, 5(4) :276–287, April 2004. ISSN 1471-0064. doi : 10.1038/nrg1315. URL https://www.nature.com/articles/nrg1315. Number : 4 Publisher : Nature Publishing Group. 19

Paula Weidemüller, Maksim Kholmatov, Evangelia Petsalaki, and Judith B. Zaugg. Transcription factors : Bridge between cell signaling and gene regulation. PROTEOMICS, 21 (23-24) :2000034, 2021. ISSN 1615-9861. doi : 10.1002/pmic.202000034. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.202000034. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.202000034. 23

Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G. Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J. M. Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes. Determination and inference of eukaryotic transcription factor sequence specificity. Cell, 158(6) :1431–1443, September 2014. ISSN 1097-4172. doi : 10.1016/j.cell.2014.08.009. 30

Adriano V. Werhli and Dirk Husmeier. Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge. Statistical Applications in Genetics and Molecular Biology, 6(1), May 2007. ISSN 1544-6115. doi : 10.2202/1544-6115.1282. URL https://www.degruyter.com/document/doi/10.2202/1544-6115.1282/html. Publisher : De Gruyter. 21

John W. Whitaker, Zhao Chen, and Wei Wang. Predicting the human epigenome from DNA motifs. Nature Methods, 12(3) :265–272, March 2015. ISSN 1548-7105. doi : 10.1038/nmeth.3065. URL https://www.nature.com/articles/nmeth.3065. Number : 3 Publisher : Nature Publishing Group. 26, 29

J. Whittaker. Graphical Models in Applied Multivariate Statistics. Wiley, 1990. ISBN 978-0-471-91750-2. Google-Books-ID : MAfvAAAAMAAJ. 11

Anja Wille and Peter Bühlmann. Low-order conditional independence graphs for inferring genetic networks. Statistical Applications in Genetics and Molecular Biology, 5 :Article1, 2006. ISSN 1544-6115. doi : 10.2202/1544-6115.1170. 11

Marvin N. Wright and Andreas Ziegler. ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77 :1–17, March 2017. ISSN 1548-7660. doi : 10.18637/jss.v077.i01. URL https://doi.org/10.18637/jss.v077.i01. 34

Chuhu Yang, Eugene Bolotin, Tao Jiang, Frances M. Sladek, and Ernest Martinez. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene, 389(1) :52–65, March 2007. ISSN 0378-1119. doi : 10.1016/j.gene.2006.09.029. 27

James Yong, A Poi, Eugene Van Someren, Domenico Bellomo, and Marcel Reinders. Adaptive least absolute regression network analysis improves genetic network reconstruction by employing prior knowledge. 2008. 22

Ahrim Youn, David J. Reiss, and Werner Stuetzle. Learning transcriptional networks from the integration of ChIP–chip and expression data in a non-parametric model. Bioinformatics, 26 (15) :1879–1886, August 2010. ISSN 1367-4803. doi : 10.1093/bioinformatics/btq289. URL https://doi.org/10.1093/bioinformatics/btq289. 21

Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterProfiler : an R Package for Comparing Biological Themes Among Gene Clusters. OMICS : A Journal of Integrative Biology, 16(5) :284–287, March 2012. doi : 10.1089/omi.2011.0118. URL https://www.liebertpub.com/doi/10.1089/omi.2011.0118. 17

Han Yuan, Meghana Kshirsagar, Lee Zamparo, Yuheng Lu, and Christina S. Leslie. BindSpace decodes transcription factor binding signals by large-scale sequence embedding. Nature Methods, 16(9) :858–861, September 2019. ISSN 1548-7105. doi : 10.1038/s41592-019-0511-y. URL https://www.nature.com/articles/s41592-019-0511-y. 30

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 68(1) :49–67, February 2006. ISSN 1369-7412, 1467-9868. doi : 10.1111/j.1467-9868.2005.00532.x. URL http://doi.wiley.com/10.1111/j.1467-9868.2005.00532.x. 18, 34

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. Biometrika, 94 :19–35, February 2007. doi : 10.1093/biomet/asm018. 12

Neda Zarayeneh, Euiseong Ko, Jung Hun Oh, Sang Suh, Chunyu Liu, Jean Gao, Donghyun Kim, and Mingon Kang. Integration of multi-omics data for integrative gene regulatory network inference. International journal of data mining and bioinformatics, 18(3) :223–239, 2017. ISSN 1748-5673. doi : 10.1504/IJDMB.2017.10008266. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5771269/. 20

Kenneth S. Zaret and Jason S. Carroll. Pioneer transcription factors : establishing competence for gene expression. Genes & Development, 25(21) :2227–2241, November 2011. ISSN 0890-9369. doi : 10.1101/gad.176826.111. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3219227/. 30

Julia Zeitlinger. Seven myths of how transcription factors read the cis-regulatory code. Current Opinion in Systems Biology, 23 :22–31, October 2020. ISSN 2452-3100. doi : 10.1016/j.coisb.2020.08.002. URL http://www.sciencedirect.com/science/article/pii/S2452310020300305. 28

Arnold Zellner. Applications of Bayesian Analysis in Econometrics. Journal of the Royal Statistical Society. Series D (The Statistician), 32(1/2) :23–34, 1983. ISSN 0039-0526. doi : 10.2307/2987589. URL https://www.jstor.org/stable/2987589. Publisher : [Royal Statistical Society, Wiley]. 16, 22

Qinhu Zhang, Zhen Shen, and De-Shuang Huang. Predicting in-vitro Transcription Factor Binding Sites Using DNA Sequence + Shape. IEEE/ACM transactions on computational biology and bioinformatics, 18(2) :667–676, 2021. ISSN 1557-9964. doi : 10.1109/TCBB.2019.2947461. 30

Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 12(10) :931–934, October 2015. ISSN 1548-7105. doi : 10.1038/nmeth.3547. URL https://www.nature.com/articles/nmeth.3547. Number : 10 Publisher : Nature Publishing Group. 26, 29, 30

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature Genetics, 50(8) :1171–1179, August 2018. ISSN 1546-1718. doi : 10.1038/s41588-018-0160-6. URL https://www.nature.com/articles/s41588-018-0160-6. 26, 29

Yang Zhou, Rong Jin, and Steven Chu-Hong Hoi. Exclusive Lasso for Multi-task Feature Selection. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pages 988–995, March 2010. URL http://proceedings.mlr.press/v9/zhou10a.html. 34

Zhengze Zhou and Giles Hooker. Unbiased Measurement of Feature Importance in Tree-Based Methods. ACM Transactions on Knowledge Discovery from Data, 15(2) :26 :1–26 :21, January 2021. ISSN 1556-4681. doi : 10.1145/3429445. URL https://dl.acm.org/doi/10.1145/3429445. 35

J. Zhu, P. Y. Lum, J. Lamb, D. GuhaThakurta, S. W. Edwards, R. Thieringer, J. P. Berger, M. S. Wu, J. Thompson, A. B. Sachs, and E. E. Schadt. An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenetic and Genome Research, 105(2-4) :363–374, 2004. ISSN 1424-8581, 1424-859X. doi : 10.1159/000078209. URL https://www.karger.com/Article/FullText/78209. Publisher : Karger Publishers. 21

Jun Zhu, Matthew C. Wiener, Chunsheng Zhang, Arthur Fridman, Eric Minch, Pek Y. Lum, Jeffrey R. Sachs, and Eric E. Schadt. Increasing the Power to Detect Causal Associations by Combining Genotypic and Expression Data in Segregating Populations. PLOS Computational Biology, 3(4) :e69, April 2007. ISSN 1553-7358. doi : 10.1371/journal.pcbi.0030069. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030069. Publisher : Public Library of Science. 21

Jun Zhu, Bin Zhang, Erin N. Smith, Becky Drees, Rachel B. Brem, Leonid Kruglyak, Roger E. Bumgarner, and Eric E. Schadt. Integrating Large-Scale Functional Genomic Data to Dissect the Complexity of Yeast Regulatory Networks. Nature genetics, 40(7) :854–861, July 2008. ISSN 1061-4036. doi : 10.1038/ng.167. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2573859/. 21

Wencan Zhu, Céline Lévy-Leduc, and Nils Ternès. A variable selection approach for highly correlated predictors in high-dimensional genomic data. Bioinformatics, 37(16) :2238–2244, August 2021. ISSN 1367-4803. doi : 10.1093/bioinformatics/btab114. URL https://doi.org/10.1093/bioinformatics/btab114. 41

Pietro Zoppoli, Sandro Morganella, and Michele Ceccarelli. TimeDelay-ARACNE : Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinformatics, 11(1) :154, March 2010. ISSN 1471-2105. doi : 10.1186/1471-2105-11-154. URL https://doi.org/10.1186/1471-2105-11-154. 12

Hui Zou. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association, 101(476) :1418–1429, December 2006. ISSN 0162-1459. doi : 10.1198/016214506000000735. URL https://doi.org/10.1198/016214506000000735. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1198/016214506000000735. 33

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 67(2) :301–320, 2005. ISSN 1467-9868. doi : 10.1111/j.1467-9868.2005.00503.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2005.00503.x. 15, 17, 22, 29

Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. The Annals of Statistics, 37(4) :1733–1751, August 2009. ISSN 0090-5364, 2168-8966. doi : 10.1214/08-AOS625. URL https://projecteuclid.org/journals/annals-of-statistics/volume-37/issue-4/On-the-adaptive-elastic-net-with-a-diverging-number-of/10.1214/08-AOS625.full. Publisher : Institute of Mathematical Statistics. 22

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, 15(2) :265–286, 2006. ISSN 1061-8600. URL https://www.jstor.org/stable/27594179. Publisher : [American Statistical Association, Taylor & Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America]. 11

Jan Zrimec, Filip Buric, Mariia Kokina, Victor Garcia, and Aleksej Zelezniak. Learning the Regulatory Code of Gene Expression. Frontiers in Molecular Biosciences, 8, 2021. ISSN 2296-889X. URL https://www.frontiersin.org/article/10.3389/fmolb.2021.673363. 26

Yiming Zuo, Yi Cui, Guoqiang Yu, Ruijiang Li, and Habtom W. Ressom. Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO. BMC Bioinformatics, 18(1) :99, February 2017. ISSN 1471-2105. doi : 10.1186/s12859-017-1515-1. URL https://doi.org/10.1186/s12859-017-1515-1. 22

Quatrième partie

Annexe : Publications choisies

Annexe A

Roméro, R. , Menichelli, C., Marin, J.-M. , Lèbre, S., Lecellier, C.-H. , Bréhélin., L. (Soumis)

Systematic analysis of the genomic features involved in the binding preferences of transcription factors

Raphaël Romero[1,2] Christophe Menichelli[1] Jean-Michel Marin[2]
Sophie Lèbre[2,4†] Charles-Henri Lecellier[3†] Laurent Bréhélin[1†]

[1] LIRMM, Univ Montpellier, CNRS, Montpellier, France
[2] IMAG, Univ. Montpellier, CNRS, Montpellier, France
[3] Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France
[4] Univ. Paul-Valéry-Montpellier, Montpellier, France
[†] Corresponding authors: sophie.lebre@umontpellier.fr, charles.lecellier@igmm.cnrs.fr, brehelin@lirmm.fr

Abstract

Transcription factors (TFs) orchestrate gene expression and are at the core of cell-specific phenotypes and functions. One given TF can therefore have different binding sites depending on cell type and conditions. However, the TF core motif, as represented by Position Weight Matrix for instance, are often, if not invariably, cell agnostic. Likewise, paralogous TFs recognize very similar motifs while binding different genomic regions. We propose a machine learning approach called TFscope aimed at identifying the DNA features explaining the binding differences observed between two ChIP-seq experiments targeting either the same TF in two cell types or treatments or two paralogous TFs. TFscope systematically investigates differences in i) core motif, ii) nucleotide environment around the binding site and iii) presence and location of co-factor motifs. It provides the main DNA features that have been detected, and the contribution of each of these features to explain the binding differences. TFscope has been applied to more than 350 pairs of ChIP-seq. Our experiments showed that the approach is accurate and that the genomic features distinguishing TF binding in two different settings vary according to the TFs considered and/or the conditions. Several samples are presented and discussed to illustrate these findings. For TFs in different cell types or with different treatments, co-factors and nucleotide environment often explain most of the binding-site differences, while for paralogous TFs, subtle differences in the core motif seem to be the main reason for the observed differences in our experiments.

The source code (python), data and results of the experiments described in this article are available at `https://gite.lirmm.fr/rromero/tfscope`.

Introduction

The programming of gene expression is the primary mechanism that controls cellular phenotype and function. At the DNA level, transcription factors (TFs) are supposed to play a key role in this control. These proteins bind DNA sequence through specialized DNA binding domains (DBDs) to enhance or repress the transcription of their target genes. DBDs bind preferentially to specific DNA sequences, which are resumed in statistical models known as Position Weight Matrices (PWMs) [49]. Most of the times, PWMs are obtained from dedicated probabilistic models—the Position Probability Matrices (PPMs)—, which are available for many TFs in databases like JASPAR [15] and HOCOMOCO [28]. PWMs can be used to compute binding affinities and to identify potential binding sites in genomes. However, contrary to bacterial DBDs which recognize sequences that

often have sufficient information content to target particular genomic positions, most eukaryotic DBDs recognize short binding motifs (around 10bp) that are not sufficient for specific targeting in the usually large (*e.g.* 10^9bp) eukaryotic genomes [52]. This purely statistical analysis has been corroborated by genome-wide studies based on sequencing approaches (ChIP-seq, ChIP-exo, CUT&RUN) that have been applied to hundreds of TFs in order to determine their binding profiles in various cell types and conditions [13]. These studies showed that most TFs only associate with a small subset of their potential genomic sites *in vivo* [48], and that the binding sites of a given TF often vary substantially between cell types and conditions [44]. Furthermore, as the number of DBD families in a genome is small with regard to the number of TFs, TFs paralogs from the same DBD family often share very similar binding motifs, yet they usually show distinct binding sites *in vivo* [21, 42, 27]. Thus, it is now evident that DBD motifs as resumed by PWMs are not sufficient to completely determine TF binding in a specific cell or condition. On the other hand, several studies revealed that a substantial part of the *in vivo* binding sites lack an obvious match with the known binding motif of the target TF [48, 27].

At this point, it is important to emphasize the strong links that exist between TF binding and histone marks [14]. Also, ChIP-seq experiments revealed that most TF binding sites (TFBSs) lie within highly accessible (*i.e.*, nucleosome-depleted) DNA regions [45]. However, it remains unclear whether these chromatin states are a cause or a consequence of TF binding [20]. Moreover, recent approaches based on machine learning, and specifically convolutional neural networks (CNNs), have shown that transcription factor binding but also gene expression as well as histone modifications and DNase I-hypersensitive sites can be predicted from DNA sequence only, often with surprisingly high accuracy [50, 54, 36, 24, 47, 2]. The good predictive performances of these approaches suggest that a large part of the instructions for gene regulation and TF binding lies at the level of the DNA sequence.

Several mechanisms based on specific DNA features have thus been proposed to complement DBD motifs and to explain how TFs target precise genome locations. The current view is that TF combinations underlie the specificity of eukaryotic gene expression regulation [11], with several TFs competing and collaborating to regulate common target genes. Multiple mechanisms can lead to TF cooperation [34, 37]. In its simplest form, cooperation involves direct TF-TF interactions before any DNA binding. But cooperation can also be mediated through DNA, either with DNA providing additional stability to a TF-TF interaction [22], or without any direct protein-protein interaction, as in the pioneer/settler hierarchy described in Sherwood et al. [43] or in a non-hierarchical cooperative system such as the billboard model for enhancers [3, 33].

Besides TF combination, other studies have investigated the role the genomic environment around TFBS may have on binding specificity, revealing that some TFs have a preferential nucleotide content in the flanking positions of their core binding sites [30, 12]. Other studies have proposed that much larger regions containing repetitive sequences or multiple occurrences of low-affinity motifs may play an active role in TF binding [1, 9, 27]. Finally, another possibility that may be underestimated and that could also explain binding specificity in certain cases is that, depending on cell, condition, or TF paralog, the binding motif may actually differ, showing globally the same PWM to our eyes, but slightly differing on specific positions.

All these mechanisms have been independently studied on specific cases, but a global computational approach is still missing to investigate their role and relative importance in an automatized manner. The above-mentioned deep learning approaches are able to capture and combine the different DNA features involved, but identifying them from the CNNs remains a difficult task [26, 17]. Although interesting methods are being developed to post-analyze CNN predictions and to identify single nucleotides and motifs (see *e.g.* [54, 4, 26]), disentangling all mechanisms/features captured by a CNN remains unreliable.

Here, we propose a machine learning approach called TFscope specially designed to explain the binding differences observed between two settings: two cell types, two treatments, or two paralogous TFs. Our method directly compares the two ChIP-seq data associated with the two settings, by considering only regions bound either in one or the other experiment. This strategy

has two advantages. First, by focusing on the binding differences, we obviously gain sensitivity for identifying the genomic features that best explain these differences. Second, we circumvent the common problem of the background definition which arises in all studies that aim to distinguish bound (foreground) versus unbound (background) genomic regions in a given cell type. While the definition of the foreground is straightforward, the definition of the background is often much more challenging and strongly influences the results and the conclusions (see for example references [51, 50, 35, 53] for interesting considerations about the background issue).

Given two ChIP-seq data, our method systematically investigates the importance of i) the core motif, ii) the genomic environment and iii) the cooperative TFs for predicting the binding differences between two data. TFscope is based on three different modules that capture these three levels of information. The first module captures the potential differences in the core motif. This module is based on a new method that learns discriminative PWMs. It is worth noting that well known approaches such as DREME/STREME [5], DAMO [39], Homer [19], etc. have been already proposed for this task. These methods are however designed for a slightly different and computationally more complex problem, that is not exactly the same as ours. As a consequence, they rely on sub-optimal heuristics while an optimal algorithm exists for our problem. The second module captures the nucleotidic environment in the form of short k-mers (2-4 bps) enriched in specific regions around the core motif and is based on our DExTER method [32]. The third module is a refinement of our TFcoop method that identifies co-factors and TF combinations involved in the binding of a target TF [47]. In a final step, these data are used together in a global predictive model that is used to quantify the relative importance of each information for the problem at hand. Hence, in contrast to CNN based methods [53, 4], our approach completely controls the predictive features inputted into the model. This allows to easily measure the importance of each feature by computing the loss of accuracy induced by its withdrawal from the model, something very challenging to do with classical CNN approaches.

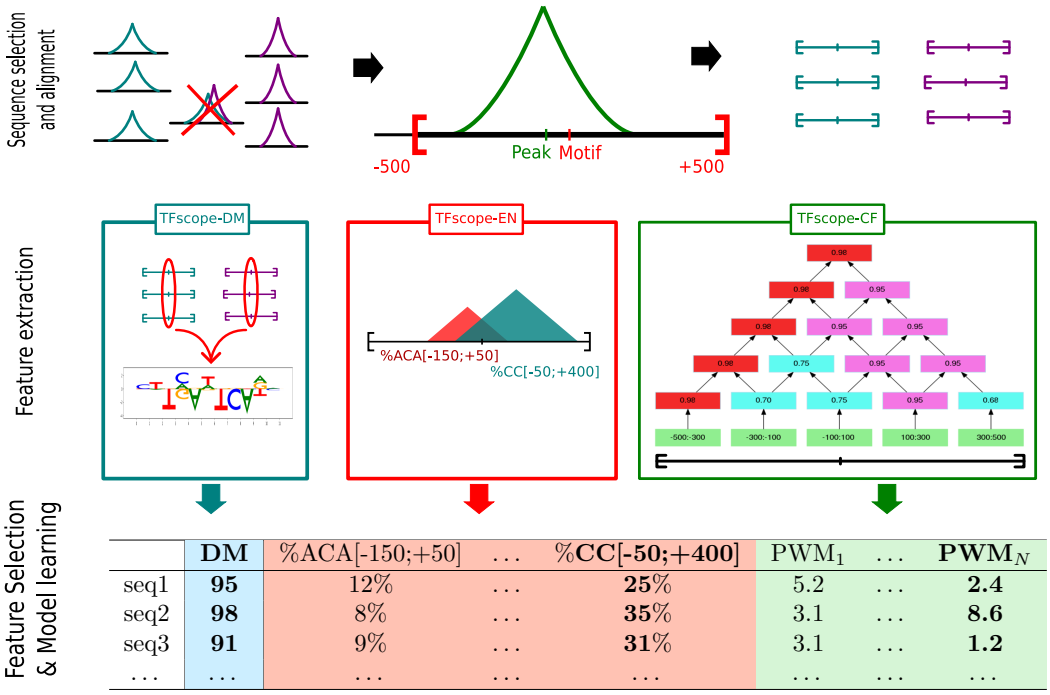
We applied TFscope to more than 350 pairs of ChIP-seq targeting either a common TF in two different cell types or treatments, or two paralogous TFs in the same cell type. Our results showed that classification is very often accurate and that the most important DNA features greatly vary depending on TFs and conditions. For TFs in different cell types or with different treatments, either co-factors or the nucleotidic environment often explainsmost of the binding-site differences. Moreover, when co-factors are involved, which is the most frequent case, their position on the DNA relative to the core motif is also important. On the contrary, for paralogous TFs the core motif seems to be the most important factor in our experiments. Although the motifs of paralogous TFs show very similar PWMs, subtle differences at specific positions explain most of the binding differences.

Results

TFscope overview

TFscope aims to identify the genomic features responsible for the binding differences observed between two ChIP-seq experiments. Typically, TFscope can be used to identify the differences between two experiments targeting the same TF in different cell types or conditions, or two experiments targeting two paralogous TFs that share similar motifs. TFscope takes in input two sets of ChIP-seq peaks corresponding to the two ChIP-seq experiments and then runs the three steps illustrated on Figure 1: sequence selection & alignment, feature extraction, and model learning. In the sequence alignment step, TFscope first identifies the peaks that are unique either to the first or the second set (see Material and Methods). All common peaks are discarded for the analysis (this point is discussed in the experiments below). Then, TFscope identifies the most likely binding site using a strategy similar to Centrimo [6] and UniBind [16], and parses the sequence around the peak summit with the PWM associated with the target TF (if several versions of the motif are available or if the analysis involves two paralogous TFs with similar motifs, the most discriminative PWM is chosen to scan the two sets of sequences; see Methods).

The FIMO tool [18] is used for this analysis, and the position with the highest PWM score is used as an anchor point to extract the 1Kb long sequence centered around this position. At the end of the alignment step, we get two classes of sequences centered on the most likely TFBSs of the ChIP-seq peaks given in input. Sequences with no occurrence of the motif around the peak summit are discarded.



	DM	%ACA[-150;+50]	...	**%CC[-50;+400]**	PWM_1	...	**PWM_N**
seq1	**95**	12%	...	**25%**	5.2	...	**2.4**
seq2	**98**	8%	...	**35%**	3.1	...	**8.6**
seq3	**91**	9%	...	**31%**	3.1	...	**1.2**
...	...	...	...	...	...	...	...

Figure 1 — The TFscope approach. In the first step (sequence selection and alignment), peaks associated with both ChIP-seq experiments are removed. The most likely TFBSs of the remaining peaks are identified and used to extract the 1Kb sequences centered on these sites. All sequences are then used for the second step (feature extraction). Three dedicated modules extract three kinds of genomic features that can be useful for discriminating the two classes. The TFscope-DM module learns a new PWM that can discriminate the sequences on the basis of the core motif solely. The TFscope-EN module searches for specific nucleotidic environments (*i.e.* frequency of specific k-mers in specific regions) that are different in the two classes. The TFscope-CF module searches for binding sites of specific co-factors whose presence in specific regions differs between the two classes. All these features (variables) are then gathered into a long table, and a logistic model (Expression (1)) is learned on the basis of these data (Feature selection and model learning). A special penalty function (LASSO) is used during training, for selecting only the best variables in the model (in bold in the table).

TFscope then runs three modules detailed below to extract three kinds of DNA features that are discriminative of the two sequence classes (feature-extraction step). The first module (TFscope-DM) learns a new PWM of the core motif (see below). This PWM is different from the original PWM used to parse the sequence, as it focuses on the potential differences of core motif that may exist between the two sequence sets. This module returns a single variable $\mathrm{DM}(s)$, which is the score of the new PWM on each sequence s. The second module (TFscope-NE) searches for pairs of (k-mer,region) for which the frequency of the k-mer in the defined region is different between the two sets of sequences. For example, we may observe that the frequency of the 3-mer ACA in region $[-150 : +500]$ (0 being the anchor point of the sequences) is globally higher in sequences of the first class than in those of the second class. The idea is to capture the differences in nucleotidic environment that may exist between the two classes. We use for this module a slight modification

of the DExTER method recently proposed to identify long regulatory elements [32]. This module returns a potentially large set of variables $\mathrm{NE}_i(s)$ that corresponds to the frequency of ith k-mer in the associated ith region for each sequence s. The third module (TFscope-CF) uses a library of PWMs (in the experiments below the JASPAR2020 library is used [15]) and searches for pairs of (PWM,region) for which the score of the PWM in the identified region is different between the two sets of sequence (see below). The idea is to identify all co-factors of the target TF whose binding sites differ between the two classes: either because these binding sites are in majority present in one class and not the other, or because the locations of these binding sites are different between the two classes. This third module returns a set of variables $\mathrm{CF}_j(s)$ that corresponds to the score of the jth PWM in the identified jth region for each sequence s.

All variables are then integrated into a global model that aims to predict if a sequence belongs to the first or the second class (learning step). We used a logistic regression model:

$$P(1|s) = S\left(a \cdot \mathrm{DM}(s) + \sum_i b_i \cdot \mathrm{NE}_i(s) + \sum_j c_j \cdot \mathrm{CF}_j(s)\right), \tag{1}$$

where $P(1|s)$ is the probability that sequence s belongs to the first class, S is the sigmoid function, $\mathrm{DM}(s)$ is the score of the discriminative motif for sequence s, $\mathrm{NE}_i(s)$ is the value of the ith nucleotidic-environment variable for sequence s, $\mathrm{CF}_j(s)$ is the value of the jth co-factor variable for sequence s, and a, b_i and c_j are the regression coefficients which constitute the parameters of the model. Because the set of variables identified by the last two modules is usually large and variables are often correlated, the model is trained with a LASSO penalty function [46] that selects the most relevant variables—*i.e.* many regression coefficients (a, b_i and c_j) are set to zero. Finally, once a model has been trained, its accuracy is evaluated by computing the area under the ROC (AUROC) on several hundred sequences. To avoid any bias, this is done on a set of sequences that have not been used in the previous steps.

TFscope-DM: Identification of differences in the core motif

The first TFscope module learns a new discriminative PWM. Recall that at the end of the alignment step, the most likely binding site of each ChIP-seq peak has been identified with the JASPAR PWM associated with the TF, and all sequences are aligned on these sites. If several versions of the PWM are available, the most discriminative PWM is used (see Methods). We then extract the K-length sub-sequence corresponding to the occurrence of the motif in each sequence (K being the size of the PWM). The first module aims to learn a new PWM that could discriminate these two sets of K-length sequences. First, each sequence s is one-hot encoded in a $K \times 4$ matrix $\mathbf{s}$. Then, a logistic model with $K \times 4$ parameters is learned to discriminate the two classes of sequences:

$$P(1|s) = S\left(\sum_{k=1}^{K}\sum_{j=1}^{4} a_{k,j} \cdot \mathbf{s}_{k,j}\right), \tag{2}$$

with $P(1|s)$ the probability that sub-sequence s belong to the first class, S the sigmoid function, $\mathbf{s}_{k,j}$ the entry of the one-hot matrix $\mathbf{s}$ indicating whether the kth nucleotide of sequence s corresponds to the jth nucleotide of $\{A, T, G, C\}$ or not, and $a_{k,j}$ the regression coefficients of the model.

Once this model has been learned, it can be used to predict if a sequence belongs to the first or the second class. The sigmoid function being monotonically increasing, this can be done easily by computing the linear function inside the parenthesis of Expression (2) and using the result as a score reflecting the likelihood of class 1. Interestingly, this score function has exactly the same form as the one used to compute a score with a PWM (see Material and methods). As a consequence, the logistic model of Expression (2) is strictly speaking a regular PWM with parameters $a_{k,j}$. The interest to learn a PWM in this way is two folds. First, we take advantage of all the algorithmic and theory developed for logistic regression. Most notably, as the likelihood function of a logistic model is convex, we have the guarantee that the learned model is optimal, which means that the inferred discriminative PWM is the best PWM for our problem. This is

an important difference from the approaches already proposed to learn a discriminative PWM, such as DAMO [39] or STREME [5]. The reason for this is that these approaches do not exactly address the same problem as ours: they do not search for a PWM that discriminates two sets of sequences perfectly aligned and of the same length as the PWM. Instead, they take as input two sets of sequences usually much longer than the PWM, and their goal is to identify a motif whose presence can be used to discriminate the two sets, a problem known to be NP-hard [31]. As a consequence, these approaches rely on heuristics and do not warrant returning the best PWM for our problem (see section Discussion for more details on these differences). The second advantage to learn a PWM via a logistic regression approach is that we can include a LASSO penalty in the optimization procedure in order to obtain a model with fewer variables [46] (see Material and methods). In practice, this means that many parameters $a_{k,j}$ are set to zero, and hence that the resulting PWM is simpler and easier to interpret.

It is important to note that, as DAMO [39], the PWMs output by our method are not obtained from position probability matrices (PPMs), which are the probabilistic models that are often associated with PWMs. This avoids the constraints attached to PPMs (see section Discussion and the work of Ruan and Stormo [38] for more details) but this also impedes to represent PWMs with the classical logo graphics based on information theory [40]. Instead, our PWMs are represented by "mirror-logos" such as the one on Figure 2B (middle). These logos provide the sign of the parameters, which allows to easily distinguish the nucleotides that are more present in sequences of one or the other class.

TFscope-NE: Identification of differences in nucleotidic environment

The second TFscope module extracts features related to the nucleotidic environment around the core binding motif. More precisely, this module constructs variables defined by a pair (kmer,region) such that the frequency of the identified k-mer in the identified region is, on average, different in the two classes. We used for this a slight modification of the DExTER method initially proposed to identify pairs of (kmer,region) whose values are correlated with an expression signal. The optimization function of DExTER has thus been modified to return variables correlated with classes rather than with expression signal (see Material and method). The TFscope-NE module explores short k-mers up to length 4. To prevent this module to capture information related to the core-motif, this motif is masked before running the TFscope-NE analysis.

TFscope-CF: Identification of differences in co-factor combinations

The third TFscope module extracts features related to co-factors. This module constructs variables defined by a pair (PWM,region) such that the score of the PWM in the identified region is, on average, different between the two classes. For example, one can observe that sequences of the first class often have a potential binding site for a specific TF in region [-250,0] upstream the binding site of the target TF, while the sequences of the second class have not these potential binding sites. Hence, the goal of this module is to identify, for each PWM of the library, a specific region of the sequences in which the scores of this PWM are higher in one class than in the other one.

Sequences are first segmented in bins of the same size. We used 13 bins in the following experiments. The number of bins impacts the precision of the approach but also the computing time of the analysis. For each PWM, TFscope scans all sequences with FIMO [18], and the best score achieved on each bin of each sequence is stored. Then, TFscope searches the region of consecutive bins for which the PWM gets the most different scores depending on the class of the sequences. A lattice structure is used for this exploration (see Figure 1 and details in section Materials and Methods). For each PWM of the library, TFscope-CF selects the region that shows the highest differences and returns a variable corresponding to this PWM and region. As for TFscope-NE, the core-motif is masked before running the analysis.

Analysis of the cellular specificities of 272 ChIP-seq pairs

We first sought to apply TFscope to identify binding sites differences of TFs in different cell types using a selection of 272 pairs of ChIP-seq experiments downloaded from the GTRD database [25]. To minimize the effects linked to technical issues or indirect binding, data were filtered using the UniBind p-value score [16]. In UniBind the authors studied the distance between the ChIP-seq peaks and the position of the most likely binding site (inferred with the PFM associated with the TF studied). They showed that this binding site is sometimes far from the ChIP-seq peak, and that the peak may be a false positive. Using a dedicated method named ChIP-eat, the authors were able to determine genomic boundaries inside which the binding sites are likely true positives, and they provide a p-value measuring peak enrichment in these boundaries. We used this p-value to remove ChIP-seq experiments that could be affected by technical issues and indirect binding. Moreover, we only selected for this analysis pairs of experiments that show strong binding site differences according to Jaccard's distance (see Materials and method). The 272 pairs were chosen to provide a wide view of the ChIP-seq data in GTRD, *i.e.* pairs that were too close to another pair already selected were discarded (see the pair selection procedure in Materials and method).

TFscope learns both discriminative and informative core motifs

We first assessed the TFscope ability to identify core motif differences in the ChIP-seq experiments pairs. In this analysis, we only used the score function of the learned PWM (Expression (2)) to discriminate the two cell-types. For comparison, we also used the score of the original PWM on this problem. Accuracies were measured by AUROC on an independent set of sequences (see Figure 2A). If several versions of the original PWM were available, we used the version that provides the best AUROC. As we can see, the new PWM outperforms the original PWM most of the time. Moreover, we can also observe that for some of the 272 experiment pairs, the core motif itself is sufficient to differentiate the two cell-types with high accuracy. As already discussed, the discriminative PWM is different from the original PWM, as it specifically models the differences while removing the features common to the two classes. The "mirror-logo" representation summarizes these differences and shows which features are associated with which cell-type. For example, the Figure 2B (up) shows the original CEBPA PWM provided by JASPAR, while the middle figure shows the mirror logo of the discriminative PWM learned by TFscope for discriminating CEBPM binding sites between the SKH1 and U937 cell-types. One can see here that the canonical CEBPA motif is more often associated with ChIP-seq peaks collected in U937 than in SKH1. Although the SKH1 sequences also bear a very similar motif (recall that the motif is present at the center of the sequences for both conditions) the mirror logo indicates for example that the T nucleotides at positions 3 and 4 are more often missing in the SKH1 sequences than in the U937 sequences. Similarly, among the small differences that may exist between the motifs in the two conditions, it seems that the SKH1 sequences often have a C at position 5.

We next sought to compare these results to those obtained with another method that learns discriminative PWMs. We used the DAMO approach for this comparison, as it is one of the rare methods that do not rely on PPM to learn a PWM. Recall that DAMO, as the other classical approaches to learn PWMs, has not been designed to address exactly the same problem as our. Indeed, DAMO usually takes in input sequences that are not aligned and that are much longer than the target PWM. Nevertheless, it can also be used on our simpler problem. However, as illustrated on Figure 2C, it does not achieve the same accuracy as TFscope on this problem, which was somewhat expected as the logistic classifier used by TFscope theoretically returns the most discriminative PWM.

Another striking fact when we compare the discriminative motifs learned by DAMO to those of TFscope is that the DAMO motifs appear much more complex, with a lot of positions without clear preferences. On the contrary, thanks to the LASSO penalty used for learning, the TFscope motifs are easier to interpret, with many positions set to zero (compare the two examples provided on Figure 2B). This aspect was assessed systematically on the 272 experiments using a score function based on the Gini coefficient for measuring motif simplicity (see Material and methods). As

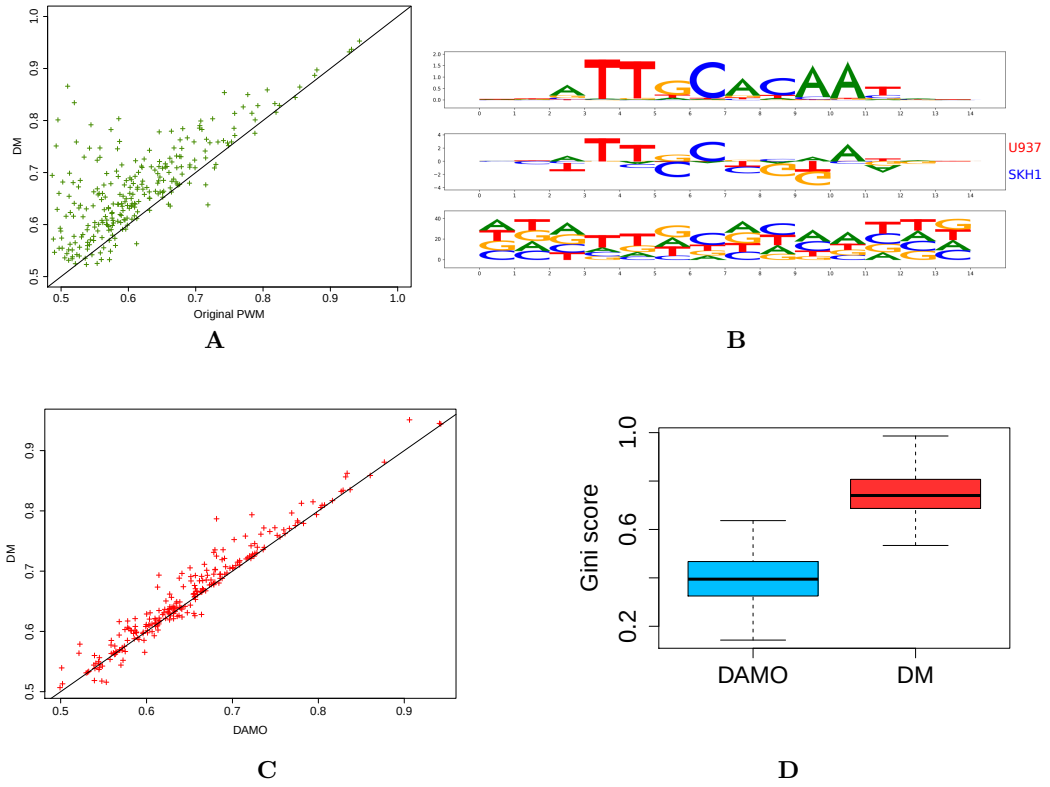


Figure 2 — TFscope learns discriminative and informative motifs. A AUROCs achieved by the TFscope PWMs *vs.* original PWMs on the 272 experiments. **B(up)** Original (JASPAR) PWM of TF USF2. **B(middle)** discriminative PWM learned by TFscope for discriminating USF2 binding between HepG2 and GM12878. **B(bottom)** discriminative PWM learned by DAMO on the same training set. **C** AUROCs achieved by the TFscope PWMs *vs.* DAMO PWMs on the 272 experiments.**D** Gini score of the PWMs learned by TFscope and DAMO. The higher the Gini score, the simpler the model.

illustrated on Figure 2D, TFscope motifs have higher Gini coefficient, and are thus simpler and easier to interpret than their DAMO counterpart.

Finally, we observed that increasing the size of the PWM until 4 nucleotides on both sides still improves the AUROC of the DM model (see Supp. Fig. 1). So, in the following, this model (denoted as DM+8) is used in the TFscope model and the experiments.

Position of co-factors helps for predicting cell-specificity

We next sought to investigate the information gained by the position of the binding sites of potential co-factors for cell-type prediction. We used for this a simplification of the model of Expression (1), which only uses the core motif and the co-factor variables for the prediction—*i.e.* the NE_i variables capturing the nucleotidic environment were removed from the model. The accuracy of this model was compared to that of a similar model that also uses the score of potential co-factors, but without integrating the information of position. This model, which strongly resembles the TFcoop approach we previously proposed [47], simply uses the best score achieved by the different PWMs in the whole sequence. Hence, the predictive variables of this model are the best scores achieved at any position of the sequence, while in Expression (1) TFscope uses the best score achieved in a specific region identified as the most informative for each co-

factor. While the two models have exactly the same number of parameters (*i.e.* the number of PWMs in the PWM library), the variables of TFscope greatly increase the accuracy of the approach (Figure 3), illustrating the fact that position of co-factors relative to the considered TFBS also carry important information. Note that, as we will see in the following, TFscope provides a graphical representation of all identified co-factors, and position information can be easily retrieved.

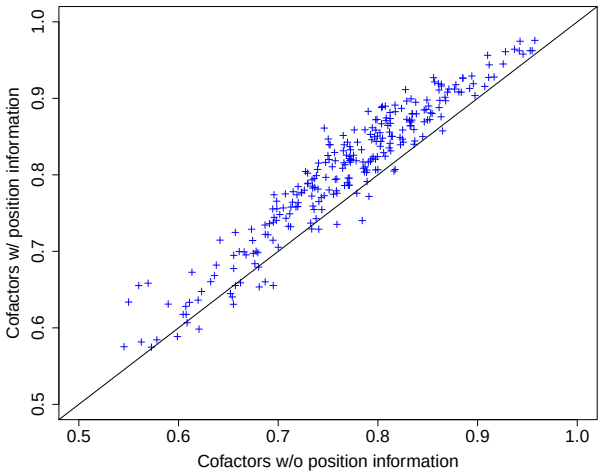


Figure 3 — Position of co-factors helps for predicting cell-specificity. AUROCs achieved by TFscope models that only use the score of co-factors for the predictions. On the x-axis, the score of a co-factor on a sequence s is the best achieved at any position of the sequence, *i.e.* the position of the co-factor is not considered in the score computation. On the y-axis, the score of each co-factor is only computed on a specific region identified by the TFscope-CF module to be the most informative for this co-factor.

TFscope assesses the relative importance of each genomic feature

We next ran TFscope with the full model of Expression (1) on the 272 pairs of experiments and compared its accuracy (AUROC) to that of different alternatives: the original PWM only, the discriminative PWM only, and three incomplete TFscope models that only use two of the three kinds of genomic information. These incomplete models were obtained by taking the full TFscope model trained with all variables, and by setting to zero either the variable DM (model TFscope w/o core motif information), or the NE_i variables (TFscope w/o nucleotidic environment information), or the CF_j variables (TFscope w/o co-factor information). Figure 4A reports the accuracy achieved by all these models. As we can see, the full TFscope model successfully integrates the three kinds of genomic information and outperforms the alternative models. We can also observe that the accuracy is often good, with a median AUROC above 80%. Moreover, there is a strong link between the accuracy of the approach and the Jaccard distance between the ChIP-seq peaks in the two cell types (Pearson r=0.51; see Figure 4B), *i.e.* experiments with low proportion of ChIP-seq peaks shared by the two cell types often get good accuracy (remember that these peaks are removed before the analyses). In other words, when the two ChIP-seq experiments are really different, TFscope accurately predicts these differences. Figure 4B also illustrates the fact that, for most analyses, the Jaccard distance is high (so the Jaccard index is low, see Supp. Fig. 2). This means that the number of common peaks is small in proportion, hence removing these peaks makes sense for our analyses.

These good performances legitimate the use of TFscope to investigate the relative importance

of each kind of genomic information in the different comparisons. For this, in addition to the logo of the discriminative PWM, TFscope outputs a radar plot that summarizes the accuracy of the different models and alternatives, and a location plot that summarizes the position of the most important variables of the model (see Material and methods). For example, Figure 5B reports the radar plot obtained when analyzing the binding differences of TF JUND between liver and lung carcinoma. For this experiment, the core motif is clearly the most discriminant information (Figure 5A), since removing this information lead to the largest drop of AUROC. Besides, peaks detected in lung harbor additional AP-1 motifs around the core motif (Figure 5C). JunD belongs to the AP-1 family of dimeric TFs, which associate members of the Jun (c-Jun, JunB and JunD) and Fos (c-Fos, FosB, Fra-1/Fosl1 and Fra-2/Fosl2) families. In contrast to the Jun family members, which can homodimerize, the Fos family members must heterodimerize with one of the Jun proteins to bind DNA. Importantly, Fos:Jun heterodimers have a stronger affinity for DNA than the Jun:Jun homodimers [7]. According to various expression data listed in the EBI Expression Atlas (`https://www.ebi.ac.uk/gxa/home`), Fos TFs are less expressed in liver than in lung. Thus, JunD binding preferences observed in liver *vs.* lung might merely be explained by the expression of Fos TFs: because the probability to form Fos:Jun heterodimers is greater in lung than in liver, JunD will bind DNA with a higher affinity in lung than in liver. For comparison, the discriminative motif and the radar plot of the CTCF experiment between CD20 and RH4 are on Figure 5D-E. Here, the most discriminative information seems to be the nucleotidic environment. The location plot provides additional information (Figure 5F). We can see that CD20 favors A/T rich environment in the vicinity of the binding motif ($\sim +/-100$bp around the motif), and C/G nucleotides in the larger surrounding region ($+/-500$bp). On the contrary, RH4 prefers a nucleotide environment rich in TG and CA dinucleotides. All results obtained on the 272 experiments are available on `https://gite.lirmm.fr/rromero/tfscope/-/tree/main/results`.

Finally, in an attempt to provide a broad picture of the genomic strategies involved in the control of binding differences between cell types, we ran a K-means clustering on the importance profiles inferred by TFscope. More precisely, all 272 experiments were described by a vector of length 3 obtained by subtracting the AUROC of TFscope w/o DM, TFscope w/o NE and TFscope w/o CF from that of the full TFscope model. Each experiment is then represented by three values representing the three AUROC losses associated with the three kinds of information. We reasoned that a maximum of 7 broad classes can be expected from these data: three classes with a single information clearly higher than the two others, three classes with two more-important information, and one class with approximately equal importance of the three information. However, by visually inspecting the results of several K-means clustering, we end up with a total of only 4 classes: the three-information class, and the DM+NE and DM+CF classes seem absent from the 272 models. Figure 4C reports the distribution of the 272 models in these 4 classes, highlighting the fact that the co-factors are by far the most common mechanism involved in the binding differences between cell types. The target motif itself appears to be the more discriminative feature in 10.6% of the 272 experiments, and the nucleotidic environment around the binding site in 14% cases.

Analysis of the binding differences induced by a specific treatment

We next sought to use TFscope to analyze the binding differences observed between two ChIP-seq experiments targeting the same cell type but with two different treatments. 79 ChIP-seq pairs were selected (see Material & Methods) and analyzed. As for the cell type comparisons, all results obtained on the 79 experiments are available on `https://gite.lirmm.fr/rromero/tfscope/-/tree/main/results`. We got globally similar results than for the cell type experiments (see the plot of accuracy in Supp. Fig.3A), although the Jaccard distance between treatments is often smaller than between cell types (Supp. Fig.3B), *i.e.* two treatments often show more similar binding sites than two cell types. However, for several experiments there is a clear difference in the binding sites and TFscope indentifies interesting features.

For instance, TFscope confirms the cross-talk between GR signaling and NF-κB reported in [23] and proposes additional features. Specifically, analyzing NR3C1 ChIP-seq upon Dexamethasone (Dex) and Dex+TNF treatments with TFscope reveals that the main features distinguishing the binding sites in these two conditions are cooperating TFs (Figure 6A). While motifs of NFI-related TFs are enriched in NR3C1 peaks upon Dex treatment alone, as observed in [23], motifs of NF-κB-related TFs are enriched in NR3C1 peaks upon Dex+TNF (Figure 6B).

Similarly, cooperating TFs appear as the main features distinguishing RELA ChIP-seq peaks upon TNF and Dex+TNF treatments (Figure 6C). TFscope confirms that, in the presence of Dex, RELA peaks are associated with steroid receptor TFs (NR3C1, NR3C2 and AR) but it also suggests that GR signaling abolishes cooperation with AP-1 related TFs observed preferentially in RELA peaks in pro-inflammatory conditions (TNF alone) (Figure 6D).

Analysis of the binding differences of paralogous TFs

We showed in a previous work [8] that the binding of two paralogous TFs, namely FOSL1 and FOSL2 (also called as FRA1 and FRA2), can be distinguished primarily by the scores of their motif: FOSL2 preferentially binds sequences harboring high scores for the canonical AP-1 motif, while FOSL1 binds sequences with some degenerate positions (lower scores). We then thought to use TFscope to distinguish FOSL1 from FOSL2 binding on the same dataset. TFscope-DM is indeed sufficient to classify the two peak classes (Figure 7A) and the typical AP-1 motif is more frequently found in FOSL2 peaks (Figure 7B), confirming our previous results obtained with another approach [47]. Moreover, the discriminative motif also brings new information. For example, FOSL1 favors nucleotides that are inverse from those of the canonical motif in positions 2 and 10.

To confirm the applicability of TFscope in this sort of classification task, we considered another pair of paralogous TFs, NR3C1 and AR, and ChIP-seq data collected in MCF-7 cells [41]. As shown in Figure 7C, TFscope is able to accurately distinguish NR3C1 from AR ChIP peaks, and TFscope-DM shows that the main differences lie in the core motif itself. The output of TFscope-DM reveals that dinucleotides AC at position 4 and GT at position 11 in the canonical NR3C1/AR motif are more frequent for NR3C1 than for AR (Figure 7D). Moreover, AR ChIP-seq peaks appear more GC-rich than NR3C1 peaks (Figure 7E). These results are in full agreement with that obtained by Kulik *et al.*, who compared AR and GR binding preferences in U2OS cells [29]. Together these results illustrate the possibility to use TFscope to distinguish the binding of paralogous TFs.

Discussion

We proposed here a new machine learning approach to identify the DNA features that can explain the binding preferences of a TF in two settings: two cell types, two conditions, or two paralogous TFs. Our approach uses three modules that identify three kinds of DNA features related to TF binding. The first one is a new method to learn a discriminative PWM. Among the numerous approaches already proposed to learn a PWM, it is important to note that most of them actually learn a PPM which is then converted into a PWM with a simple log ratio formula (see for example reference [49]). The problem with this procedure is that it potentially impedes the accuracy of the PWM. Indeed, PPMs being probabilistic models, they are subject to strong constraints (notably, the sum of a PPM column must be equal to one) which inevitably also constrains the weights of the PWM. For example, the log-ratio operation cannot produce a PWM in which one of the columns has all but one weight equal to zero (the log ratio gives zero when the probability of the nucleotide at this position equals the probability of the nucleotide in the background; but if it is the case for 3 nucleotides it is also necessarily the case for the 4th nucleotide). Interested readers can refer to the work of Ruan and Stormo [38] for more arguments about the limits of PPMs for PWM learning. Our approach based on logistic regression avoids this problem and has moreover the advantage of allowing us to include a LASSO penalty to get simpler and more

readable PWMs. Another important difference with previous approaches is that in TFscope the discriminative PWM is only used to discriminate the two classes, but not to scan the sequences. Hence TFscope needs two PWMs: the JASPAR PWM is used to scan the sequences and identify the binding sites in both classes, and the discriminative PWM is used to score the binding sites and differentiate the two classes.

In addition to the specificity of the core motif that is captured by our discriminative PWM, the two other modules extract DNA features related to the nucleotidic environment around the TFBS, and the presence and position of every potential co-factor. Then, a learning algorithm is run to both train a model and select the most discriminative features at the same time. Hence, contrary to the CNNs based methods that have been recently proposed, our approach completely controls the predictive features used by the model. This allows us to easily assess the global importance of each feature, by measuring the loss of accuracy induced by its withdrawal, something very challenging to do with CNN approaches.

Our results on different TFs and different cell types show that co-factors are often the most important determinant associated with the cell-specific binding sites, and that their position relative to the TFBS considered is key. However, for several experiments such as CTCF in CD20 *vs.* RH4 the large nucleotidic environment around the binding sites also explains a part of the observed differences. For some other experiments such as JUND in lung *vs.* liver the main differences lie directly in specific nucleotides of the binding site. When comparing two treatments the picture is globally the same, while for paralogous TFs the main differences are associated with the core motifs themselves in our experiments. In this latter case, although the binding motifs globally show very similar PWMs for both TFs, subtle differences at specific positions actually explain most of the binding differences.

Our approach could be improved in different ways. Notably, one drawback that can sometimes hamper a straight interpretation of the TFscope results is the correlation between predictive variables. Scores of TF motifs especially may be highly correlated, as several TFs often share very similar motifs. Hence, although the PWM highlighted by TFscope is the one that shows the highest link with the predicted signal, other PWMs could also have a high correlation, and thus other TFs are potential co-factors. We therefore encourage users to refer to PWM clusters as defined for example in the RSAT-matrix clustering [10]. Similarly, there are sometimes correlations between the nucleotidic composition captured by the TFscope-NE module and the co-factor motifs identified by TFscope-CF. Here again, the linear model and the LASSO penalty ensure that the variables selected by TFscope are those with the strongest link with the predicted signal. Nevertheless, it is important to keep in mind that other variables may actually be involved in the studied process. We are thus working on a way to identify and present all alternative variables in a friendly interface. Another improvement would be to integrate additional DNA features into our model. Specifically, the number of repeats of a given PWM could be an interesting variable for discriminating two ChIP-seq experiments. Such information is not directly accounted for in the current model but could potentially explain binding differences in some experiments.

Material and Methods

Sequence extraction and alignment

TFscope takes in input two sets of ChIP-seq peaks provided as BED files. First, all peaks common to the two files are removed. This is done with the BED tools using

```
bedtools window -v -w 500 -a class0.bed -b class1.bed > class0_no_overlap.bed
bedtools window -v -w 500 -a class1.bed -b class0.bed > class1_no_overlap.bed
```

Then the sequences corresponding to each peak are extracted and aligned on the most likely occurrence of the TFBS. We use for this the PWM associated with the target TF in JASPAR 2020 [15]. FIMO is used to parse the sequences with the command

```
fimo --thresh 0.001 --max-strand --text --bfile background_fimo.txt
   PPM_jaspar2020.meme fasta.fa > occurrences.dat
```

The best occurrence of the motif around the ChIP-seq peak (in a limit of 500kb) is identified and used as an anchor point around which the 1Kb sequences are centered (see Figure 1). Sequences for which no occurrence of the motif is found around the peak summit are discarded. Finally, the number of sequences of the two classes are rebalanced, *i.e.* some sequences of the larger class are randomly selected and removed, in order to get two classes with an equal number of sequences.

If several versions of the PWM are available in JASPAR, we used the PWM that is the most discriminative for the problem at hand. Namely, for each PWM, the best occurrence of the motif is identified on each sequence, and these scores are used to discriminate the two classes (this corresponds to the AUROC of the original PWM in the radar plots). The PWM version with the highest AUROC is used for the rest of the analysis.

The formatted data used in the experiments are available in the dedicated gite repository: `https://gite.lirmm.fr/rromero/tfscope`.

TFscope-DM

This module takes as input the K-length sub-sequence corresponding to the most likely occurrence of the motif in each sequence (K being the size of the PWM). Each sub-sequence s is one-hot encoded in a $K \times 4$ matrix $\mathbf{s}$. Then, a logistic model with $K \times 4$ parameters (see Expression 2) is learned to discriminate the two classes of sub-sequences. The parameters of the model are estimated by maximum likelihood, with a LASSO penalization [46] to favor simple and easy-to-interpret models. This is done with library glmnet on python 3.

TFscope-NE

This module takes in input the 1Kb sequences centered on the most likely TFBS (cf. Sequence extraction and alignment). The K-length sub-sequence corresponding to the TFBS is masked (replaced by K N nucleotides) to avoid capturing information related to the core-motif. Then the TFscope-NE module constructs new variables defined by a pair (kmer,region) such that the frequency of the identified k-mer in the associated region is, on average, different between the two classes. We used for this a slight modification of the DExTER method [32]: rather than searching for variables that are correlated with an expression signal, TFscope-NE extracts variables that can discriminate the two classes, as measured by the AUROC. The rest of the procedure is exactly the same as that used in DExTER (see ref. [32] for details). Sequences are first segmented into different bins. We used 7 bins in the experiments. TFscope-NE starts with 2-mer (dinucleotides) and, for each 2-mer, identifies the region of consecutive bins for which the 2-mer frequency in the region is the most discriminant. Once the best region has been identified for a 2-mer, TFscope-NE attempts to iteratively extend this 2-mer for identifying longer k-mers (up to 4-mers). At the end of the process, a set of variables corresponding to the frequency of the identified k-mers in the identified regions is returned for each sequence.

TFscope-CF

As TFscope-NE, this module takes in input the 1Kb sequences centered on the most likely TFBS (this TFBS is also masked to avoid capturing information related to the core-motif). This module constructs variables defined by a pair (PWM,region) such that the score of the PWM in the identified region is, on average, different between the two classes. For this, sequences are first segmented in bins of the same size. We used 13 bins in the experiments. The number of bins impacts the precision of the approach but also the computing time of the analysis. For each PWM, TFscope scans all sequences with FIMO, and the best score achieved in each bin of each sequence is stored. Then, TFscope uses a lattice structure (see Figure 1) to compute the best score achieved in any region made up of consecutive bins. Each node of the lattice is associated

with a specific region: the top of the lattice represents the whole sequence, while the lowest nodes represent the different bins. Once the best score achieved in every bin has been computed, the best score achieved in any node of the lattice can be easily deduced with a max() operation on its two children nodes. For example, the lattice of Figure 1 corresponds to a sequence for which the best score is obtained in the first bin ([-500;-300]). For each PWM, a lattice like this one is computed for every sequence. Then, TFscope identifies the node (region) such that the scores associated with this node in the different lattices provide the highest AUROC for discriminating the two classes.

Selection of 272 ChIP-seq pairs targeting the same TF in two different cell types

272 pairs of experiments targeting a common TF, with the same treatment, in two different cell-types were selected from the GTRD and UniBind databases. ChIP-seq data were downloaded from GTRD `http://gtrd20-06.biouml.org/downloads/20.06/bigBeds/hg38/ChIP-seq/Clusters_by_TF_and_Peakcaller/MACS2/`. Only experiments associated with a UniBind p-value below 1% were considered, which represents a total of 2815 ChIP-seq data. This data can be arranged in a total of 6553 pairs targeting a common TF, with the same treatment, in two different cell-types. We chose to select only the pairs that show highly different peaks for the analyses. This was measured with the Jaccard's distance. Let A and B be two sets of ChIP-seq peaks on the genome, the Jaccard's distance D_J is defined from the Jaccard index by:

$$D_j = 1 - \frac{|A \cap B|}{|A \cup B|}. \tag{3}$$

Peak intersections and unions were computed with Bedtools window and merge, respectively:

```
bedtools window -w 500 -a class0.bed -b class1.bed > intersection.bed
bedtools merge -d 500 -i intersection.bed
```

For a given TF and treatment, several pairs with different cell-types are often possible. In order to select only a subset of these pairs, we ran a hierarchical clustering of all the data targeting the same TF with the same treatment. The clustering was done using the Jaccard's distance and the complete-linkage agglomeration strategy. We then selected one pair of experiments for each internal node of the tree (the two experiments with the highest number of peaks were selected). Hence, if the tree has N leaves (corresponding to the N ChIP-seq experiments targeting the same TF and treatment) the number of pairs is exactly N. In this way, we end up with a total of 425 ChIP-seq pairs, which were reduced to 368 pairs by selecting only pairs with at least 1000 specific peaks in each cell-type. Among these pairs, more than 100 actually involved CTCF. We chose to keep only 7 CTCF pairs (which were chosen as the pairs with the largest Jaccard distance), and we end up with a final set of 272 pairs of experiments.

Measure of PWM simplicity

Information content derived from information theory is often used to measure the conservation of specific nucleotides at specific positions of a motif. This measure is however based on probability distribution and is thus restricted to PPMs: it does not extend to the PWM general case. Hence, we propose to use the Gini coefficient to measure PWM simplicity.

The Lorenz curve is a graphical representation of the income distribution between individuals in econometrics. It is obtained by ordering the individuals in the order of their income, and by calculating the cumulative part of income in function of the cumulative part of individuals (see Supp. Figure 4). In the case of an equal distribution between all the individuals, the curve follows the line $y = x$. Otherwise, it is found under this line. The surface A is the area between the Lorenz curve and the line of perfect equality of distribution. The surface B is the area between the Lorenz curve and the perfect inequality curve (all the income belongs to a single individual).

The Gini coefficient is defined by $A/(A+B)$. It is equal to 1 if all the incomes belong to a single individual and equal to 0 if the incomes are equally distributed.

For PWMs, we compute the Gini coefficient on the set of PWM weights. More precisely, we gather the $4 \times K$ weights of the PWM (all positions combined), order them in ascending order of their absolute value, and compute the Lorenz curve and the Gini coefficient associated with this set of weights. A small Gini coefficient implies an equal distribution of the PWM weights: the information is dispersed on many elements of the PWM. On the contrary, a large Gini coefficient (close to 1) indicates that some weights of the PWM gather all the information and that many weights are equal or close to 0. So we can see it as a measure of the interpretability of the models, where a model with a large Gini coefficient will be simpler than a model with a Gini coefficient close to 0. Importantly, the Gini coefficient does not depend on the scale of the PWM weights but only on their relative importance. Hence, it can be used to compare PWMs obtained from different methods.

Measure of variable importance

We devised an *ad hoc* procedure based on LASSO penalty and model error for measuring the individual importance of the different variables of a model. Given a penalization constraint λ, the LASSO procedure searches the model parameters that minimize the prediction error subject to the constraint. In practice, a grid of constraints of decreasing values is initialized, and a model is learned for each value. The result is a series of models with an increasing number of parameters. To identify the most important variables of a model in a given condition, we took the model with 10 parameters and estimated the importance of each of the 10 variables in the following way. Given a variable X, its importance was estimated by the AUROC difference between the complete model and the model obtained by setting β_X to 0.

Selection of 79 ChIP-seq pairs targeting the same TF with two different treatments

To compare binding preferences upon different treatments, we removed experiments associated with the 'no-condition' term in our GTRD/UniBind joint list. We sorted the remaining $1,354$ experiments according to their GTRD IDs in order to consider experiments from the same publication/study. We further removed time-course experiments and selected 100 pairs of possible comparisons. Then, the same procedure as the one used for the selection of the 272 ChIP-seq pairs was applied. This gives a total of 79 pairs of ChIP-seq experiments.

Acknowledgements

We thank Marius Gheorghe and Anthony Mathelier for their helpful comments and assistance with Unibind data.

References

[1] Ariel Afek, Hila Cohen, Shiran Barber-Zucker, Raluca Gordân, and David B. Lukatsky. Nonconsensus Protein Binding to Repetitive DNA Sequence Elements Significantly Affects Eukaryotic Genomes. *PLOS Computational Biology*, 11(8):e1004429, August 2015. Publisher: Public Library of Science.

[2] Vikram Agarwal and Jay Shendure. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Reports*, 31(7):107663, May 2020.

[3] David N. Arnosti and Meghana M. Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry*, 94(5):890–898, April 2005.

[4] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, March 2021. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 3 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Chromatin immunoprecipitation;Computational biology and bioinformatics;Genomics Subject_term_id: chromatin-immunoprecipitation;computational-biology-and-bioinformatics;genomics.

[5] Timothy L Bailey. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, (btab203), March 2021.

[6] Timothy L. Bailey and Philip Machanick. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, 40(17):e128, September 2012.

[7] Fabienne Bejjani, Emilie Evanno, Kazem Zibara, Marc Piechaczyk, and Isabelle Jariel-Encontre. The AP-1 transcriptional complex: Local switch or remote command? *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1872(1):11–23, 2019. Publisher: Elsevier.

[8] Fabienne Bejjani, Claire Tolza, Mathias Boulanger, Damien Downes, Raphaël Romero, Muhammad Ahmad Maqbool, Amal Zine El Aabidine, Jean-Christophe Andrau, Sophie Lebre, Laurent Brehelin, Hughes Parrinello, Marine Rohmer, Tony Kaoma, Laurent Vallar, Jim R Hughes, Kazem Zibara, Charles-Henri Lecellier, Marc Piechaczyk, and Isabelle Jariel-Encontre. Fra-1 regulates its target genes via binding to remote enhancers without exerting major control on chromatin architecture in triple negative breast cancers. *Nucleic acids research*, 49(5):2488–2508, 2021. Publisher: Oxford University Press.

[9] Milagros Castellanos, Nivin Mothi, and Victor Muñoz. Eukaryotic transcription factors can track and control their target genes using DNA antennas. *Nature Communications*, 11, January 2020.

[10] Jaime Abraham Castro-Mondragon, Sébastien Jaeger, Denis Thieffry, Morgane Thomas-Chollier, and Jacques van Helden. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Research*, 45(13):e119, July 2017.

[11] Hemangi G. Chaudhari and Barak A. Cohen. Local sequence features that influence AP-1 cis-regulatory activity. *Genome Research*, 28(2):171–181, February 2018.

[12] Iris Dror, Tamar Golan, Carmit Levy, Remo Rohs, and Yael Mandel-Gutfreund. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Research*, 25(9):1268–1280, January 2015.

[13] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696):636–640, October 2004.

[14] Jason Ernst and Manolis Kellis. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Research*, 23(7):1142–1154, July 2013.

[15] Oriol Fornes, Jaime A. Castro-Mondragon, Aziz Khan, Robin van der Lee, Xi Zhang, Phillip A. Richmond, Bhavi P. Modi, Solenne Correard, Marius Gheorghe, Damir Baranašić, Walter Santana-Garcia, Ge Tan, Jeanne Chèneby, Benoit Ballester, François Parcy, Albin Sandelin, Boris Lenhard, Wyeth W. Wasserman, and Anthony Mathelier. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 48(D1):D87–D92, January 2020.

[16] Marius Gheorghe, Geir Kjetil Sandve, Aziz Khan, Jeanne Chèneby, Benoit Ballester, and Anthony Mathelier. A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Research*, 47(4):e21, February 2019.

[17] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of Neural Networks Is Fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, July 2019. Number: 01.

[18] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011.

[19] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, 38(4):576–589, May 2010.

[20] Łukasz Huminiecki and Jarosław Horbańczuk. Can We Predict Gene Expression by Understanding Proximal Promoter Architecture? *Trends in Biotechnology*, 0(0), April 2017.

[21] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-Binding Specificities of Human Transcription Factors. *Cell*, 152(1–2):327–339, January 2013.

[22] Arttu Jolma, Yimeng Yin, Kazuhiro R. Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, Teemu Kivioja, Ekaterina Morgunova, and Jussi Taipale. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578):384–388, November 2015.

[23] Vineela Kadiyala, Sarah K Sasse, Mohammed O Altonsy, Reena Berman, Hong W Chu, Tzu L Phang, and Anthony N Gerber. Cistrome-based cooperation between airway epithelial glucocorticoid receptor and NF-κB orchestrates anti-inflammatory effects. *Journal of Biological Chemistry*, 291(24):12673–12687, 2016. Publisher: ASBMB.

[24] David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, May 2018.

[25] Semyon Kolmykov, Ivan Yevshin, Mikhail Kulyashov, Ruslan Sharipov, Yury Kondrakhin, Vsevolod J Makeev, Ivan V Kulakovskiy, Alexander Kel, and Fedor Kolpakov. GTRD: an integrated view of transcription regulation. *Nucleic Acids Research*, 49(D1):D104–D111, January 2021.

[26] Peter K. Koo and Sean R. Eddy. Representation learning of genomic sequence motifs with convolutional neural networks. *PLOS Computational Biology*, 15(12):e1007560, December 2019.

[27] Judith F. Kribelbauer, Chaitanya Rastogi, Harmen J. Bussemaker, and Richard S. Mann. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annual Review of Cell and Developmental Biology*, 35(1):357–379, October 2019.

[28] Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan S. Yevshin, Ruslan N. Sharipov, Alla D. Fedorova, Eugene I. Rumynskiy, Yulia A. Medvedeva, Arturo Magana-Mora, Vladimir B. Bajic, Dmitry A. Papatsenko, Fedor A. Kolpakov, and Vsevolod J. Makeev. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, 2017.

[29] Marina Kulik, Melissa Bothe, Gözde Kibar, Alisa Fuchs, Stefanie Schöne, Stefan Prekovic, Isabel Mayayo-Peralta, Ho-Ryun Chung, Wilbert Zwart, Christine Helsen, Frank Claessens, and Sebastiaan H Meijsing. Androgen and glucocorticoid receptor direct distinct transcriptional programs by receptor-specific and shared DNA binding sites. *Nucleic acids research*, 49(7):3856–3875, 2021. Publisher: Oxford University Press.

[30] Michal Levo, Einat Zalckvar, Eilon Sharon, Ana Carolina Dantas Machado, Yael Kalma, Maya Lotam-Pompan, Adina Weinberger, Zohar Yakhini, Remo Rohs, and Eran Segal. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research*, 25(7):1018–1029, January 2015.

[31] Ming Li, Bin Ma, and Lusheng Wang. Finding similar regions in many strings. In *Proceedings of the thirty-first annual ACM symposium on Theory of Computing*, STOC '99, pages 473–482, New York, NY, USA, May 1999. Association for Computing Machinery.

[32] Christophe Menichelli, Vincent Guitard, Rafael M. Martins, Sophie Lèbre, Jose-Juan Lopez-Rubio, Charles-Henri Lecellier, and Laurent Bréhélin. Identification of long regulatory elements in the genome of Plasmodium falciparum and other eukaryotes. *PLOS Computational Biology*, 17(4):e1008909, April 2021. Publisher: Public Library of Science.

[33] Leonid A. Mirny. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22534–22539, December 2010.

[34] Ekaterina Morgunova and Jussi Taipale. Structural perspective of cooperative transcription factor binding. *Current Opinion in Structural Biology*, 47:1–8, December 2017.

[35] Gherman Novakovsky, Manu Saraswat, Oriol Fornes, Sara Mostafavi, and Wyeth W. Wasserman. Biologically relevant transfer learning improves transcription factor binding prediction. *Genome Biology*, 22(1):280, September 2021.

[36] Daniel Quang and Xiaohui Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11):e107–e107, June 2016.

[37] Franziska Reiter, Sebastian Wienerroither, and Alexander Stark. Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics & Development*, 43:73–81, April 2017.

[38] Shuxiang Ruan and Gary D. Stormo. Inherent limitations of probabilistic models for protein-DNA binding specificity. *PLOS Computational Biology*, 13(7):e1005638, July 2017. Publisher: Public Library of Science.

[39] Shuxiang Ruan and Gary D. Stormo. Comparison of discriminative motif optimization using matrix and DNA shape-based models. *BMC Bioinformatics*, 19(1):86, March 2018.

[40] T D Schneider and R M Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, October 1990.

[41] Tesa M Severson, Yongsoo Kim, Stacey EP Joosten, Karianne Schuurman, Petra Van Der Groep, Cathy B Moelans, Natalie D Ter Hoeve, Quirine F Manson, John W Martens, Carolien HM Van Deurzen, Ellis Barbe, Ingrid Hedenfalk, Peter Bult, Vincent T. H. B. M. Smit, Sabine C. Linn, Paul J. van Diest, Lodewyk Wessels, and Wilbert Zwart. Characterizing steroid hormone receptor chromatin binding landscapes in male and female breast cancer. *Nature communications*, 9(1):1–12, 2018. Publisher: Nature Publishing Group.

[42] Ning Shen, Jingkang Zhao, Joshua L. Schipper, Yuning Zhang, Tristan Bepler, Dan Leehr, John Bradley, John Horton, Hilmar Lapp, and Raluca Gordan. Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding. *Cell Systems*, 6(4):470–483.e8, April 2018.

[43] Richard I. Sherwood, Tatsunori Hashimoto, Charles W. O'Donnell, Sophia Lewis, Amira A. Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K. Gifford. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, 32(2):171–178, February 2014.

[44] Divyanshi Srivastava and Shaun Mahony. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1863(6):194443, June 2020.

[45] Robert E. Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, Andrew B. Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K. Canfield, Morgan Diegel, Douglas Dunn, Abigail K. Ebersol, Tristan Frum, Erika Giste, Audra K. Johnson, Ericka M. Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D. Nguyen, Hongzhu Qu, Alex P. Reynolds, Vaughn Roach, Alexias Safi, Minerva E. Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M. Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O. Dorschner, R. Scott Hansen, Patrick A. Navas, George Stamatoyannopoulos, Vishwanath R. Iyer, Jason D. Lieb, Shamil R. Sunyaev, Joshua M. Akey, Peter J. Sabo, Rajinder Kaul, Terrence S. Furey, Job Dekker, Gregory E. Crawford, and John A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, September 2012.

[46] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[47] Jimmy Vandel, Océane Cassan, Sophie Lèbre, Charles-Henri Lecellier, and Laurent Bréhélin. Probing transcription factor combinatorics in different promoter classes and in enhancers. *BMC Genomics*, 20(1):103, February 2019.

[48] Jie Wang, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, Oliver J. Rando, Ewan Birney, Richard M. Myers, William S. Noble, Michael Snyder, and Zhiping Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22(9):1798–1812, January 2012. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[49] Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, April 2004.

[50] John W. Whitaker, Zhao Chen, and Wei Wang. Predicting the Human Epigenome from DNA Motifs. *Nature methods*, 12(3):265–272, March 2015.

[51] Rebecca Worsley Hunt, Anthony Mathelier, Luis del Peso, and Wyeth W. Wasserman. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics*, 15(1):472, June 2014.

[52] Zeba Wunderlich and Leonid A. Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in genetics: TIG*, 25(10):434–440, October 2009.

[53] An Zheng, Michael Lamkin, Hanqing Zhao, Cynthia Wu, Hao Su, and Melissa Gymrek. Deep neural networks identify sequence context features predictive of transcription factor binding. *Nature machine intelligence*, 3(2):172–180, February 2021.

[54] Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934, October 2015.

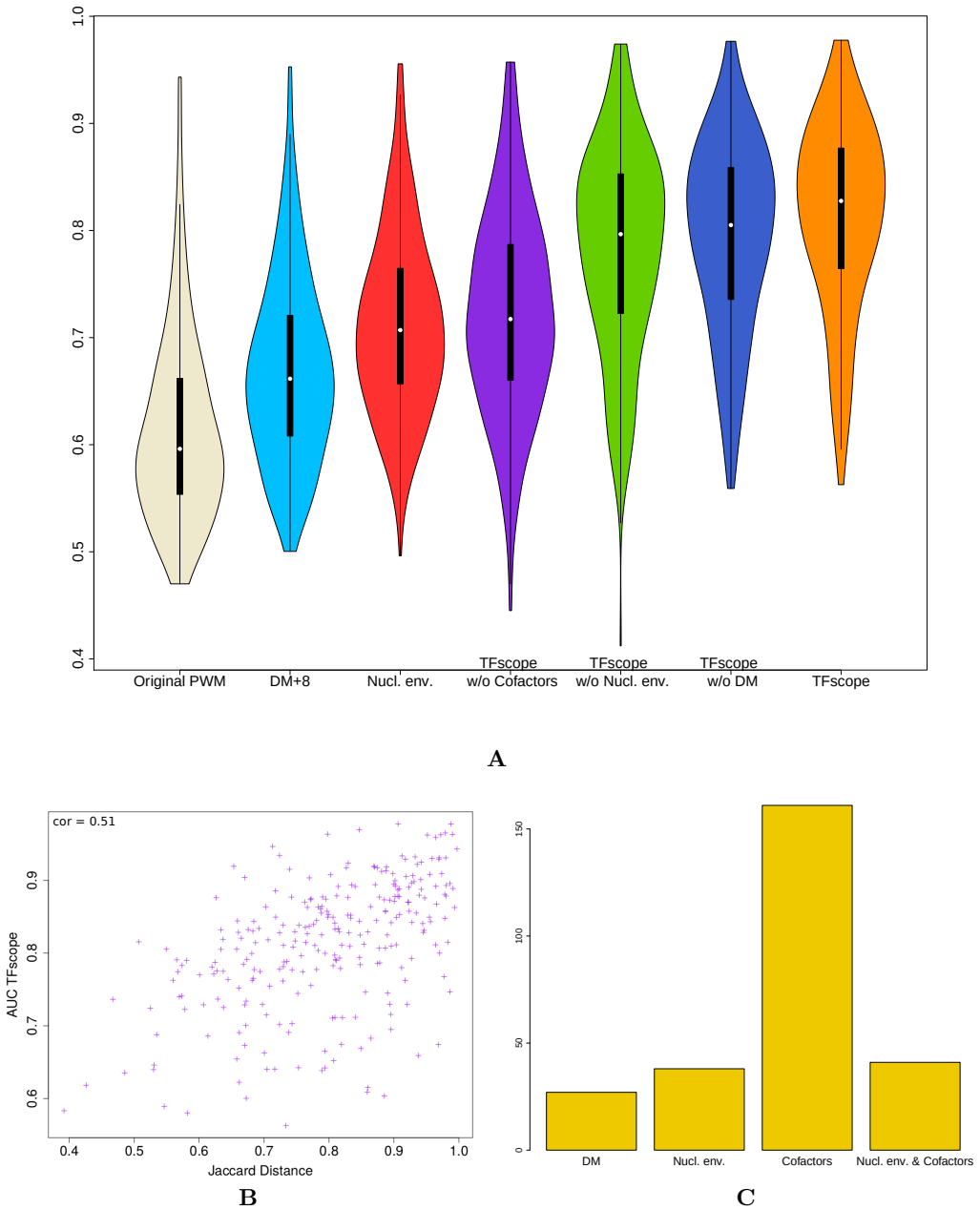


Figure 4 — Accuracy achieved by TFscope for discriminating binding sites of different cell types. **A** Distribution of AUROCs achieved by TFscope and several alternative models for discriminating binding sites of one TF in two different cell types. **B** Link between TFscope accuracy and the similarity of ChIP-seq peaks in the two cell types. ChIP-seq experiments that have a high proportion of peaks in common have low Jaccard distance (Jaccard distance = 1 - Jaccard index). **C** Distribution of TFscope models according to what are the most discriminative features: the discriminative motif (DM), the nucleotidic environment, the co-factors, or the nucleotidic environment + co-factors.

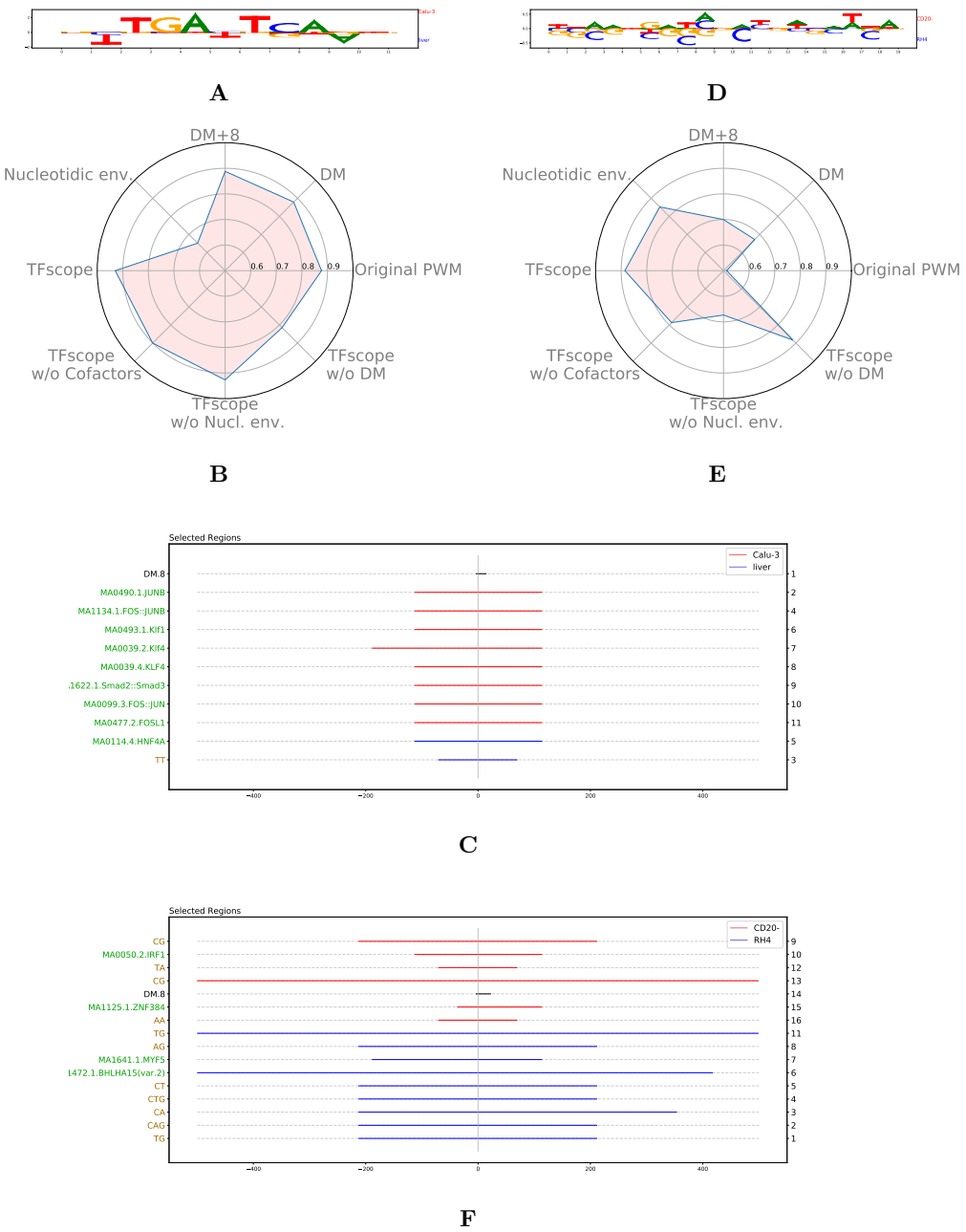


Figure 5 — Core motif, nucleotidic environment and co-factors together determine cell specificity. A-B-C Discriminative PWM, radar plot and location of the most important variables in the JUND comparison between liver and lung carcinoma. **D-E-F** Discriminative PWM, radar plot and location of the most important variables in the CTCF comparison between B lymphocite and rhabdomyosarcoma. Radar plots (B & E) summarize the AUROC achieved by TFscope and several alternative models. Location plots (C & F) provide the identity and location of the most important variables (black: DM; green: co-factors; brown: nucleotidic environment). The numbers on the right hand indicate the ranking of the variables, from the most important (rank 1) to the least important. The color of segments indicates the cell-type associated with each variable.

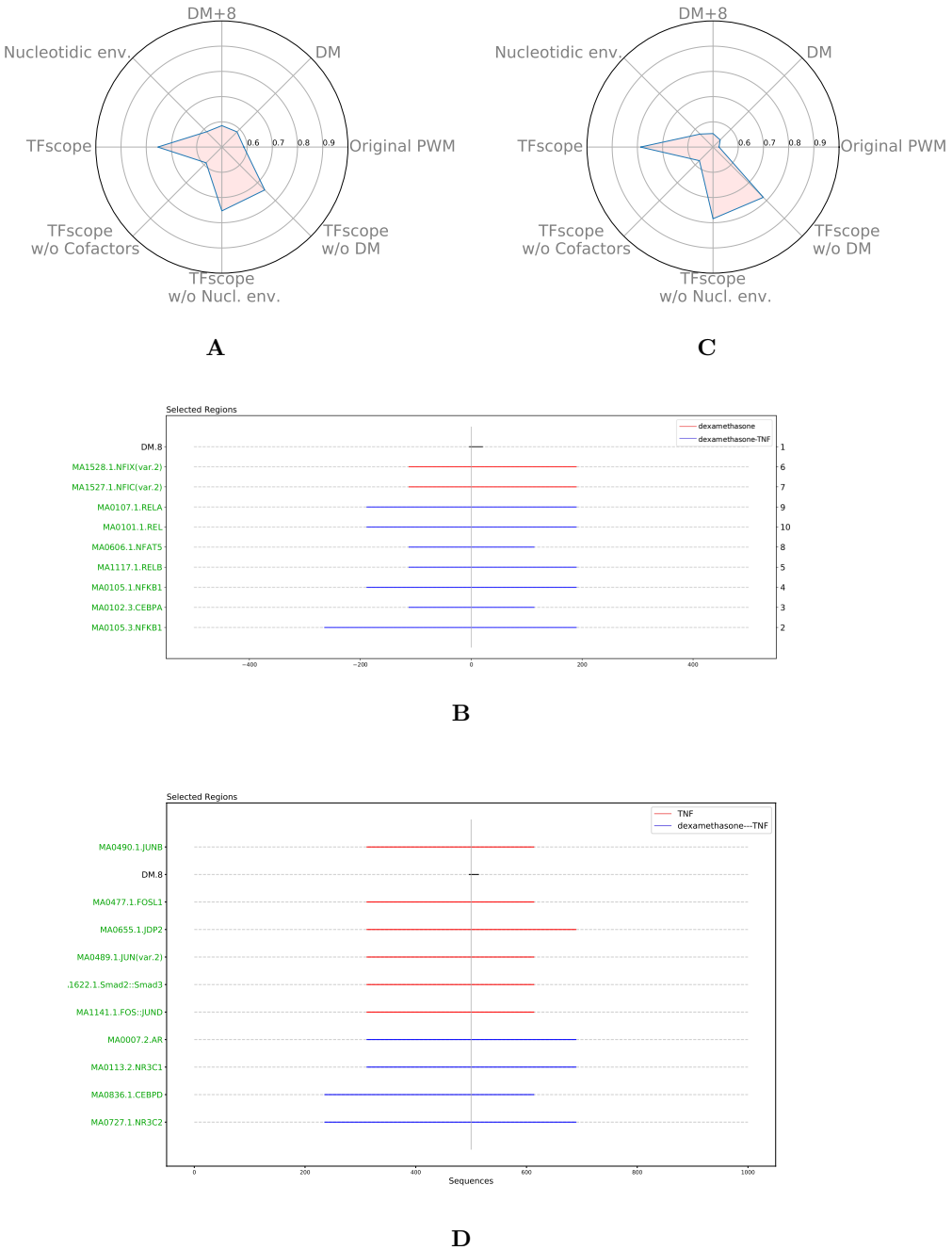
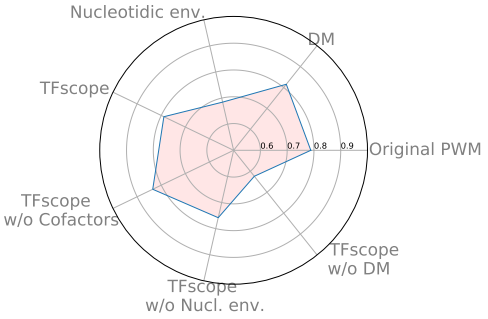
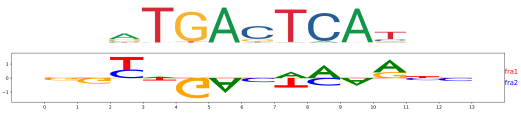


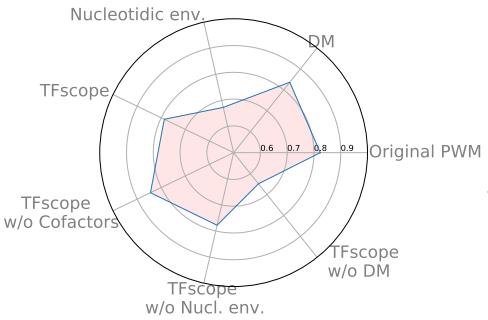
Figure 6 — Analysis of the binding differences induced by a specific treatment A-B Radar plot and location plot of the most important variables in the NR3C1 comparison between Dex and Dex+TNF treatments. **C-D** Radar plot and location plot of the most important variables in the RELA comparison between TNF and Dex+TNF treatments.



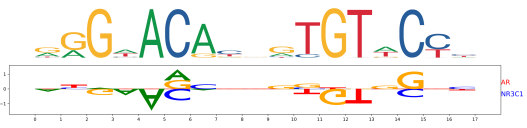
A



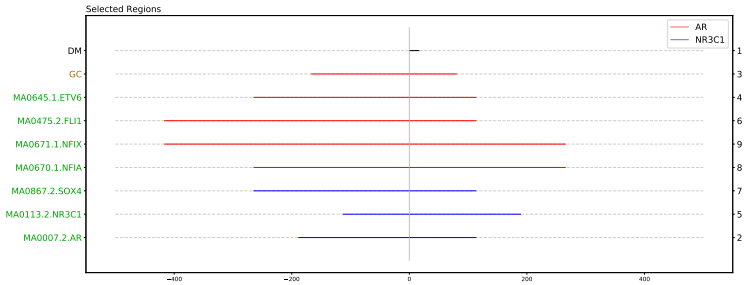
B



C



D



E

Figure 7 — Analysis of the binding differences of paralogous TFs A Radar plot of the most important variables discriminating FOSL1 and FOLS2 binding sites. **B** JASPAR FOSL1 motif (up) and TFscope-DM motif discriminating FOSL1 and FOLS2 binding sites (down) **C** Radar plot of the most important variables discriminating AR and GR binding sites. **D** JASPAR AR motif (up) and TFscope-DM motif discriminating AR and GR (NR3C1) binding sites (down). **E** Location plot of variables discriminating AR and GR binding sites.

Annexe B

Cassan, O., Lèbre, S., Martin.,A (2021)

Cassan *et al. BMC Genomics* (2021) 22:387
https://doi.org/10.1186/s12864-021-07659-2

BMC Genomics

SOFTWARE Open Access

Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite

Océane Cassan[1*], Sophie Lèbre[2,3] and Antoine Martin[1]

Abstract

Background: High-throughput transcriptomic datasets are often examined to discover new actors and regulators of a biological response. To this end, graphical interfaces have been developed and allow a broad range of users to conduct standard analyses from RNA-seq data, even with little programming experience. Although existing solutions usually provide adequate procedures for normalization, exploration or differential expression, more advanced features, such as gene clustering or regulatory network inference, often miss or do not reflect current state of the art methodologies.

Results: We developed here a user interface called DIANE (Dashboard for the Inference and Analysis of Networks from Expression data) designed to harness the potential of multi-factorial expression datasets from any organisms through a precise set of methods. DIANE interactive workflow provides normalization, dimensionality reduction, differential expression and ontology enrichment. Gene clustering can be performed and explored via configurable Mixture Models, and Random Forests are used to infer gene regulatory networks. DIANE also includes a novel procedure to assess the statistical significance of regulator-target influence measures based on permutations for Random Forest importance metrics. All along the pipeline, session reports and results can be downloaded to ensure clear and reproducible analyses.

Conclusions: We demonstrate the value and the benefits of DIANE using a recently published data set describing the transcriptional response of Arabidopsis thaliana under the combination of temperature, drought and salinity perturbations. We show that DIANE can intuitively carry out informative exploration and statistical procedures with RNA-Seq data, perform model based gene expression profiles clustering and go further into gene network reconstruction, providing relevant candidate genes or signalling pathways to explore. DIANE is available as a web service (https://diane.bpmp.inrae.fr), or can be installed and locally launched as a complete R package.

Keywords: Gene regulatory network inference, Graphical user interface, Multifactorial transcriptomic analysis, Model-based clustering, Analysis workflow

*Correspondence: oceane.cassan@cnrs.fr
[1]BPMP, CNRS, INRAE, Institut Agro, Univ Montpellier, 34060 Montpellier, France
Full list of author information is available at the end of the article

Background

Analyzing gene expression to uncover regulatory mechanisms

A multitude of regulatory pathways have evolved in living organisms in order to properly orchestrate development, or to adapt to environmental constraints. Much of these regulatory pathways involve a reprogramming of genome expression, which is essential to acquire a cell identity corresponding to given internal and external environments. To characterize these regulatory pathways, and translate these changes in gene expression at the genome-wide level, global transcriptome study under various species, tissues, cells and biological conditions has become a fundamental and routinely performed experiment for biologists. To do so, sequencing of RNA (RNA-Seq) is now the most popular and exploited technique in next-generation sequencing (NGS) methods, and underwent a great expansion in the field functional genomics. RNA-seq will generate fragments, or short reads, that match to genes and quantitatively translate their level of expression. Standard analysis pipelines and consensus methodological frameworks have been established for RNA-Seq. Following quality control of data, reads mapping to a reference genome, and quantification on features of interest are performed, several major steps are commonly found in RNA-Seq data analysis. They usually consist in proper sample-wise normalization, identification of differential gene expression, ontology enrichment among sets of genes, clustering, co-expression studies or regulatory pathways reconstruction.

However, these analysis procedures often require important prior knowledge and skills in statistics and computer programming. In addition, tools dedicated to analysis, exploration, visualization and valorization of RNA-Seq data are very often dispersed. Most of RNA-Seq data are therefore not properly analyzed and exploited at their highest potential, due to this lack of dedicated tools that could be handled and used by (almost) anyone.

Current tools for facilitating the exploitation of RNA-seq data

Over the last few years, several tools have emerged to ease the processing of RNA-Seq data analysis, by bringing graphical interfaces to users with little programming experience. Among those tools are DEBrowser [1], DEApp [2], iGEAk [3], DEIVA [4], Shiny-Seq [5], IRIS-DEA [6], iDEP [7], or TCC-GUI [8]. All of them propose normalization and low count genes removal, exploratory transcriptome visualizations such as Principal Component Analysis (PCA), and per-sample count distributions plots. They also provide functions for interactive Differential Expression Analysis (DEA) and corresponding visualizations such as the MA-plot. Gene Ontology (GO) enrichment analysis can be performed in those applications, apart from IRIS-DEA, DEApp, and TCC-GUI.

However, when it comes to further advanced analyses such as gene expression profiles clustering or network reconstruction, solutions in those tools are either absent, or sub-optimal in terms of statistical framework or adequacy with certain biological questions. For instance, most of those applications perform clustering using similarity based methods such as k-means and hierarchical clustering, requiring both the choice of metric and criterion to be user-optimized, as well as the selection of the number of clusters. Probabilistic models such as Mixture Models are a great alternative [9–11], especially thanks to their rigorous framework to determine the number of clusters, but they are not represented in currently available tools.

Regarding Gene Regulatory Networks (GRN) inference, only three of the applications cited above propose a solution. Two of them, iDEP and Shiny-Seq rely on the popular WGCNA framework (WeiGhted Correlation Network Analysis) [12], which falls into the category of correlation networks. This inference method have the disadvantage of being very vulnerable to false positives as it easily captures indirect or spurious interactions. When the number of samples in the experiment is low or moderate, high correlations are often accidentally found [13]. Besides, linear correlations like Pearson coefficient can miss complex non-linear effects. Lastly, WGCNA addresses the question of co-expression networks, more than GRN. To infer GRN, which should link Transcription Factors (TF) to target genes, iGEAK retrieves information from external interaction databases and binding motives. This allows to exploit valuable information, but makes this step extremely dependent on already publicly available datasets. An exhaustive comparison with respect to the features and methods handled by the described interfaces for RNA-Seq analysis is given in Fig. 1.

Other frameworks focus on gene network reconstruction and visualization only. For instance, the web server GeNeCK [14] makes the combination of several probabilistic inference strategies easily available, but there is no possibility to select a subset of genes to be considered as regulators during inference. The online tool ShinyBN [15] performs Bayesian network inference and visualization. This Bayesian approach is however prohibitive when large scale datasets are involved. Lastly, neither ShinyBN nor GeNecK allow for upstream analyses and exploration of RNA-Seq expression data.

Consequently, efficient statistical and machine learning approaches for GRN inference (like for instance GENIE3 [16], TIGRESS [17], or PLNModels [18], see [19] for a review) are not available, to our knowledge, as a graphical user interfaces allowing necessary upstrem operations like normalization or DEA.

	DEBrowser	iDEP	Genavi	iGEAK	TCC-GUI	ShinySeq	IRIS-EDA	DEApp	DIANE
Normalisation-filtering									
PCA-MDS									
Distributions plot									
Differential expression analysis									
MA-volcano plots									
GO enrichment analysis									
Expression based gene clustering	Non parametric approaches: k-means, hierarchcal clustering on heatmaps. None or limited parametrization for models/number of clusters.								
Clusters advanced exploration									
Network inference		WGCNA		binding databases		WGCNA + binding			
Network analysis and statistics									
Module detection and analysis									
Reports generation									
WEB Deployment									
Local use		Not free							

Sample homogeneity and exploration
Comparing transcriptomes
Clustering genes
Pathways reconstruction
Ease of use / reproducibility

Feature implemented
Feature implemented but room for improvment (insufficient tuning possibilities, sub-obtimal methodology)
Feature is absent

Fig. 1 Comparison of tools for facilitating the valorization of expression datasets. Eight interactive tools for analysis of count data from RNA-Seq are presented here and compared in terms of features and methodological choices. The features included are the ones we believe are the expectation from most users willing to exploit RNA-Seq experiments and understand regulatory mechanisms, and that we included to DIANE. Although not reported here for clarity reasons, many compared tools had their own features and specificities of interest. For instance, IRIS-DEA handles single cell RNA-Seq and facilitates GEO submission of the data, iDEP enables to build protein-protein interaction network and has an impressive organisms database, while Shiny-Seq can summarise results directly into power point presentations

Besides, all of the cited applications are available as online tools or as local packages with source code, although the useful possibility to provide both solutions simultaneously, in order to satisfy advanced users as much as occasional ones, is not always available. It is also worth noting that availability of organisms in current services varies a lot. Some of them like iGEAK are restricted to human or mouse only.

Proposed approach

In this article, we propose a new R-Shiny tool called DIANE (Dashboard for the Inference and Analysis of Networks from Expression data), both as an online application and as a fully encoded R package. DIANE performs gold-standard interactive operations on RNA-Seq datasets, possibly multi-factorial, for any organism (normalization, DEA, visualization, GO enrichment, data exploration, etc.), while pushing further the clustering and network inference possibilities for the community. Clustering exploits Mixture Models including RNA-seq data prior transformations [11] and GRN inference uses Random Forests [16, 20], a non-parametric machine learning method based on a collection of regression trees. In addition, a dedicated statistical approach, based on both the biological networks sparsity and the estimation of empirical p-values, is proposed for the selection of the edges. Step-by-step reporting is included all along the analyses, allowing reproducible and traceable experiments.

In order to illustrate the different features of DIANE, we have used a recently published RNA-seq data set, describing the combinatorial effects of salt (S), osmotic (M), and heat (H) stresses in the model plant *Arabidopsis thaliana* [21]. RNA-seq were performed under single (H, S, M), double (SM, SH, MH), and triple (SMH) combinations of salt, osmotic, and heat stresses. In the course of our paper, we will demonstrate that DIANE can be a simple and straightforward tool to override common tools for transcriptome analyses, and can easily and robustly lead to GRN inference and to the identification of candidate genes.

Implementation and results

DIANE is an R Shiny [22, 23] application available as an online web service, as well as a package for local use. To perform relevant bioinformatic and bio-statistical work, different existing CRAN and Bioconductor packages as well as novel functions are brought together. Its development was carried out via the golem [24] framework, allowing a modular and robust package-driven design for complex production-grade Shiny applications. Each main feature or analysis step is programmed as a shiny module, making use of the appropriate server-side functions. In the case of local use, those functions are exported by the package so they can be called from any R script to be part of an automated pipeline or more user-specific analyses. We also provide a Dockerfile [25] and instructions so that interested users can deploy DIANE to their own team servers. Figure 2 presents the application workflow and main possibilities. The analysis steps in DIANE are shown in a sequential order, from data import, pre-preprocessing and exploration, to more advanced studies such as co-expression or GRN inference.

Fig. 2 DIANE's workflow. The main steps of the pipeline available in the application -data import, normalization, exploration, differential expression analyses, clustering, network inference- alongside with some chosen visual outputs

Data upload

Expression file and design

To benefit from the vast majority of DIANE's features, the only required input is an expression matrix, giving the raw expression levels of genes for each biological replicate across experimental samples. It is assumed that this expression matrix file originates from a standard bioinformatics pipeline applied to the raw RNA-Seq fastq files. This typically consists in quality control followed by reads mapping to the reference genome, and quantification of the aligned reads on loci of interest.

Organism and gene annotation

Several model organisms are included in DIANE to allow for a fast and effortless annotation and pathway analysis. For now, automatically recognized model organisms are *Arabidopsis thaliana*, *Homo sapiens*, *Mus musculus*, *Drosophilia melanogaster*, *Ceanorhabditis elegans*, and *Escherichia coli*. DIANE takes advantage of the unified annotation data for those organisms offered by the corresponding Bioconductor organisms database packages [26–31]. Other plant species are annotated such as white lupin, and users can easily upload their custom files to describe any other organism whenever it is needed or possible along the pipeline. Organism specific information needed can be common gene names and descriptions, gene - GO terms associations, or known transcriptional regulators.

Normalization and low count genes removal

DIANE proposes several strategies of normalization to account for uneven sequencing depth between samples. One step normalization can be performed using either the Trimmed Mean of M values method (TMM) [32] or the median of ratios strategy from DESeq2 [33]. The TCC package [34] also allows to perform a prior DEA to remove

potential differentially expressed genes (DEG), and then compute less biased normalization factors using one of the previous methods. DIANE also includes a user-defined threshold for low-abundance genes, which may reduce the sensitivity of DEG detection in subsequent analyses [35]. The effect of normalization and filtering threshold on the count distributions can be interactively observed and adjusted.

Exploratory analysis of RNA-seq data

PCA - MDS

Dimensionality reduction techniques are frequently employed on normalized expression data to explore how experimental factors drive gene expression, and to estimate replicate homogeneity. In particular, the Multi-Dimensional Scaling (MDS) plot takes samples in a high dimensional space, and represents them as close in a two-dimensional projection plane [36] depending on their similarity. Principal Component Analysis (PCA) is also a powerful examination of expression data. Through linear algebra, new variables are built as a linear combination of the initial samples, that condense and summarize gene expression variation. By studying the contribution of the samples to each of these new variables, the experimenter can assess the impact of the experimental conditions on gene expression. DIANE offers those two features on expression data, where each gene is divided by its mean expression to remove the bias of baseline expression intensity.

As presented in Fig. 3a, we applied PCA to the normalized transcriptomes after low gene counts removal. No normalization was applied in DIANE as raw data

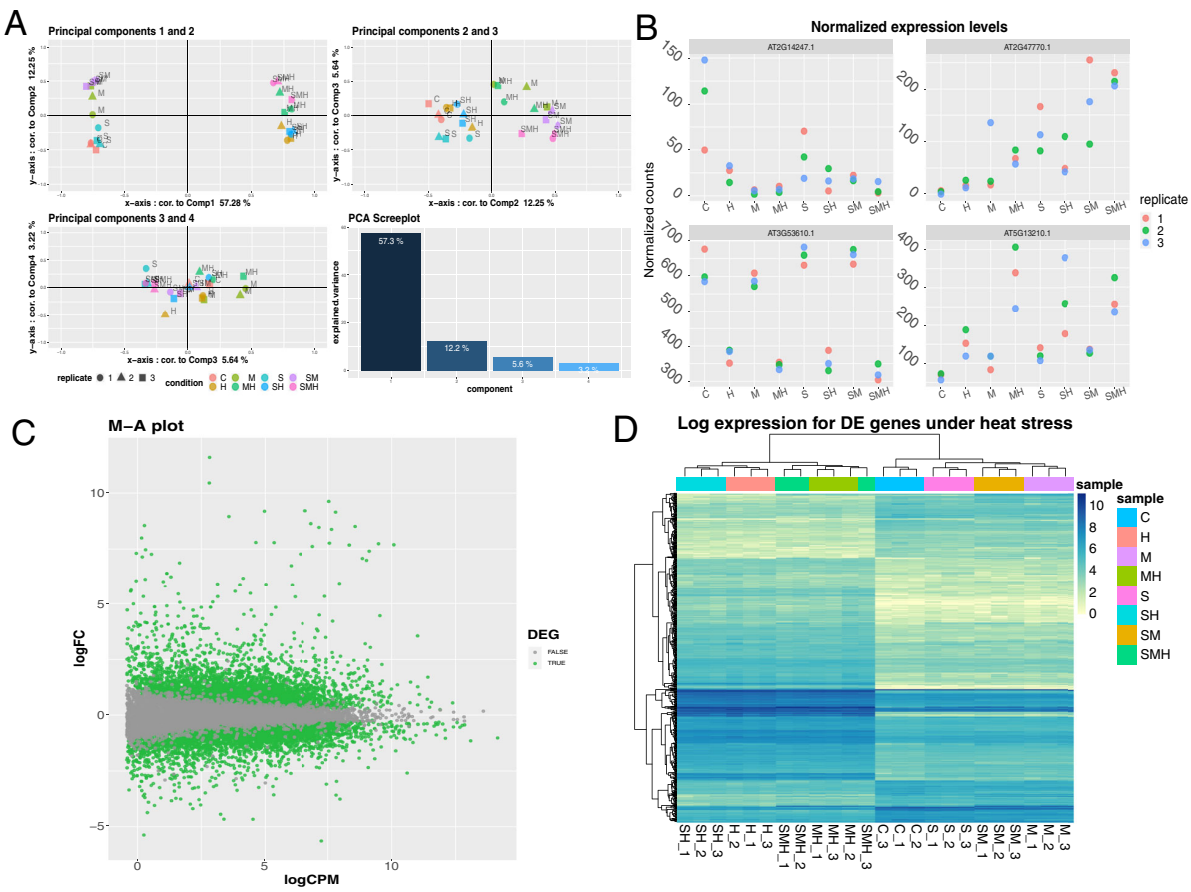


Fig. 3 Normalization and exploration of RNA-seq dataset with DIANE. **a** PCA analysis for the normalized expression table. The experimental conditions have for coordinates their contributions (correlations) to the first four principal components. The scree-plot shows, for each principal component, the part of global variability explained. **b** Example of normalized gene expression levels across all seven perturbations and control. **c** MA-plot for the DEG in response to a single heat stress. The x-axis is the average expression, and the y-axis is the LFC in expression between heat stress and control. DEG with FDR < 0.05 and an absolute LFC > 2 appear in green. **d** Log normalized expression heatmap for the DEG under heat across all perturbations and control

was presented as Tags Per Millions. We found consistent conclusions regarding how heat, salinity and osmotic stresses affect gene expression. The first principal component, clearly linked to high temperature, discriminates the experimental conditions based on heat stress while explaining 57% of the total gene expression variability. The second principal component, to which mannitol-perturbed conditions strongly contributes, accounts for 12% of gene expression variability. The effect of salinity is more subtle and can be discerned in the third principal component.

Normalized gene expression profiles

The "expression levels" tab of the application is a simple exploratory visualization, that allows the user to observe the normalized expression levels of a several genes of interest, among the experimental conditions of its choice. Each replicate is marked as different shapes. Besides rapidly showing the behavior a desired gene, it can provide valuable insights about a replicate being notably different from the others.

Using this feature of DIANE, we represented in Fig. 3b four genes showing different behaviors in response to the combination of stresses, and illustrating the variation that can be found among biological replicates.

Differential expression analysis

DEA in DIANE is carried out through the EdgeR framework [37], which relies on Negative Binomial Modelling. After gene dispersions are estimated, Generalized Linear Models are fitted to explain the log average gene expressions as a linear combination of experimental conditions. The user can then set the desired contrasts to perform statistical tests comparing experimental conditions. The adjusted p-value (FDR) threshold and the minimal absolute Log Fold Change (LFC) can both be adjusted on the fly. A data table of DEG and their description is generated, along with descriptive graphics such as MA-plot, volcano plot, and interactive heat-map. The result DEG are stored to be used as input genes for downstream studies, such as GO enrichment analysis, clustering or GRN inference.

Figure 3c and d represent DEG under heat perturbation. Selection criteria were adjusted p-values greater than 0.05, and an absolute log-fold-change over 2. The 561 up-regulated genes and 175 down-regulated genes are indicated in green in the MA-plot, and correspond to the rows of the heatmap. The high values of LFC for those genes, along with their expression pattern in the heatmap across all conditions confirm the strong impact of heat stress on the plants transcriptome.

In the case where several DEA were performed, it might be useful to compare the resulting lists of DEG. DIANE can perform gene lists intersection, and provide visualizations through Venn diagrams, as well as the possibility to download the list of the intersection. This feature is available for all genes, or specifically for up or down regulated genes.

GO enrichment analysis

Among a list of DEG, it is of great interest to look for enriched biological processes, molecular functions, of cellular components. This functionality is brought to DIANE by the clusterProfiler R package [38], that employs Fischer-exact tests on hypergeometric distribution to determine which GO terms are significantly more represented. Results can be obtained as a downloadable data table, a dotplot of enriched GO terms with associated gene counts and p-values, or as en enrichment map linking co-occurring GO terms.

Gene clustering

Method

In order to identify co-expressed genes among a list of DEGs, DIANE enables gene expression profiles clustering using the statistical framework for inferring mixture models through an Expectation-Maximisation (EM) algorithm introduced by [9, 10]. We chose to use the approach implemented in the Bioconductor Coseq package [11]. Coseq makes it possible to apply transformation to expression values prior to fitting either Gaussian or Poisson multivariate distributions to gene clusters. A penalized model selection criterion is then used to determine the best number of clusters in the data. With DIANE, users simply have to select which DEG should be clustered among previously realized DEA, the experimental conditions to use for clustering, as well as the range of number of clusters to test.

Exploring the clusters

Once clustering was performed, a new tab enables a detailed exploration of the created clusters. It includes interactive profiles visualization, downloadable gene data table, GO enrichment analysis. In addition, if the experimental design file was uploaded, Poisson generalized linear models are fitted to the chosen cluster in order to characterize the effect of each factor on gene expression.

To validate and extend the work done around our demonstration dataset, we performed clustering analysis similarly to what was done in the original paper [21]. We considered all genes from the seven DEA computed between control and perturbation treatments, with a 0.05 FDR threshold and an absolute LFC above 2.

Figure 4 presents the clusters of interest as given by the Poisson Mixtures estimation. They provide a gene partitioning representative of all behaviors in the dataset. In particular, we found that the 3 biggest clusters (2, 3, 6) were composed of heat responsive genes. Among those clusters, statistically enriched GO terms are in majority

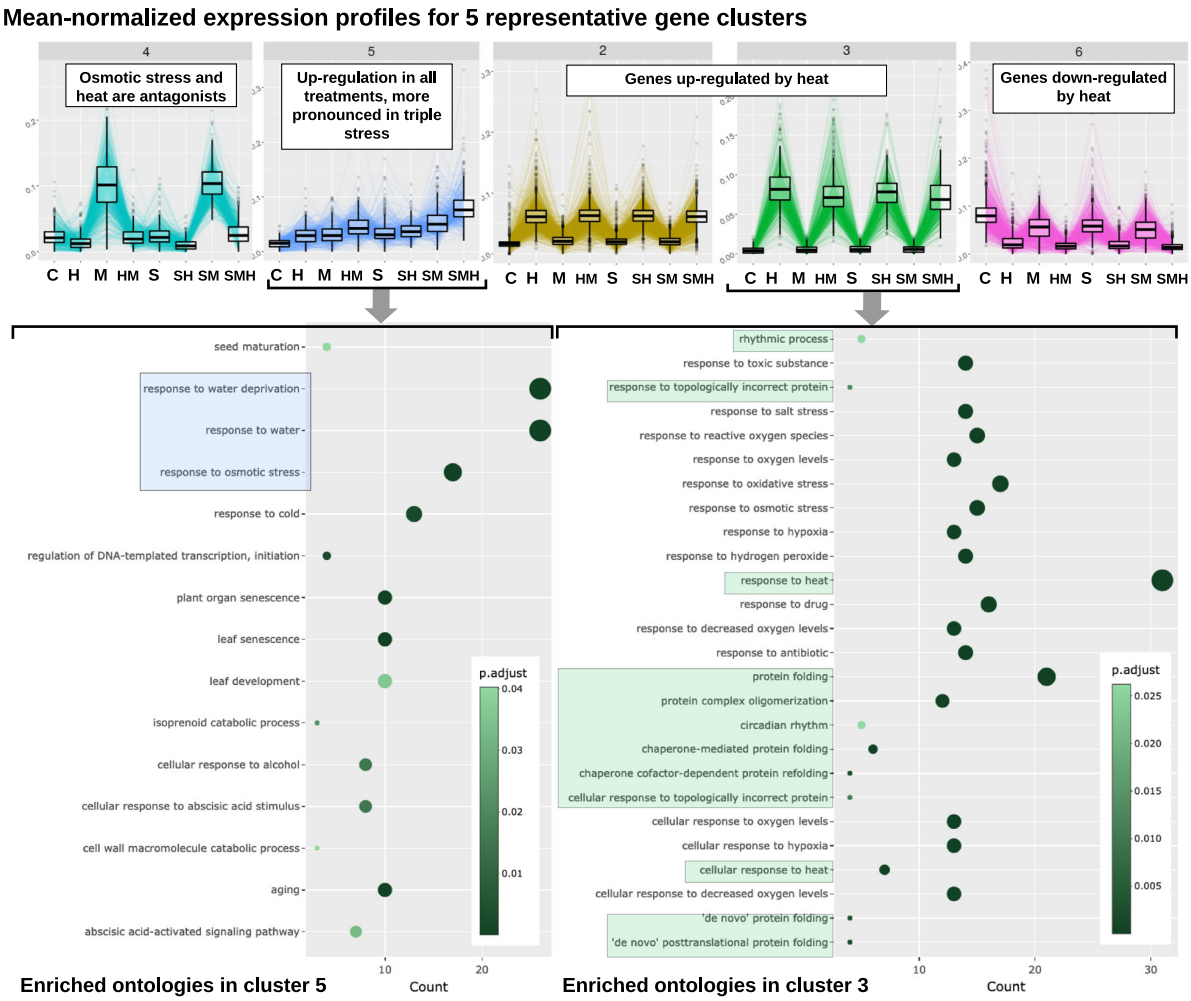


Fig. 4 Clustering of combinatorial RNA-seq data with DIANE. Clusters of interest generated by Coseq in DIANE. Gene expression profiles are defined as the normalized expression divided by the mean normalized expression across all conditions. Graphical results of ontology enrichment analysis are presented for clusters 3 and 5. Highlighted ontologies are relevant categories in line with previously published findings [21]. Ontology enrichment plots show detected GO terms (under 0.05 in Fischer's exact tests), color-coded by their adjusted *p*-value, and shifted in the x-axis depending on the number of genes matching this ontology

linked to heat and protein conformation. Indeed, proteins misfolding and degradation are direct consequences of high temperatures, thus requiring rapid expression reprogramming to ensure viable protein folding in topology control [39]. Two enriched ontologies involved in rhythmic and circadian processes also support evidence for disrupted biological clock. Second, the cluster 5 brings together genes up-regulated in all stress treatments, with the highest induction being observed in the combination of the three perturbations. Those genes, also noted in [21] to exhibit a synergistic response to mannitol and salt, contain three ontologies related to osmotic stress and water deprivation. Lastly, cluster 4 corroborates the existence of genes characterized by opposite reactions to osmotic stress and heat. They are specifically induced in all mannitol perturbations, except under high temperature, where they are strongly repressed.

Gene regulatory network inference

GRN inference is a major contribution of DIANE compared to similar existing applications, the latter offering either no possibility for such task, or either limited ones, as described in the "Background" section.

Estimating regulatory weights

GRN inference aims to abstract transcriptional dependencies between genes based on the observation of their resulting expression patterns. Each gene is represented by

a node in the network. The aim is to recover a weight associated with each edge (i.e. pair of nodes). This is a complex retro-engineering process, challenged by the Curse of dimensionality. Many methods are available, and can be divided into two main categories : statistical and data-driven approaches [13]. Statistical strategies rely on assumptions regarding the data distribution, whose parameters are estimated by maximum-likelihood techniques, often in the case of Bayesian [40] or Lasso inference [17, 41]. However, the underlying modelling assumptions may be inaccurate or difficult to verify in practice. In the second category, the objective is to quantify interaction strengths between pairs of nodes directly from the data. This is typically achieved by using similarity measures such as correlation [12], information theory metrics [42, 43], or feature importances extracted from regression contexts [16]. This second category is less restrictive in terms of hypothesis. However, once the inference is performed, the problem of defining a threshold above which an interaction will be part of the network is far from easy.

There is a large variety of tools available for the task of network inference. Many of them have been benchmarked against one another at the occasion of the DREAM challenges [44, 45]. Those challenges aim at comparing state of the art network inference methods on both simulated and validated biological data. They provide performance metrics for 27 methods based on regression techniques, mutual information metrics, correlation or Bayesian framework among other methods. The performance metrics gathered by DREAM5 [45] (i.e Area Under Precision and Recall curves or overall scores), as well as more recent efforts to compare new methods on those gold standards (i.e F-meausres, ROC curves) are useful resources to help making a choice. For example, existing methods to learn GRN structures are WGCNA [12], ARACNE, CLR, TIGRESS, GENIE3 (see [45] for an exhaustive and referenced list of methods), or also SORDER [46] or CMI2NI [47].

In DIANE, the package chosen for GRN reconstruction is GENIE3 [16], a machine learning procedure that was among the best performers of the DREAM challenges. GENIE3 uses Random Forests [20] which is a machine learning method based on the inference of a collection of regression trees. It has the advantage of being a non-parametric procedure, requiring very few modelling or biological priors, while being able to capture interactions and high order combinatorics between regulators. After having defined a set of regulators among the genes under study, the regression framework allows to infer oriented edges from regulators to targets. With GENIE3, for each target gene, a Random Forest determines the predictive power of each regulator on the target gene expression. The regulatory interactions can then be thresholded according to their importance, so that the strongest links are kept to build a sparse final network. However, choosing such a threshold is not trivial, left as an open question by GENIE3's authors and ever since.

Selecting meaningful regulatory weights

Proposed approach To avoid the unsatisfying hard-thresholding solution, some researchers make use of TF binding experiments, TF-perturbation assays, or literature data to select a threshold influence measure maximizing prediction precision [48–50]. Network backboning [51, 52] and BRANE Cut [53] are mathematical frameworks that try to extract an informative structure from weighted fully connected networks, but they rely on mathematical modelling and assumptions that we suppose might be too strong or not valid in the precise case of gene regulatory network topology. Feeling the lack of an appropriate model-agnostic strategy with no need for external data, we conceived a method that provides a statistical testing framework for weighted regulator-gene pairs. The main steps of the method, as schematized in Fig. 5, are:

Inference of the importance values for all regulator-target gene pairs using Random Forests according to GENIE3's strategy [16] on a chosen list of DEG as input. Transcriptional regulators with a very high value of non linear correlation (typically 0.9 or 0.95) can lead to spurious or missed connections in the final network, and cause robustness issues during the regression procedure. DIANE allows to group them together and to consider them as unique genes.

Selection of the strongest inferred regulatory influences. As biological networks are known for their pronounced sparsity [54–56], testing all possible regulator-target pairs would be of very little interest, as well as a waste of computation time. We therefore create a first graph, topologically consistent with biological network density standards, which will be further refined by statistical tests.

Empirical *p*-values are computed for the selected regulatory weights. To assess weather the importance value of a pair is significant or not, the rfPermute package [57] fits Random Forests and repeatedly shuffles the target gene expression profile so that the null distribution of each regulator influence is estimated. Hence, the empirical *p*-value of a regulator-gene pair is given by the extremeness of its importance as compared to the estimated null distribution. For a faster and more exploratory-oriented network inference, it is possible to skip edges testing (this step and the following).

FDR correction for multiple testing [58] is applied to the *p*-values, and only the edges above an FDR threshold are kept to form the final network. After edges statistical testing, graphics that show the *p*-values distribution and the final number of edges depending on the FDR choice

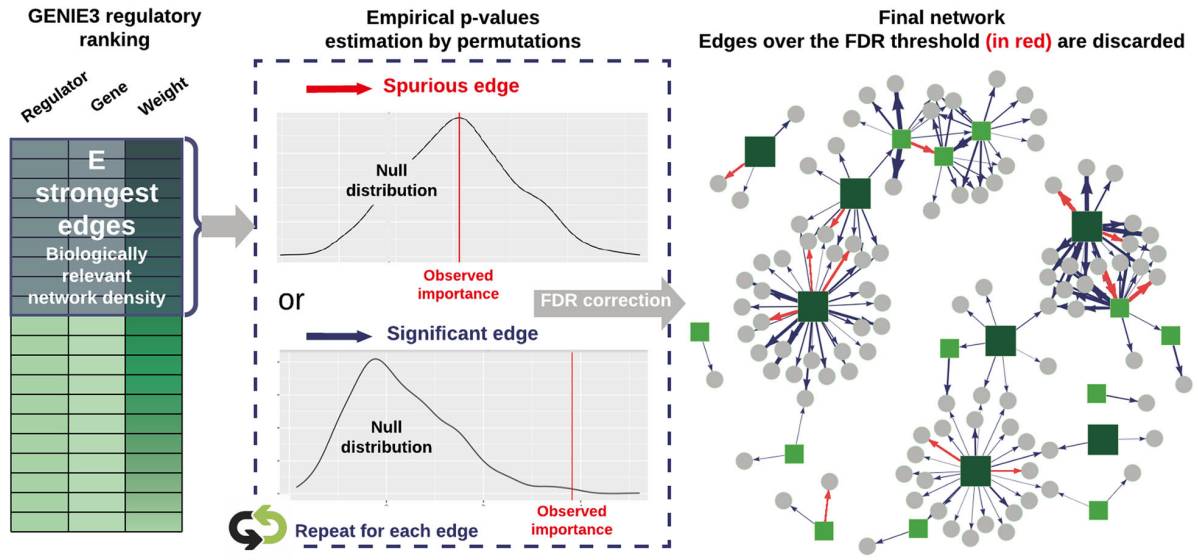


Fig. 5 Statistical testing procedure for edges selection among weighted regulatory links. A first network is built from the influence measures resulting from network inference, by choosing an appropriate network connectivity density. The statistical significance of all its edges is then assessed by empirical tests based on permutation for Random Forest importance metrics

are displayed, providing the user with additional decision guidance.

See Additional file 1 for more details on the statistical procedure and implementation. Thanks to this procedure, the main user-defined parameters are the network density prior to statistical tests, and the FDR cut-off. Together, they bring much more biological meaning and decision help than an arbitrary importance threshold.

Benchmark of the proposed approach We benchmark this novel procedure designed to keep the most significant interactions from a complete GRN. As GENIE3's performance was already assessed in several comparative studies, we focus here only on the edges testing strategy, that we compare to a more naive approach, hard thresholding. To do so, we applied our edges selection strategy to GENIE3 edges ranking on two different datasets, for which robust regulator-gene validation information is available.

The first expression dataset is the RNA-Seq experiment on *Arabidopsis thaliana* we present in this article. We inferred a GRN of heat responsive genes in all experimental conditions (1497 genes from C versus H DEA, LFC $\geq$ 1.5, FDR $\leq$ 0.05, containing 118 regulators). To validate the inferred connections, we made use of connecTF [59], a recent database containing regulatory interactions in Arabidopsis thaliana obtained from in vitro and in vivo binding experiments, as well as in planta regulation experiments. We specifically chose to use the interactions in connecTF obtained from CHIP-Seq and TARGET experiments that represent the most robust data in order to validate connections.

The second dataset is an experiment on *Escherichia coli*, generated by the authors of the "Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles" [60]. We restricted ourselves to a subset of this compendium of experimental conditions corresponding to a single combinatorial experiment. In the latter, bacteria were exposed to a control treatment or to norfloaxacin for different amounts of time, for a total of 24 experimental conditions. The 4345 genes of the organism provided in the dataset, containing 154 transcription factors, are used for GRN inference followed by edges testing. In order to validate the connections of the networks generated in DIANE, we used RegulonDB [61], a database of regulatory interactions built from classic molecular biology experiments and more recently high throughput genomics such as CHIP-Seq and gSELEX.

For each organism, we compared the validity of network predictions between two strategies. The first one corresponds to a network obtained by applying a hard threshold to GENIE3's weighted regulatory associations, to achieve a desired network connectivity density. The second strategy corresponds to that same network, but after removing the edges deemed spurious by our empirical testing procedure for edges selection. By doing so, we aim at determining weather refining edges with our testing procedure leads to networks of higher quality.

The performance metric we chose to assess our method's performance is the precision. It is computed as the fraction of edges in the final network that are present in the set of validated interactions, among those for which the regulator possesses validation information in the gold standard (for example, not all regulators were studied in CHIP-Seq nor TARGET experiments, thus are not present in the validated pairs from connecTF).

To provide some parameter exploration, we compare the two strategies for two different initial connectivity densities, and three FDR thresholds to remove spurious interactions. For all the following benchmarks, we used Random Forests made of 1000 trees, and grouped regulators correlated over 90%, as discussed in the previous paragraph "Proposed approach". In order to evaluate robustness while giving an overview of the variability inherent to Random Forest inference and statistical testing by permutations, we launched the two strategies 20 times for each set of parameters and performed non parametric tests for group mean comparisons.

The results are gathered in Fig. 6a and b. They demonstrate that a significant increase of precision can be achieved on both datasets when choosing stringent adjusted *p*-values for edges removal, independently of prior density. This finding supports that *p*-values obtained from permutations on Random Forest importance metrics can allow more confidence in the inferred edges than hard thresholding GENIE3's fully connected network. Figure 6a and b also illustrate the order of magnitude of the number of connections removed by the testing strategy.

After using our empirical testing procedure for edges removal, we stored the number of remaining edges. We then applied hard-thresholding to GENIE3's ranking in order to create networks containing those same number edges. We observed that the precision of such networks was not as high as with our empirical testing

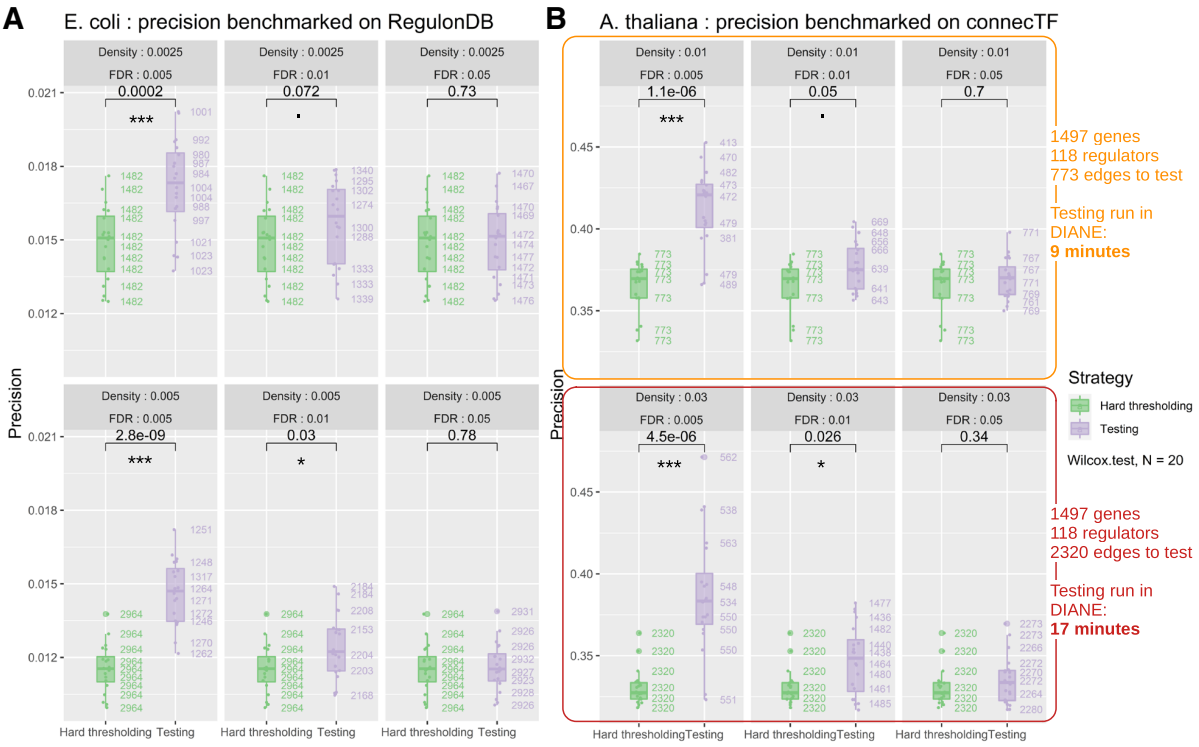


Fig. 6 Benchmark of the proposed testing method on the *E. coli* and *A. thaliana* datasets. Boxplots compare the distributions of precision between hard-thresholding (green) and hard-thresholding followed by the removal of non significant edges as predicted by the testing procedure (purple). The 20 replicates for each configuration provide an estimation of the precision dispersion caused by randomness in GENIE3 and testing by permutations. For each organism, we investigate two appropriate connectivity densities, and three adjusted *p*-value thresholds (FDR). On the right of the boxplots, the number of edges kept in the final network are displayed. *P*-values significance of non parametric mean comparisons between the strategies are encoded as follows : $0 \leq^{***} < 0.001 \leq^{**} < 0.01 \leq^{*} < 0.05 \leq . < 0.1$. The results demonstrate that the proposed testing strategy offers a robust gain in precision when using a stringent adjusted *p*-value threshold for edges removal. **a** Results for the GRN inferred on *E. coli* genes, validated on the regulonDB database. **b** Results for the GRN inferred on *A. thaliana* heat-responsive genes, validated on the connecTF database. Additional metrics about the number of genes, interactions to test, and computation time on DIANE's interface are shown

procedure. This reveals that our adjusted *p*-values bring more information than GENIE3's ranking only, even with a hard-thresholding resulting in the same number of final interactions.

Figure 6b shows computation times required to perform statistical testing on *A. thaliana* dataset, as permitted by DIANE's online interface. DIANE's online version is hosted on a Debian 9.13 server with a 256Go RAM, and 2 Intel(R) Xeon(R) Gold 6130 2.10GHz CPUs. The parallel computing for online use allows up to 16 CPU cores (computation time reported in Fig. 6b uses 16 cores).

Altogether, this benchmarking analysis demonstrates an added-value in terms of network precision when edges selection is performed on the basis of *p*-values rather than by hard thresholding, for a limited time of computation.

Interactive network analysis and community discovery

The last tab of the application is dedicated to network manipulation and exploration. An interactive view of the network is proposed, showing connections between regulatory genes and their predicted targets. By clicking one of the genes, its inward and outward interactions are shown, as well as its annotation and expression profile across samples.

Network-related statistics are automatically generated, delivering topological insights on genes behaviors and network structure. For instance, in and out degree distribution are displayed, and genes can be ranked based on their number of connections. This ranking might then be used for further identification of hub genes and candidate key regulators in the response of interest. In addition, DIANE extracts gene modules, making use of the Louvain algorithm [62]. The experimenter is then free to visualize the results in the network as color-coded communities, while exploring module-specific expression profiles and GO enrichment analyses. At last, it is possible to download edges and node information as csv dataframes, to be further investigated or opened in popular network visualization tools such as Cytoscape.

We used the GRN features of DIANE in order to infer a GRN of the response to heat under osmotic stress, environmental conditions that plants are supposed to face more frequently under climate change circumstances. The input list of genes is obtained in DIANE, by calculating DEG between simple osmotic stress and the double heat-osmotic perturbation (M versus HM, FDR $<$ 0.01, LFC $>$ 2). 640 DEG are detected, among which 363 are up-regulated, 277 are down-regulated, and 45 are transcriptional regulators. Regulators with Spearman correlations over 90% in all available experimental conditions were grouped before network inference, so that a total of 27 regulators are used as predictive variables during inference. For GRN reconstruction, we used Random Forests composed of 4000 trees. A prior network density of 0.03 was defined to select the strongest edges for permutation testing, and edges under a 0.01 FDR were kept in the final network. This network, presented in Fig. 7a, is composed of 289 nodes and 438 edges.

The M versus MH GRN provided by DIANE revealed two interesting groups of regulators, acting as central nodes in their topological modules, and being connected to a large number of target genes.

The most connected regulator of the network is composed by the WRKY47-WRKY8 grouping. Along with other top-ranked WRKY transcription factors (WRKY30, WRKY6, WRKY55), they belong to the topological community of genes that exhibit antagonistic behavior between heat and osmotic stress. The expression values of WRKY8 and WRKY47 in the experiment are presented in Fig. 7b. As already pointed out by our clustering analysis in Fig. 4, those genes undergo a strong induction after mannitol treatment while being repressed by all high temperature conditions. This behavior can also be observed in the intra-module expression profiles in Fig. 7c. Such a module is of high biological interest, as these opposite interactions between drought and high temperature might explain the increased damages observed in the combination of those perturbations [21], and help to understand how heat can suppress the adaptive response of plants to water deficit. Given that WRKY47 and WRKY8 act as a hub in the inferred network, they would be a relevant choice of candidates for experimental pathway validation. Interestingly, WRKY47 has already been identified in rice as a positive regulator of the response to drought [63], strongly reinforcing the validity of the candidate genes from GRN inference in DIANE.

The second most connected node is formed by the regulators TRFL3-AT5G57150-PRR3-BBX8-PIL2-BT1-DREB2B-FRG5-ASY1-ARR15. Those genes, sharing highly correlated profiles across the 24 experimental samples, respond to heat in a clear manner, as well as the other genes inside their community as shown in Fig. 7c. It is worthy to note that PIL2 is a member of a transcription factor family known to be involved in the response to temperature [64] and that DREB2B is a regulator already characterized to act at the interaction between drought and heat stress [65]. The other mentioned regulators offer thus promising leads to be further explored. Three members of the Heat Stress Transcription Factor family (HSFA2 grouped with HSFB2B, and HSFA3) are also found within the genes of the module.

Inside each module, both correlated and anti-correlated expression patterns coexist, which can indicate negative regulation between their gene members. Such opposite variations are captured by the Random Forest algorithm, and allow to go beyond co-expression analysis provided by a clustering approach alone.

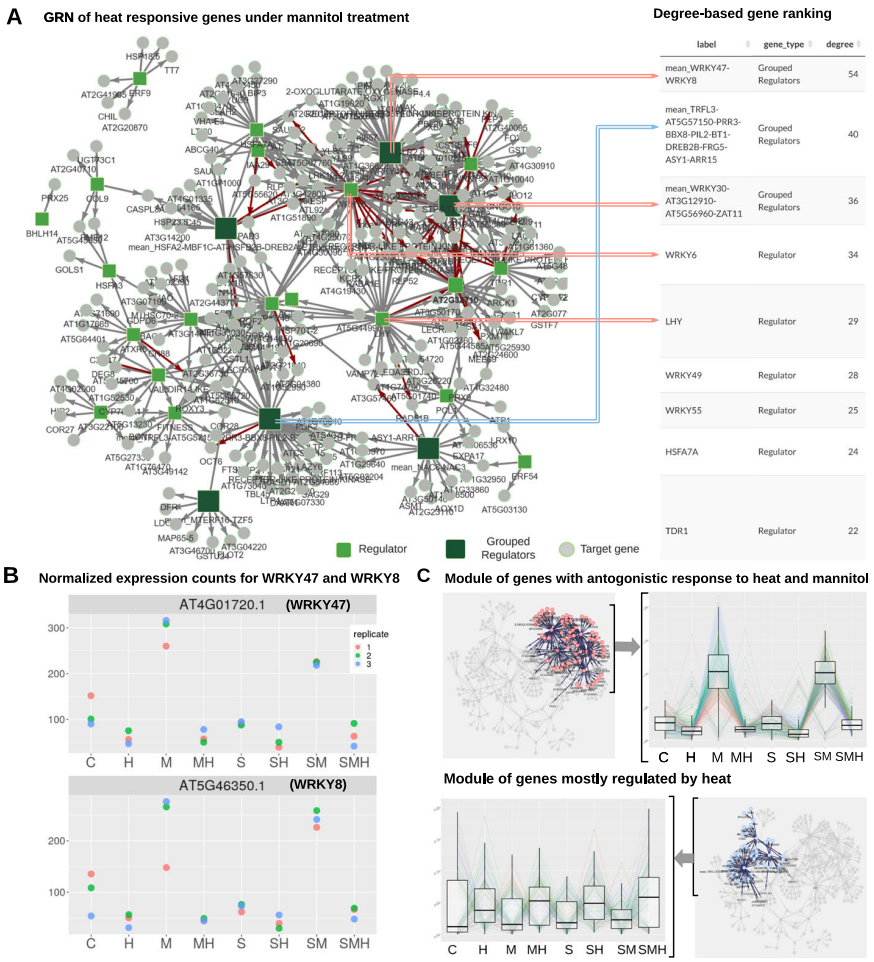


Fig. 7 Network inference and exploration with DIANE. **a** GRN on M versus MH DEG using DIANE's method for GNR inference, and the corresponding degree-based ranking of the nodes. The 11 most connected nodes are presented by order of importance. The regulators mentioned in the network analysis are pointed out by an arrow, the color of the arrow depending on the topological cluster. **b** Expression profiles for WRKY47 and WRKY8, representing the most connected node of the network. **c** Topological modules containing the two most connected groups of regulators are highlighted, juxtaposed to their genes expression profiles

Research reproducibility

For each step of the pipeline, automatically generated reports can be downloaded, rendered on the fly in RMarkdown. They store the users settings, chosen strategies, and display previews of the results. In that way, analysis can be re-run, shared across users, and their settings can be backed-up. The chosen format for those reports is HTML, as it keeps a possibility to interact with data tables, or even manipulate network objects outside of the application. Additional file 2 is an example of report as generated for the network inference described in previous section. Besides, a seed can be set as a global setting of the application, to ensure reproducible runs of the pipeline steps making use of randomness.

Accessibility

DIANE is a tool designed to be as accessible as possible. However, it can be challenging for users with little programming and command line experience to process raw RNA-Seq data into the expression matrix needed in DIANE. Services such as quality control, read mapping and quantification require to handle large files transfers and intensive computations, which are much less easily set up on online applications. However, local programs such as the Tuxedo suite [66], RMTA [67] or GenePattern [68] represent well documented and adequate solutions to most users in order to produce the expression matrices required in DIANE.

Conclusions

To summarise this work, we presented an online graphical user interface to easily conduct in-depth analyses on gene expression data from multi-factorial experiments,

including gene expression profile clustering and GRN inference. It can be downloaded and installed seamlessly as any R package to run the pipeline locally or from R scripts. Given that all other graphical interface tools found in the literature are (i) more oriented toward co-expression rather than regulation and (ii) do not provide recent advanced methodological frameworks for pathway reconstruction, our application positions itself as a tool of first choice to explore regulatory mechanisms.

The demonstration of DIANE on its companion dataset allowed to better understand the effect of combined heat, osmotic and salinity perturbations on Arabidopsis thaliana, consistently with the original analysis [21]. Similar patterns in gene behaviors were highlighted, such as the predominant influence of heat, and its aggravating effect when combined to dehydration. Moreover, DIANE provided new leads through its network inference features : key genes involved in the response to high temperature under drought were pointed out to be promising candidate regulators for improving crops resistance to arid conditions and climate change.

In terms of computational cost, the final step of DIANE's pipeline, i.e. the statistical testing of TF-target edges, could be improved. The R implementations of Random forests and permutations in rfPermute are currently being used, but a C++ version could be envisioned to shorten the method's execution time. Besides, the inference method itself could be subject to improvement in the future. First, combining the results of several inference methods has proven to be as a robust and powerful approach on validated datasets [45, 52]. Second, our strategy is particularly well-suited for multi-factorial and perturbation designs, but is not optimal for time series RNA-Seq. Other inference methods specific to time series RNA-Seq data [69] could be available in DIANE, to bring closer to causality in the inferred transcriptionnal interactions. Lastly, it would be valuable to add further functional features in DIANE, notably in order to integrate external information, such as interaction databases, or data from TF binding or chromatin accessibility experiments.

Availability and requirements

Project name: DIANE
Project home pages: https://oceanecsn.github.io/DIANE https://github.com/OceaneCsn/DIANE
Operating system(s): Platform independent
Programming language: R
Other requirements: Web use : none. Local use: R >4.0.1
License: GNU GPL
Any restrictions to use by non-academics: none

Abbreviations

H: High temperature perturbation M: Mannitol perturbation S: Salinity perturbation SM: Salinity and Mannitol perturbations SH: Salinity and High temperature perturbations MH: Mannitol and High temperature perturbations SMH: Salinity, Mannitol and High temperature perturbations DEA: Differential Expression Analysis DEG: Differentially Expressed Genes DIANE: Dashboard for the Inference and Analysis of Networks from Expression data FDR: False Discovery Rate GENIE3: GEne Network Inference with Ensemble of trees GO: Gene Ontology GRN: Gene Regulatory Network LFC: Log Fold Change NGS: Next-Generation Sequencing PCA: Principal Component Analysis RNA-Seq: Sequencing of RNA TF: Transcription Factors TMM: Trimmed Mean of M values WGCNA: WeiGhted Correlation Network Analysis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07659-2.

Additional file 1: Full description of the procedure of importance measures empirical testing. the files gives more details about the methodological choices for the procedure.

Additional file 2: Network inference report from the M versus MH GRN. Interactive report generated after network inference and edges testing in DIANE. Slight changes might be observed from the textual description of the network because of the stochasticity inherent to the Louvain, Random Forest, and permutations procedures.

Acknowledgements

We thank Alexandre Soriano, Cécile Fizames, Adrien Jarretier-Yuste for help, comments and suggestions during the development of this application. We thank Benjamin Péret for his support in the initial web deployment of DIANE.

Authors' contributions

SL, AM, OC defined the application concepts, searched scientific literature for appropriate methods, tools, biological findings, and redacted the article. SL, OC developed the empirical testing procedure on edges importance measures. AM, OC chose the demonstration dataset, used DIANE on it, and performed biological interpretations. OC carried out the programming and benchmarking of DIANE. All authors have read and approved the manuscript.

Funding

OC, SL and AM are supported by a 80 Prime fellowship from the National Center of Scientific Research (CNRS, France).

Availability of data and materials

The RNA-Seq experiment we included to DIANE for demonstration purposes corresponds to the GEO accession GSE146206 and can be found at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE146206.
The code and benchmark scripts of DIANE are available in the github repositories https://github.com/OceaneCsn/DIANE and https://github.com/OceaneCsn/Benchmarking_DIANE.
DIANE largely relies on the CRAN https://cran.r-project.org/ and Bioconductor https://bioconductor.org/ packages repositories.
The datasets queried to retrieve validated regulatory interactions are connecTF https://connectf.org/ and RegulonDB http://regulondb.ccg.unam.mx/.
The expression data used to infer regulatory networks on *Escherichia coli* were taken from the Many Microbe Microarrays Database at http://m3d.mssm.edu/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]BPMP, CNRS, INRAE, Institut Agro, Univ Montpellier, 34060 Montpellier, France. [2]IMAG, Univ. Montpellier, CNRS, Montpellier, France. [3]Université Paul-Valéry-Montpellier 3, Montpellier, France.

Received: 17 November 2020 Accepted: 28 April 2021
Published online: 26 May 2021

References

1. Kucukural A, Yukselen O, Ozata DM, Moore MJ, Garber M. DEBrowser: Interactive differential expression analysis and visualization tool for count data 06 Biological Sciences 0604 Genetics 08 Information and Computing Sciences 0806 Information Systems. BMC Genomics. 2019;20(1):6. https://doi.org/10.1186/s12864-018-5362-x.
2. Li Y, Andrade J. DEApp: An interactive web interface for differential expression analysis of next generation sequence data. Source Code Biol Med. 2017;12(1):10–3. https://doi.org/10.1186/s13029-017-0063-4.
3. Choi K, Ratner N. IGEAK: An interactive gene expression analysis kit for seamless workflow using the R/shiny platform. BMC Genomics. 2019;20(1):177. https://doi.org/10.1186/s12864-019-5548-x.
4. Harshbarger J, Kratz A, Carninci P. DEIVA: A web application for interactive visual analysis of differential gene expression profiles. BMC Genomics. 2017;18(1):47. https://doi.org/10.1186/s12864-016-3396-5.
5. Sundararajan Z, Knoll R, Hombach P, Becker M, Schultze JL, Ulas T. Shiny-Seq: advanced guided transcriptome analysis. BMC Res Notes. 2019;12(1):432. https://doi.org/10.1186/s13104-019-4471-1.
6. Monier B, McDermaid A, Wang C, Zhao J, Miller A, Fennell A, Ma Q. IRIS-EDA: An integrated RNA-seq interpretation system for gene expression data analysis. PLoS Comput Biol. 2019;15(2):. https://doi.org/10.1371/journal.pcbi.1006792.
7. Ge SX, Son EW, Yao R. iDEP: An integrated text application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics. 2018;19(1):1–24. https://doi.org/10.1186/s12859-018-2486-6.
8. Su W, Sun J, Shimizu K, Kadota K. TCC-GUI: A Shiny-based application for differential expression analysis of RNA-Seq count data. BMC Res Notes. 2019;12(1):133. https://doi.org/10.1186/s13104-019-4179-2.
9. Rau A, Celeux G, Martin-Magniette M-L, Maugis-Rabusseau C. Clustering high-throughput sequencing data with poisson mixture models. [Research Report] RR-7786, INRIA. 2011, p. 36. hal-01193758v2.
10. Rau A, Maugis-Rabusseau C, Martin-Magniette M-L, Celeux G. Co-expression analysis of high-throughput transcriptome sequencing data with poisson mixture models. Bioinformatics. 2015;31(9):1420–7.
11. Rau A, Maugis-Rabusseau C. Transformation and model choice for RNA-seq co-expression analysis. Brief Bioinforma. 2018;19(3):425–36. https://doi.org/10.1093/bib/bbw128.
12. Langfelder P, Horvath S. Wgcna: an r package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1):559.
13. Sanguinetti G, Huynh-Thu VA. Gene regulatory networks. New York: Springer, Humana Press; 2019.
14. Zhang M, Li Q, Yu D, Yao B, Guo W, Xie Y, Xiao G. Geneck: a web server for gene network construction and visualization. BMC Bioinformatics. 2019;20(1):1–7.
15. Chen J, Zhang R, Dong X, Lin L, Zhu Y, He J, Christiani DC, Wei Y, Chen F. shinybn: an online application for interactive bayesian network inference and visualization. BMC Bioinformatics. 2019;20(1):711.
16. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE. 2010;5(9):12776. https://doi.org/10.1371/journal.pone.0012776.
17. Haury A-C, Mordelet F, Vera-Licona P, Vert J-P. Tigress: trustful inference of gene regulation using stability selection. BMC Syst Biology. 2012;6(1):145.
18. Chiquet J, Robin S, Mariadassou M. Variational inference for sparse network reconstruction from count data. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97. PMLR; 2019. p. 1162–71.
19. Mochida K, Koda S, Inoue K, Nishii R. Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets. Front Plant Sci. 2018;9:1770.
20. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
21. Sewelam N, Brilhaus D, Bräutigam A, Alseekh S, Fernie AR, Maurino VG. Molecular plant responses to combined abiotic stresses put a spotlight on unknown and abundant genes. J Exp Bot. 2020.
22. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2020. https://www.R-project.org/.
23. Chang W, Cheng J, Allaire J, Xie Y, McPherson J, et al. Shiny: web application framework for r. R package version 1(5). 2017.
24. Guyader V, Fay C, Rochette S, Girard C. Golem: A Framework for Robust Shiny Applications. 2020. R package version 0.2.1. https://CRAN.R-project.org/package=golem. Accessed 04 May 2021.
25. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux J. 2014;2014(239):2.
26. Carlson M. org.At.tair.db: Genome Wide Annotation for Arabidopsis. 2020. R package version 3.11.4.
27. Carlson M. org.Ce.eg.db: Genome Wide Annotation for Worm. 2020. R package version 3.11.4.
28. Carlson M. org.Dm.eg.db: Genome Wide Annotation for Fly. 2020. R package version 3.11.4.
29. Carlson M. org.EcK12.eg.db: Genome Wide Annotation for E Coli Strain K12. 2020. R package version 3.11.4.
30. Carlson M. org.Hs.eg.db: Genome Wide Annotation for Human. 2020. R package version 3.11.4.
31. Carlson M. org.Mm.eg.db: Genome Wide Annotation for Mouse. 2020. R package version 3.11.4.
32. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):25. https://doi.org/10.1186/gb-2010-11-3-r25.
33. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
34. Sun J, Nishiyama T, Shimizu K, Kadota K. TCC: An R package for comparing tag count data with robust normalization strategies. BMC Bioinformatics. 2013;14(1):219. https://doi.org/10.1186/1471-2105-14-219.
35. Sha Y, Phan JH, Wang MD. Effect of low-expression gene filtering on detection of differentially expressed genes in rna-seq data. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). New York: IEEE; 2015. p. 6461–4.
36. Kruskal JB. Multidimensional Scaling, vol. 11. Thousands Oaks, California: Sage; 1978.
37. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288–97. https://doi.org/10.1093/nar/gks042.
38. Yu G, Wang L-G, Han Y, He Q-Y. clusterprofiler: an r package for comparing biological themes among gene clusters. Omics J Integr Biol. 2012;16(5):284–7.
39. Wang W, Vinocur B, Shoseyov O, Altman A. Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. Trends Plant Sci. 2004;9(5):244–52.
40. Ko Y, Kim J, Rodriguez-Zas SL. Markov chain monte carlo simulation of a bayesian mixture model for gene network inference. Genes Genomics. 2019;41(5):547–55.
41. Omranian N, Eloundou-Mbebi JM, Mueller-Roeber B, Nikoloski Z. Gene regulatory network inference using fused lasso on multiple data sets. Sci Rep. 2016;6:20533.
42. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):8.
43. Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A. Reverse engineering cellular networks. Nat Protoc. 2006;1(2):662.
44. Greenfield A, Madar A, Ostrer H, Bonneau R. DREAM4: Combining genetic and dynamic information to identify biological networks and Dynamical Models. PLoS ONE. 2010;5(10):. https://doi.org/10.1371/journal.pone.0013397.
45. Marbach D, Costello JC, Küffner R, Vega N, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G, Performed GSM. Wisdom of crowds for robust gene network inference the DREAM5 Consortium HHS Public Access. Nat Methods. 2016;9(8):796–804. https://doi.org/10.1038/nmeth.2016.
46. Aghdam R, Ganjali M, Zhang X, Eslahchi C. Cn: a consensus algorithm for inferring gene regulatory networks using the sorder algorithm and conditional mutual information test. Mol BioSyst. 2015;11(3):942–9.

47. Zhang X, Zhao J, Hao J-K, Zhao X-M, Chen L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. Nucleic Acids Res. 2015;43(5):31–31.
48. Anwar M, Tambalo M, Ranganathan R, Grocott T, Streit A. A gene network regulated by FGF signalling during ear development. Sci Rep. 2017;7(1):. https://doi.org/10.1038/s41598-017-05472-0.
49. Shibata M, Breuer C, Kawamura A, Clark NM, Rymen B, Braidwood L, Morohashi K, Busch W, Benfey PN, Sozzani R, Sugimoto K. GTL1 and DF1 regulate root hair growth through transcriptional repression of ROOT HAIR DEFECTIVE 6-LIKE 4 in Arabidopsis. Development (Cambridge). 2018;145(3):. https://doi.org/10.1242/dev.159707.
50. Brooks MD, Cirrone J, Pasquino AV, Alvarez JM, Swift J, Mittal S, Juang C-L, Varala K, Gutiérrez RA, Krouk G, et al. Network walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions. Nat Commun. 2019;10(1):1–13.
51. Coscia M, Neffke FMH. Network backboning with noisy data; 2017. p. 425–436. https://doi.org/10.1109/ICDE.2017.100.
52. Schiffthaler B, Serrano A, Delhomme N, Street NR. Seidr: A toolkit for calculation of crowd networks: Cold Spring Harbor Laboratory; 2018, p. 250696. https://doi.org/10.1101/250696.
53. Pirayre A, Couprie C, Bidard F, Duval L, Pesquet JC. BRANE Cut: Biologically-related a priori network enhancement with graph cuts for gene regulatory network inference. BMC Bioinformatics. 2015;16(1):368. https://doi.org/10.1186/s12859-015-0754-2.
54. Koutrouli M, Karatzas E, Paez-Espino D, Pavlopoulos GA. A Guide to Conquer the Biological Network Era Using Graph Theory. Front Media S.A. 2020. https://doi.org/10.3389/fbioe.2020.00034.
55. Leclerc RD. Survival of the sparsest: Robust gene networks are parsimonious. Mol Syst Biol. 2008;4:. https://doi.org/10.1038/msb.2008.52.
56. Hayes W, Sun K, Pržulj N. Graphlet-based measures are suitable for biological network comparison. Bioinformatics. 2013;29(4):483–91. https://doi.org/10.1093/bioinformatics/bts729.
57. Archer E. rfPermute: Estimate Permutation *p*-values for Random Forest Importance Metrics. 2020. R package version 2.1.81. https://CRAN.R-project.org/package=rfPermute.
58. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol). 1995;57(1):289–300.
59. Brooks MD, Juang C-L, Katari MS, Alvarez JM, Pasquino A, Shih H-J, Huang J, Shanks C, Cirrone J, Coruzzi GM. Connectf: A platform to integrate transcription factor-gene interactions and validate regulatory networks. Plant Physiol. 2020;185(1):49–66. https://doi.org/10.1093/plphys/kiaa012. https://academic.oup.com/plphys/article-pdf/185/1/49/36389080/kiaa012.pdf.
60. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):8.
61. Santos-Zavaleta A, Salgado H, Gama-Castro S, Sánchez-Pérez M, Gómez-Romero L, Ledezma-Tejeida D, García-Sotelo JS, Alquicira-Hernández K, Muñiz-Rascado LJ, Peña-Loredo P, et al. Regulondb v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli k-12. Nucleic Acids Res. 2019;47(D1):212–20.
62. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):10008.
63. Raineri J, Wang S, Peleg Z, Blumwald E, Chan RL. The rice transcription factor oswrky47 is a positive regulator of the response to water deficit stress. Plant Molecular Biol. 2015;88(4-5):401–13.
64. Lin L, Liu X, Yin R. Pif3 integrates light and low temperature signaling. Trends Plant Sci. 2018;23(2):93–5.
65. Lata C, Prasad M. Role of drebs in regulation of abiotic stress responses in plants. J Exp Bot. 2011;62(14):4731–48.
66. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nat Protocol. 2012;7(3):562–78.
67. Peri S, Roberts S, Kreko IR, McHan LB, Naron A, Ram A, Murphy RL, Lyons E, Gregory BD, Devisetty UK, Nelson ADL. Read mapping and transcript assembly: A scalable and high-throughput workflow for the processing and analysis of ribonucleic acid sequencing data. Front Genet. 2020;10:1361. https://doi.org/10.3389/fgene.2019.01361.
68. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. Genepattern 2.0. Nat Genet. 2006;38(5):500–1.
69. Geurts P, et al. dyngenie3: dynamical genie3 for the inference of gene networks from time series expression data. Sci Rep. 2018;8(1):1–12.

Publisher's Note

Annexe C

Bousquet, F. , Lèbre, S., Lavergne, C. (2020)

2017-2020 : Thèse de Faustine Bousquet (2020) `www.theses.fr/2020MONTS095`
Titre : Prédiction des réponses utilisateurs à une campagne de publicité mobile
Financement : Contrat CIFRE avec la société TabMo, Montpellier.
Co-encadrement 50% avec Christian Lavergne (Professeur, IMAG, UPVM).

Résumé : La prédiction du taux de clics (CTR) est l'un des défis majeurs de la publicité en ligne au cours de ces dernières années. L'objectif de notre travail est de répondre à un encart publicitaire disponible via un système d'enchère en proposant la publicité la plus pertinente possible. En d'autres termes : il s'agit d'être en mesure de proposer la bonne publicité à la bonne personne au bon moment. Cet objectif prend en considération deux enjeux principaux. Le premier concerne la caractérisation des données à disposition qui sont de natures volumineuses, hétérogènes et clairsemées. Le second objectif concerne la mise en production du modèle : le modèle doit pouvoir être utilisé en temps réel et son déploiement doit être simple à mettre en œuvre. Nous introduisons ici une nouvelle méthode de prédiction du CTR qui repose sur un mélange de modèles linéaires généralisés (GLM). Nous développons tout d'abord une méthode de clustering basée sur un modèle prenant en considération l'aspect longitudinal (afin d'exploiter l'historique de chaque campagne) et non gaussien (la métrique d'intérêt est un taux) des observations du CTR dans les campagnes publicitaires. Cette étape préliminaire permet de grouper les campagnes ayant des profils similaires et offre ainsi une meilleure description des données. Le package R `binomialMix` implémente cette approche pour le mélange de données binomiales et longitudinales. Par la suite, en s'appuyant directement sur les clusters inférés, nous proposons un modèle prédictif qui permet de répondre au sujet central de notre problématique métier : estimer une probabilité de clic pour toute campagne en temps réel. Plusieurs modèles sont mis en compétition : des modèles naïfs et un modèle simple de GLM sont ainsi comparés à plusieurs modèles qui se basent sur les résultats du clustering. Deux modèles (parmi ceux qui utilisent les résultats du modèle de mélange) se distinguent par leurs performances prédictives. Des expérimentations menées sur données simulées et réelles ont montré l'importance de l'étape préliminaire de classification non supervisée sur la qualité de la prédiction. L'ensemble de ces étapes a ainsi pu être industrialisé et intégré dans le processus d'enchère déjà existant. Cette intégration est la succession d'un ensemble d'étapes : la récupération des données, leur prétraitement, l'estimation des paramètres du mélange à partir de variables explicatives soigneusement choisies et enfin, la mise en place du modèle prédictif. Un dernier travail a permis l'exploitation des prédictions à partir des probabilités de clic obtenues en sortie des modèles prédictifs. Ainsi, nous avons pu prédire le CTR en temps réel sur la plateforme d'enchère et pour chaque espace publicitaire disponible qui y transite. L'analyse des premiers résultats en production montre que, pour certains contextes d'enchère, l'utilisation du modèle prédictif, couplé à l'étape de clustering au préalable, a permis une amélioration significative du taux de clics.

Cette thèse a fait l'objet d'un article publié dans les actes de la conférence internationale ECAI 2020 (Bousquet (2020), disponible en Annexe C), d'une communication orale (JdS 2019), d'un poster (ECML 2018) et d'un package R (BinomialMix).

ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200197

From Mixture of Longitudinal and Non-Gaussian Advertising Data to Click-Through-Rate Prediction

Faustine Bousquet [1,2] and **Sophie Lebre** [3,1] and **Christian Lavergne** [3,1]

Abstract. Click-Through-Rate (CTR) prediction is one of the most important challenges in the advertisement field. Nevertheless, it is essential to understand beforehand the voluminous and heterogeneous data structure. Here we introduce a novel CTR prediction method using a mixture of generalized linear models (GLMs). First, we develop a model-based clustering method dedicated to publicity campaign time-series, i.e. non-Gaussian longitudinal data. Secondely, we consider two CTR predictive models derived from the inferred clustering. The clustering step improves the CTR prediction performance both on simulated and real data. An R package *binomialMix* for mixture of binomial and longitudinal data is available on CRAN.

1 Introduction

1.1 Context

The field of advertising, and more particularly online advertising, has been disrupted by the development and success of Real-Time Bidding (RTB) [20, 21]. This process connects advertisers and publishers in real time and gives them the advantage of personalization via an auction system: the publisher provides a set of information about the ad slot context and the advertiser can decide whether he is interested in the auction or not. This system reduces ineffective targeting. We call "impression" the advert's display on the end user's device. The Click-Through Rate (CTR) is the most common way to estimate the efficiency of an advertising campaign. It measures the ratio of the number of times the user has clicked the ad and the number of times the ad has been displayed. Many statistical challenges emerged from online advertising including CTR prediction [2, 3, 9, 19].
In this paper, we address a real-world online issue from TabMo[2] advertising platform. TabMo is an adtech company managing the campaigns for the advertisers. Its business objective is to provide the best ad slot context for every campaign and increase Key Point of Interest as CTR for advertisers. Constraints coming from production context make the issue interesting in many ways. The data volume is huge. Every second the predictive model has to answer around 1 million auctions with the most adequate advertising campaign among all available. Also, we are in case of rare events: number of click is very low compared to the number of impressions (around 1 click every 1000 impressions). The very imbalanced data makes predictions difficult. Despite all the studies (see Section 1.2) conducted on the prediction of user response in an online advertising context, the prediction of CTR remains an open and still relevant issue.

[1] IMAG Institut Montpellierain Alexander Grothendieck, Univ. Montpellier, CNRS, France
[2] TabMo Labs, Montpellier, France - https://hawk.tabmo.io/
[3] Universit Paul Valry Montpellier 3, France

1.2 Related work

The last three years, Neural Network emerged in the online advertising state-of-art for CTR prediction. At the same time, the dimension of features and the data volume has increase. That is one of the reasons why a lot of research and continuous improvements have been done in model structure Neural Network. [3] developed an hybrid predictive model using both the advantages of linear model and deep network architecture. Predictions from both component are combined to produce a final prediction. In [19], the proposed architecture of the DNN focuses on taking into account interactions between variables beyond order 2. Without the need of a manualy preprocessing data and with a quite simple implementation for this kind of modeling, authors assume that there is a significant decrease of the logloss value compared to a classical DNN architecture. Neural networks take advantage of their multi-layered architecture and achieve good predictive results. However, their complexity makes it difficult to understand the model.
Factorization Machines (FM) [16] models second order polynomial with a latent vector for each feature. The interaction between two features is calculated by the inner product of the latent vector from those variables. The advantage of this method is that it reduces the complexity of the model when interactions are taken into account. A lot of extensions emerged from FM [6, 8] in the context of CTR prediction where capturing the information of feature conjunctions can be relevant. [8] proposed a modeling based on levels features interactions while [6] combined the pertinence of Neural Network with FM in the same architecture. However, until now, the Factorization Machines are mostly used for recommender systems.
Predicting CTR using logistic regression is also one of the most studied models in the literature [2, 9]. Logistic regression models present the advantage of an easy implementation for real-world predictive problem. However, to model more complex and heterogeneous data structures from real case studies, the use of logistic regression may be limited. The use of mixture models allows for better consideration of heterogeneity and data specificity.

Mixture model of Generalized Linear Model (GLM) is a well-studied statistical problem in the literature; [12] gives a complete overview of different existing methodologies for model-based clustering. The number of R packages have grown significantly in various domains of application (such as biology, physics, economics...) to model data with a finite number of subpopulations. *Mixmod* [18] is a popular tool for partitioning observations into groups. The package allows to fit Gaussian mixture models with different volume, shape and orientation of the clusters. It can also model mixtures of multinomial distribution. The *mclust* [5] package is another tool for finite

normal mixture modeling. In the case of longitudinal data, the package *flexmix* [10] provides a mixture modeling when there are M individuals with respectively n_M observations. But to the best of our knowledge, the existing packages cannot model longitudinal data in a binomial context

1.3 Contribution

This paper addresses a real-world online advertising issue. Using the traffic log of a set of campaigns, we build an efficient method that predicts a context specific click probability for each campaign. Our main contributions are the following:

- We describe a GLM-based clustering approach for longitudinal data in a binomial context. Campaigns are clustered according to their CTR depending on the advertising context described by continuous and categorical variables. The longitudinal data is defined as repeated observations for each campaign with a specific length. The temporal periodicity is captured via 2 categorical variables: day and time slot. Time slot separation was established with domain experts. This model-based clustering allows us to group advertising campaigns with similar profiles.
- We predict context specific CTR for each campaign using this model-based clustering as a preliminary step. Various clustered-based prediction schemes are considered.
- Using GLM mixture for CTR prediction leads to good performance while preserving a simple model architecture and a rapid deployment process. Experiments are performed on data extracted from a real-world online advertising context.
- We developed an R package: *binomialMix* (available on CRAN) implementing a GLM-based clustering approach for longitudinal data in a binomial context.

1.4 Outline

Section 2 presents the two-step statistical modeling and the resulting implementation. The first step describes the mixture of binomial for longitudinal data estimated with an Expectation-Maximization (EM) algorithm. The second step builds a predictive model to provide a probability of click using the clustering step. In Section 3, we (i) present the dataset, the evaluation metrics, the results on simulated data and (ii) challenge five predictive models on real data. Three are considered as baseline. The two others use mixture model estimations to predict a probability of click in a given context. We evaluate the relevance of clustering and the performance of the predictions.

2 Proposed Approach

2.1 A binomial model for the CTR

The proposed approach to address the problem of CTR prediction is described in two steps. First, as the observed metric is the click ratio (CTR), we propose a mixture model of Generalized Linear Model (GLM) to model longitudinal data. Then, taking into account the resulting ad campaigns clusters, we develop a methodology to predict a probability of click. The proposed model describes each advertising campaign c as longitudinal data. The CTR is the target variable.

Each day is divided into H time slots. Each campaign c is observed J_c days and some slots could be missing. Let us consider Y_{cjh} the number of clicks for each campaign c at a specific time slot (j,h), for $j = 1,\dots,J_c$ and $h = 1,\dots,H$. We assume that each variable Y_{cjh} follows a distribution of the exponential family as introduced by [11], defined in Equation (1).

$$\forall c = 1,\dots,C, \forall j = 1,\dots,J_c, \forall h = 1,\dots,H$$
$$f_{Y_{cjh}}(y_{cjh},\theta_{cjh},\psi) = \exp\left(\frac{y_{cjh}\theta_{cjh} - b(\theta_{cjh})}{a_{cjh}(\psi)} + d(y_{cjh},\psi)\right) \quad (1)$$

where θ_{cjh} is the canonical parameter, ψ the dispersion parameter. b and d are specific functions depending on the exponential distribution chosen. We can define a_{cjh} as $a_{cjh}(\psi) = \frac{\psi}{\omega_{cjh}}$ with ω_{cjh} a weighted parameter for each observation.
Let us consider a binomial distribution for Y_{cjh} with paramaters n_{cjh} and $p_{hs(c,j)}$:

$$Y_{cjh} \sim \mathcal{B}(n_{cjh}, p_{hs(c,j)}) \quad (2)$$

where n_{cjh} is the number of observed impressions, $p_{hs(c,j)}$ the click probability of campaign c at time slot h and day of week $s(c,j)$, with $s(c,j) = 1,...,S$. We have a focus on the ratio $\frac{Y_{cjh}}{n_{cjh}}$ for the following.
In the case of the binomial, $\theta_{cjh} = log\left(\frac{p_{hs(c,j)}}{1-p_{hs(c,j)}}\right)$. We can define a, b and d functions as follows: $a_{cjh}(\psi) = \frac{1}{n_{cjh}}$, $b(\theta_{cjh}) = \log(1+\exp\theta_{cjh})$ and $d(y_{cjh},\psi_{cjh}) = \log\binom{n_{cjh}}{y_{cjh}}$.
We define the logit function $\eta = g(\mu) = \log\frac{\mu}{1-\mu}$ where μ is the expectation of ratio Y/n. Then, the function links the linear combination of the β parameters with the expectation $E(\frac{Y_{cjh}}{n_{cjh}})$:

$$\log\left(\frac{E(Y_{cjh}/n_{cjh})}{1-E(Y_{cjh}/n_{cjh})}\right) = \beta_0 + \beta_h^H + \beta_{s(c,j)}^S \quad (3)$$

In Equation (3), β_h^H and $\beta_{s(c,j)}^S$ coefficients are associated to the 2 categorical variables: time slot and day of the week. In the following, the model will be extended by other advertising context variables (see Section 3.4).

2.2 Mixture of binomial for ads clustering

The objective of our approach is to obtain a mixture model of binomial distributions. Considering that C campaigns come from K subpopulations, the mixture allows to model the heterogeneity of an overall population. We denote for each campaign c the density function $f_k(y_c;\phi_k), k = 1,...,K$ with the model parameters $(\phi_1,...,\phi_K)$. Campaign density function can also be written as following: $f(y_c) = \prod_{j=d_c}^{f_c}\prod_{h=1}^{H} f(y_{cjh})$, for all $k = 1,\dots,K$ where d_c and f_c respectively are the first and last day of diffusion observed for the campaign c. We assume that a campaign belongs to the same subpopulation throughout the time.
The considered mixture model is:

$$f(y_c;\phi,\lambda) = \sum_{k=1}^{K}\lambda_k f_k(y_c;\phi_k) \quad (4)$$

where $(\lambda_1,...,\lambda_K)$ are the mixing proportion with $\lambda_k \in (0,1)$ for all k and $\sum_{k=1}^{K}\lambda_k = 1$. The log-likelihood $\log L$ is written:

$$\log Ln(Y;\phi,\lambda) = \sum_{c=1}^{C}\log\left\{\sum_{k=1}^{K}\lambda_k f_k(y_c;\phi_k)\right\} \quad (5)$$

For parameter estimation, the mixture model defined in Equation (5) can be considered as an incomplete data structure model. We introduce the hidden variable Z_{kc} where $Z_{kc} = 1$ when campaign c belongs to cluster k and 0 otherwise. Using the hidden variables, the

log-likelihood for complete data is easier to manipulate for estimation:

$$\log Ln(Y, Z; \phi, \lambda) = \sum_{c=1}^{C} \left\{ \sum_{k=1}^{K} Z_{kc} \log \left(\lambda_k f(y_c; \phi_k) \right) \right\} \quad (6)$$

The most popular way to estimate model parameters is to solve iteratively likelihood equations in order to obtain maximum likelihood estimation for each parameter. When we do not have the analytic expression of the log likelihood, the most efficient algorithms to obtain parameters estimation are the Expectation-Maximization (EM) type algorithms introduced by [4].

E-Step For each iteration, the E-Step calculates the expectation of complete-data likelihood conditionally to observed data y and current parameters $\{\lambda_k,\ \phi_k\}_{k=1,\dots,K}$. We consider $Q(\phi|\phi^{(m)}) = E(\log Ln(Y, Z; \phi, \lambda)|Y = y, \phi^{(m)})$ at iteration m. As $E(Z_{kc}|Y_c, \phi^{(m)}) = P(Z_{kc} = 1|Y_c, \phi_k^{(m)})$, thanks to Bayes formula, we calculate :

$$\begin{aligned} \pi_{kc} &= P(Z_{kc} = 1|Y_c, \phi_k^{(m)}) && (7) \\ &= \frac{P(Y_c|Z_{kc} = 1, \phi^{(m)})P(Z_{kc} = 1)}{\sum_{l=1}^{K} P(Y_c|Z_{lc} = 1, \phi^{(m)})P(Z_{lc} = 1)} && (8) \\ &= \frac{f_{\phi_k^{(m)}}(y_c)\lambda_k}{\sum_{l=1}^{K} f_{\phi_l^{(m)}}(y_c)\lambda_l} && (9) \end{aligned}$$

The probability π_{kc} represents the probability that the campaign c belongs to the cluster k at iteration m. $\pi \in \mathbf{M}_{k \times n}$ is a matrix of probabilities where the sum of each column is equal to one.

M-Step The M Step consists in maximizing $Q(\phi|\phi^{(m)})$ in order to update the model parameters. As we model a mixture of binomial, there is no explicit solution for the β_k parameters. We use the iterative Fisher algorithm [14] to estimate β_k at each M Step :

$$\beta_k^{(m+1)} = \beta_k^{(m)} - \left(E\left[\frac{\partial^2 Q(\phi|\phi^{(m)})}{\partial \beta_k \partial \beta_{k'}} \right] \right)^{-1} \frac{\partial Q(\phi|\phi^{(m)})}{\partial \beta_k} \quad (10)$$

This algorithm is based on Newton-Raphson algorithm, in which the search direction of the new value $\left(-\frac{\partial^2 Q(\phi|\phi^{(m)})}{\partial \beta_k \partial \beta_{k'}} \right)$ is replaced by its expectation. We recognize the Fisher Information expression.
Using the mixture model of binomial defined in Equation (1) and (3), a Fisher algorithm iteration from Equation (10) leads to the following estimation of parameters $\beta_k^{(m+1)}$:

$$\begin{aligned} \beta_k^{(m+1)} = &\left(\sum_{c=1}^{C} \pi_{kc} M_c^t W^{-1}_{c\beta_k^{(m)}} M_c \right)^{-1} \times \\ &\sum_{c=1}^{C} \pi_{kc} M_c^t W^{-1}_{c\beta_k^{(m)}} \left[M_c \beta_k^{(m)} + \frac{\partial \eta_{kc}}{\partial \mu_k} \left(\frac{Y_c}{n_c} - \mu_k \right) \right] \end{aligned} \quad (11)$$

with M_c the design matrix of the campaign c, μ_k the ratio of click (CTR) expectation in cluster k. The diagonal matrices are defined : $W_{c\beta_k^{(m)}} = diag\left(\frac{1}{n_{cjh}} \frac{(1+\exp M_{cjh}\beta_k^{(m)})^2}{\exp M_{cjh}\beta_k^{(m)}} \right)_{cjh}$ and $\frac{\partial \eta_k}{\partial \mu_k} = diag\left(\frac{(1+\exp M_{cjh}\beta_k^{(m)})^2}{\exp M_{cjh}\beta_k^{(m)}} \right)_{cjh}$
For each step M, we repeat a few iterations of the Fisher scoring algorithm.

This GLM model-based clustering is implemented in the R package *binomialMix* (available on CRAN[2]). Note that the number of observations n_c for each campaign c do not need to be equal.

2.3 Predict CTR from clustering results

According to the problem statement described in the Section 1, the final goal is to predict the probability of click for each campaign c. The predictions are made in order to choose one campaign c as soon as there is an advertising position available. We consider five predictive approaches. Three are considered as baseline: two naive baseline predictions and a standard GLM (without mixture). The other two predictive models are based on GLM mixture estimations.

A) A vector of zeros Data is very imbalanced: 71% of the CTR value equals zero. The most naive baseline is to consider a CTR prediction always equal to 0 no matter what the context is.

B) Yesterday's CTR This second approach is an other naive way to model the prediction. We consider that for each observation Y_{cjh}, the observed CTR at exactly the same time slot but one day before is the predicted CTR for the current moment. We make the hypothesis that from one day to another, CTR values remain stable in a similar context.

C) Binomial predictive model We consider a classical Generalized Linear Model with a binomial distribution as described in equations (2) and (3). With this simple modeling, we analyze if there is a relevant linear combination of features able to predict a click probability for any given context for all the campaigns.

We consider these three models A), B) and C) as baselines.

D) Mixture of binomial The most intuitive methodology to implement from clustering results is to use the estimated β_k from each cluster. With these estimations, we can naturally obtain prediction for each cluster of campaigns.

E) Mixture of binomial + individual random effect for each cluster We now assume that the n_c observations ($n = \sum_{c=1}^{C} n_c$) from one campaign c are no longer independent. For each target CTR value y, we define a random effect ξ_c to model dependence of observations from a same campaign c. Lets consider η the logit function defined for Equation (3). The Generalized Linear Mixed Model (GLMM) for the C campaigns can be written

$$\eta_\xi = M\beta + U\xi \quad (12)$$

where η_ξ is defined from linked function $g : \eta_\xi = g(\mu_\xi)$ with $\mu_\xi = E(Y|\xi)$. $\beta \in \mathbf{R}^B$ is the vector of the B fixed effects. M is the design matrix associated. We denote $\xi \in \mathbf{R}^C$ the random effect vector of size C and U the design matrix. We suppose that ξ_c follows a Gaussian distribution $\mathcal{N}(0, \sigma^2 \mathbf{I}_c)$. Conditionally to ξ, the model has the same properties as the GLM from Equation (2), (3).
In this approach, we run the GLMM model (see Equation (12)) for each cluster using the R package *lme4*. The model estimates fixed effects β as well as the random effect ξ, i.e. specific coefficients for each campaign present in the cluster in question. Once all the parameters are estimated, we can calculate predictions for each campaign and for each cluster.

[2] https://cran.r-project.org/web/packages/binomialMix/index.html

3 Experiments

3.1 Dataset

We consider a dataset from TabMo's[3] real traffic. The platform provides us with a very large volume of data as the incoming traffic hits a million bid requests per second. The data has first been preprocessed in order to obtain the expected format. We aggregate by campaign, time slot, day of week and ad slot features the observed number of clicks and impressions (number of times a given advertising is seen). Some campaigns last few days and others are displayed for months. The whole dataset contains 70123 rows and 12 columns. An extract of 4 rows randomly chosen is available in Table 1. We differentiate 2 types of variables:

1. The response variable CTR (in bold) is calculated as the ratio of the number of clicks (Y_{cjh}) and the number of impressions (n_{cjh}). The data is very imbalanced with around one click for 1000 impressions.

2. The other variables are the explainable features used for the modeling. Most of them are categorical as described in Table 2. *Time slot* contains 6 different labels (00h-7h,7h-12h,12h-14h,14h-18h,18h-00h) defined by domain experts and extracted from *timestamp* variable. The *ID* column is a distinct of all the 373 observed campaigns for the dataset that we want first to classify and then predict CTR.

Table 1. Extracted rows from Dataset which is composed of 70123 rows, 12 columns and 373 advertising campaign

ID	*Timestamp*	DayWeek	TimeSlot	OS	Support
622	*2018-11-20*	3	3	Android	Site
622	*2018-11-20*	3	4	Android	Site
377	*2019-01-26*	7	2	Android	App
101	*2018-12-02*	1	4	iOS	App

Ad Type	Ad Length	Ad Height	**Impressions**	**Clicks**	**CTR**
banner	320	480	**31**	**0**	**0**
banner	320	480	**57**	**1**	**0.017**
custom	320	480	**180**	**2**	**0.011**
banner	300	250	**64**	**0**	**0**

Table 2. Description of explainable features used for the model-based clustering

	Type	#Label	Description
1-Day of Week	Categorical	7	Monday to Sunday
2-Time Slot	Categorical	6	Slots of few hours
3-OS Type	Categorical	3	iOS, Android, Other
4-Support Type	Categorical	2	Application, Site
5-Advertising Type	Categorical	3	Example : Type 1
6-Advertising Length	Numerical		Pixels dimension
7-Advertising Height	Numerical		Pixels dimension

3.2 Evaluation metrics

Model choice metrics To select the right number of clusters in the mixture model, we use the BIC criterion [17]. The selected model is the one that minimizes its value. BIC is defined:

$$BIC = -2 \times (\log \hat{L}) + m \times \log(n) \tag{13}$$

where $(\log \hat{L})$ is the maximized value of the incomplete log-likelihood defined in Equation (5). m is the global number of parameters for the model and n the total number of observations in the dataset.
We can also use the Integrated Complete Likelihood (ICL) criterion [1] which is an adaptation of the BIC dedicated to clustering.

Clustering robustness metrics In order to compare clustering results, we use Adjusted Rand Index (ARI) introduced by [7]. It is based on Rand Index (RI) [15] which is a measure of similarity between two partitions and calculates the percentage of pairs in agreement. The RI and ARI values are between 0 and 1. A Rand Index (or Adjusted Rand Index) equal to 1 corresponds to two identical clustering partition. The Adjusted Rand Index is a corrected version of the Rand Index. The calculation of this index is presented in Equation (14) with notation from the contingency table described in Table 3. In this table, we consider two partitions P and Q with respectively k and j clusters.

$$ARI = \frac{\sum_{l,k}\binom{n_{jk}}{2} - \left[\sum_l \binom{n_{l\cdot}}{2}\sum_k \binom{n_{\cdot k}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_l \binom{n_{l\cdot}}{2} + \sum_k \binom{n_{\cdot k}}{2}\right] - \left[\sum_l \binom{n_{l\cdot}}{2}\sum_k \binom{n_{\cdot k}}{2}\right]/\binom{n}{2}} \tag{14}$$

Table 3. Contingency table for two clustering partitions P and Q

	Partition 1				
Partition 2	p_1	p_2	$\dots$	p_k	Sums
q_1	n_{11}	n_{12}	$\dots$	n_{1k}	$n_{1\cdot}$
q_2	n_{21}	n_{22}	$\dots$	n_{2k}	$n_{2\cdot}$
$\dots$	$\dots$	$\dots$	$\ddots$	$\dots$	$\vdots$
q_l	n_{l1}	n_{l2}	$\dots$	n_{lk}	$n_{l\cdot}$
Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot k}$	$\mathbf{n} = n_{\cdot\cdot}$

Predictive accuracy metric In order to evaluate the predictive accuracy, we consider the logloss [13]. The logloss is used to calculate the difference between a prediction and its associated target variable (which is between 0 and 1). For example, if a model predicts a probability $\hat{p} = 0.004$ and that true observation is 1, the logloss will be very bad. Logloss increases as the predicted probability diverges from the actual label. In our case, each observation is an aggregation. We study the number of times an ad is clicked (y_{cjh}) among the number of times the ad is displayed (n_{cjh}) for each campaign c at a specific time j, h (see Equation (2)). We define the number of "no click" : $n_{cjh} - y_{cjh}$. The logloss ($LogLoss_{cjh}$) is calculated for each observation of the dataset (Equation (15)).

$$LogLoss_{cjh} = -(y_{cjh}\log\hat{p} + (n_{cjh} - y_{cjh})\log(1-\hat{p})) \tag{15}$$

The resulting value that we want to analyze is the **mean logloss** :

$$LogLoss = \frac{\sum_c \sum_j \sum_h LogLoss_{cjh}}{\sum_c \sum_j \sum_h n_{cjh}} \tag{16}$$

where the numerator is the sum of logloss for each aggregated observations and the denominator is the total number of impressions.

[3] https://hawk.tabmo.io/

3.3 A short simulation study

Before evaluation on real data with significant size, we carry out a two-step simulation study: in the first step, we try to find the right partition when we know the model. In the second step, the objective is to find both the right model and the right partition.

We first assess the ability of our approach to find the right partition when the Generalized Linear Model is known. We simulate ratio of clicks for C (here, $C = 373$) advertising campaigns, uniformly distributed in $K = 2$ to 5 clusters. For each cluster, the CTR is simulated according to a binomial distribution with only 2 explanatory variables: day and time slot. Expectation of the rate of clicks is selected in 3 different intervals ($[0.2, 0.5]$, $[0.1, 0.2]$, and $[0.01, 0.1]$) so that we can estimate the impact of a low CTR in the modeling. 10 simulations were carried out in each situation. The results are presented in Figure 1. The number of clusters is correctly estimated for 2 and 3 simulated clusters, regardless of the click rate expectation. From 4 simulated clusters, the number of clusters is not always correctly estimated, especially since the expectation of the CTR is low. This is an expected behavior of the model since there are fewer campaigns involved in parameter estimation in each class. Looking at the

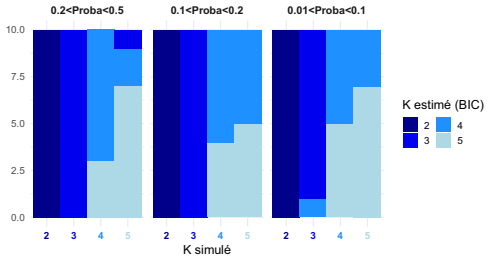


Figure 1. Comparison of the number of clusters simulated and estimated by BIC when the model is known

Adjusted Rand Index in Figure 2, conclusions are mostly identical. Indeed, up to 4 simulated clusters, for a click rate expectation greater than 0.1, the estimated partitions are very close to the simulated partitions. The partition estimation quality deteriorates for a click rate expectation of less than 0.1, even for a small number of clusters. According to these first simulations, a data set of 373 campaigns allows to correctly identify up to 3 to 4 clusters, for a CTR expectation higher than 0.1 in the case of a binomial model defined by 11 free parameters.

We are now evaluating the ability of our approach to find the right

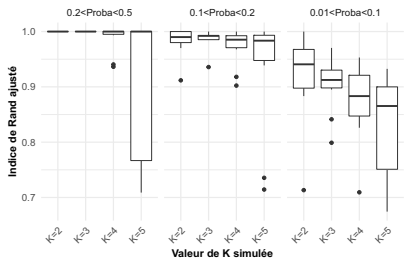


Figure 2. Adjusted Rand Index boxplot when the model is known

partition and model simultaneously. We simulated CTR for 400 advertising campaigns, uniformly distributed in K=2 to 5 clusters. The click rate is simulated according to a binomial distribution with different parameters: 2 explanatory variables (day and time slot), a single day feature, a single time slot feature, the intercept only. The expectation of the click rate is set in the interval $[0.2, 0.5]$. For each of the 8 simulations performed in each case, the model and partition chosen are the ones minimizing the BIC criterion. The results for the case K=4 and 2 explanatory variables model are presented in Table 4. The correct partition and model are found in 7 out of 8 cases, with just an error on the partition for the last case. The results leads us to the same conclusion in the other cases.

Table 4. Example of simulation where the number of clusters is equal to 4 and features used are day of week and time slot. The correct model and right number of clusters is retrieved in 7 out of 8 simulations.

	2	3	4	5
Day feature	0	0	0	0
Time Slot feature	0	0	0	0
Intercept only	0	0	0	0
Day feature + Time Slot feature	0	0	7	1

3.4 Results on Real Data

First results on real data concern mixture of binomial for advertising campaigns. Based on Equation (3) and features described in Table 2, the model (7 features, 19 free parameters) is defined in Equation (17) with β_{os}^{OS} corresponding to categorical feature OS type, β_{as}^{AS} to Support Type, β_{ad}^{AD} to Advertising Type (see Table 2). The advertising size (length and width) is measured by two numerical variables x_l and x_w.

$$\log\left(\frac{E(Y_{cjh}/n_{cjh})}{1-E(Y_{cjh}/n_{cjh})}\right) = \beta_0 + \beta_h^H + \beta_{s(c,j)}^S + \beta_{os}^{OS} + \beta_{as}^{AS} + \beta_{ad}^{AD} + \alpha_l x_l + \alpha_w x_w \quad (17)$$

For sake of clarity, we keep the index cjh which should be augmented by os, as, ad.

Optimal number of clusters The number of clusters varies from $K = 2$ to $K = 6$. We evaluate the number of clusters with the BIC criterion (Equation 13). Based on Figure 3, the optimal number of clusters is $K = 5$. Same evaluation is done with ICL and gives the same results with a minimum value for $K = 5$.

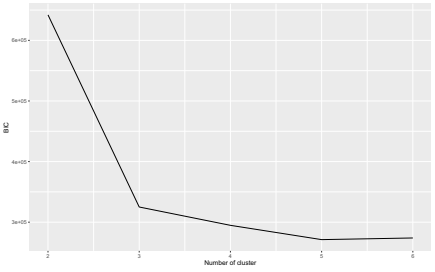


Figure 3. Evaluation of BIC for $K = 2$ to $K = 6$

Inferred profiles The resulting mixture model is composed of 5 groups divided as following: clusters 1, 2, 3, 4 and 5 contain respectively 39, 217, 29, 37 and 73 campaigns.
We analyze the inferred profiles for each cluster in Figure 4. As there are 7 *day of week* levels and 6 *time slot* levels, the x-axis represents the 42 combinations from those temporal features. The scale of the y-axis is the mean CTR estimated (in percentage).
Each figure corresponds to a cluster. From left to right on the top line, cluster 1, cluster 2 and cluster 3 are respectively displayed. On the bottom line, still from left to right, are displayed cluster 4 and cluster 5.
Figure 4 displays the mean estimated profiles for the 18 possible combinations of features levels, for banner width set to 320 and banner length to 480. The highlighted profile for cluster corresponds to the combination: Android, Application and Advertising Type 3. This is the configuration we want to compare. The scale of the y-axis is

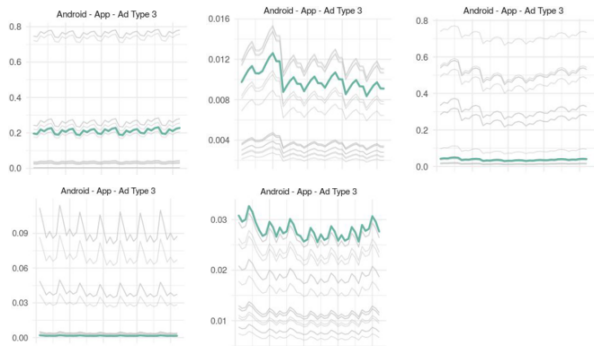


Figure 4. Estimation of inferred profiles for each cluster when Type of OS is Android, Type of support is Application, Ad Type is of Type 3 and Ad size is 320x480

different from one cluster to another. For cluster 1, the average CTR is around 0.2 while for the other clusters, the average CTR is well below 0.1. For any given combined levels, the inferred profiles are very different. We conclude that the clustering model groups advertising campaign with similar profiles and distinguishes specific types of campaigns from one group to another one. We analyze more in details two clusters. In Figure 5, eight different profiles are displayed. Dash line represents profiles when Support Type level feature is *Application*. Solid line is used for *Site* level. Red shaded lines correspond to Android profiles and blue ones to iOS.

- Cluster 1 groups campaigns with high CTR, especially for Site support and advertising of Type 3.
- Cluster 5 is composed of campaigns mainly affected by the App or Site feature, regardless of the type of advertising and the type of OS.

Prediction accuracy We evaluate the predictive performance of the models described in Section 2. Models (A) and (B) are naive modeling whose learning is respectively done from a vector of zeros or from information of the previous day. Model (C) is a generalized linear model with a logit link function with the features described in Table 2. Models (D) and (E) result from the mixture model: in model (D), we calculate the predictions based on the estimated β coefficients for each cluster. For model (E), for each cluster, we run a GLMM with a random "campaign" effect. To compare accuracy

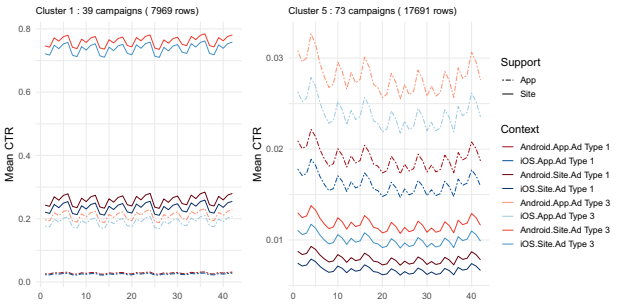


Figure 5. Inferred profiles for clusters 1 and 5. Cluster 1 groups campaign with high CTR, especially for Site support and advertising of Type 3. Cluster 5 is composed of campaigns mainly affected by the App or Site feature, regardless of the type of advertising and the type of OS.

of the models, we calculate the mean logloss (Equation 16) in two different cases:

1. **Test 1:** the dataset is extracted from one day randomly chosen (15-04-19). For models A and B, there is no learning set since we either use a vector of zeros or the vector of CTR observed the day before. For the three other models, the learning dataset is extracted from March 14th to April 14th, 2019.

   Mean Logloss for baseline Models (A) and (B) are 0.52 and 0.11. The binomial model (C) gives a mean logloss of 0.10. Models (D) and (E) resulting from the clustering step have a mean logloss respectively equal to 0.0812 and 0.0799. The addition of a random effect for each campaign ID seems to be relevant since the model (E) mean logloss outperforms other modeling.

2. **Test 2:** As the first test on one randomly chosen day provides good results, we repeat the same test procedure as before but on more days. We make the test and learning timestamp window evolve by shifting them from one day to the next one. For each new test/learning set, we run the test procedure. It allows to obtain a more global mean logloss since we learn and test on more distinct datasets. For this experiment, two periods of the year are studied: November/December and March/April. These two moments of year are very different. In November and December, activity on the bidding platform is very dense due to the end-of-year holidays, which generate a lot of advertising to display. March and April period is much calmer in terms of traffic observed on the platform. First, we shift from the first learning (01/11 - 30/11) and test set (01/12) to the last learning (30/11 - 30/12) and test set (31/12). Second, we do the learning on March/April. We shift from the first learning (01/03 - 31/03) and test set (01/04) to the last learning (30/03 - 29/04) and test set (30/04). The predictive procedure is done 30 times for both period. We analyze the resulting predictions.

For both periods, Model (A) is widely outperformed by the 4 others. The GLM mixtures outperforms the others: the mean logloss is the lowest with models (D) and (E). The clustering step with the mixture model seems to be relevant for the predicting step. Adding a random effect for each campaign in Model (E) provides a better logloss compared to the prediction with model-based clustering only. Even if it seems to be a small improvement in terms of logloss evaluation, it can lead to a significant increase for the company. In Figure 6 , we

Table 5. Mean logloss value for five models described in Section 2.3. In the first column, we calculate the mean Logloss for 30 days in December 2018. In the last column, the mean logloss is calculated for 30 days in April 2019

	Mean Logloss (December)	Mean Logloss (April)
Model (A)	0.2041	0.3468
Model (B)	0.0572	0.0948
Model (C)	0.0465	0.0711
Model (D)	0.0413	0.0598
Model (E)	**0.0405**	**0.0592**

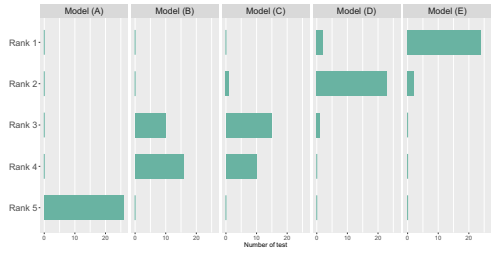


Figure 6. Ranking of the five models (see Section 2.3) obtained for each of the 30 tests performed in December 2018. Rank 1 corresponds to a model whose logloss is minimized compared to other models. Rank 5 corresponds to a model that has the worst logloss for a given test.

analyze more in details the mean logloss obtained for each models and each test day of December 2019. Rank 1 corresponds to a model whose logloss is minimized compared to other models. Rank 5 corresponds to a model that has the worst logloss for a given test. Model (A) always provides the worst mean logloss. Model (D) and (E) resulting from the clustering step almost always outperform the three other models. We obtain the same conclusions as before for Table 5: a preliminary model-based clustering step improves the prediction accuracy according to the mean logloss.

4 Conclusion and Perspectives

In this paper, we proposed a two-step methodology for the prediction of CTR in an industrial context.
First objective was to obtain a mixture model for binomial and longitudinal data. In the second step, several predictive models were in competition. Three were considered as baselines and the two others used estimated coefficients from each cluster to predict a probability of click. Using a preliminary clustering step before prediction improved the performance of prediction with a relevant logloss decrease. For future work, we want to study the optimal history window necessary for the learning step. We also want to expand the model by adding new contextual features such as the IAB category[4] for each application/site. It could be useful for the predictive task to consider second order interactions between features.
For further experiments, the model will be implemented in the large scale auction system. The objective is to evaluate its performance by A/B testing feedback in production.

[4] https://www.iab.com/wp-content/uploads/2016/03/OpenRTB-API-Specification-Version-2-5-FINAL.pdf

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments, which helped improve this paper.

REFERENCES

[1] Christophe Biernacki, Gilles Celeux, and Gérard Govaert, 'Assessing a mixture model for clustering with the integrated completed likelihood', *IEEE transactions on pattern analysis and machine intelligence*, **22**(7), 719–725, (2000).
[2] Olivier Chapelle, Eren Manavoglu, and Romer Rosales, 'Simple and scalable response prediction for display advertising', *ACM Transactions on Intelligent Systems and Technology (TIST)*, **5**(4), 61, (2015).
[3] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al., 'Wide & deep learning for recommender systems', in *Proceedings of the 1st workshop on deep learning for recommender systems*, pp. 7–10. ACM, (2016).
[4] Arthur P Dempster, Nan M Laird, and Donald B Rubin, 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22, (1977).
[5] C Fraley, AE Raftery, L Scrucca, TB Murphy, M Fop, and ML Scrucca. Gaussian mixture modelling for model-based clustering, classification, an density estimation, 2018.
[6] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He, 'Deepfm: a factorization-machine based neural network for ctr prediction', *arXiv preprint arXiv:1703.04247*, (2017).
[7] Lawrence Hubert and Phipps Arabie, 'Comparing partitions', *Journal of classification*, **2**(1), 193–218, (1985).
[8] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin, 'Field-aware factorization machines for ctr prediction', in *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 43–50. ACM, (2016).
[9] Gouthami Kondakindi, Satakshi Rana, Aswin Rajkumar, Sai Kaushik Ponnekanti, and Vinit Parakh, 'A logistic regression approach to ad click prediction', *Mach Learn Class Project*, (2014).
[10] Friedrich Leisch, 'Flexmix: A general framework for finite mixture models and latent glass regression in r', (2004).
[11] P McCullagh and John A Nelder, *Generalized Linear Models*, volume 37, CRC Press, 1989.
[12] Geoffrey McLachlan and David Peel, *Finite mixture models*, John Wiley & Sons, 2004.
[13] Kevin P Murphy, *Machine learning: a probabilistic perspective*, MIT press, 2012.
[14] John Ashworth Nelder and Robert WM Wedderburn, 'Generalized linear models', *Journal of the Royal Statistical Society: Series A (General)*, **135**(3), 370–384, (1972).
[15] William M Rand, 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical association*, **66**(336), 846–850, (1971).
[16] Steffen Rendle, 'Factorization machines', in *2010 IEEE International Conference on Data Mining*, pp. 995–1000. IEEE, (2010).
[17] Gideon Schwarz et al., 'Estimating the dimension of a model', *The annals of statistics*, **6**(2), 461–464, (1978).
[18] Mixmod Team. Mixmod statistical documentation, 2008.
[19] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang, 'Deep & cross network for ad click predictions', in *Proceedings of the ADKDD'17*, p. 12. ACM, (2017).
[20] Shuai Yuan, Jun Wang, and Xiaoxue Zhao, 'Real-time bidding for online advertising: measurement and analysis', in *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, p. 3. ACM, (2013).
[21] Robbin Lee Zeff and Bradley Aronson, *Advertising on the Internet*, John Wiley & Sons, Inc., 1999.

Annexe D

Bessière, C., Taha, M., Petitprez, F., Vandel, J., Marin, J.-M., Bréhélin, L., Lèbre, S., Lecellier, C.-H. (2018)

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Probing instructions for expression regulation in gene nucleotide compositions

Chloé Bessière[1,2☯], May Taha[1,2,3☯], Florent Petitprez[1,2], Jimmy Vandel[1,4], Jean-Michel Marin[1,3], Laurent Bréhélin[1,4‡*], Sophie Lèbre[1,3,5‡*], Charles-Henri Lecellier[1,2‡*]

1 IBC, Univ. Montpellier, CNRS, Montpellier, France, **2** Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France, **3** IMAG, Univ. Montpellier, CNRS, Montpellier, France, **4** LIRMM, Univ. Montpellier, CNRS, Montpellier, France, **5** Univ. Paul-Valéry-Montpellier 3, Montpellier, France

☯ These authors contributed equally to this work.
‡ LB, SL, and CHL also contributed equally to this work.
* brehelin@lirmm.fr (LB); sophie.lebre@umontpellier.fr (SL); charles.lecellier@igmm.cnrs.fr (CHL)

Citation: Bessière C, Taha M, Petitprez F, Vandel J, Marin J-M, Bréhélin L, et al. (2018) Probing instructions for expression regulation in gene nucleotide compositions. PLoS Comput Biol 14(1): e1005921. https://doi.org/10.1371/journal.pcbi.1005921

Editor: Zhaolei Zhang, University of Toronto, CANADA

Received: July 11, 2017

Accepted: December 10, 2017

Published: January 2, 2018

Data Availability Statement: All relevant data are within the paper, its Supporting Information files, and at http://www.univ-montp3.fr/miap/~lebre/IBCRegulatoryGenomics.

Funding: The work was supported by funding from CNRS, Plan d'Investissement d'Avenir #ANR-11-BINF-0002 Institut de Biologie Computationnelle (young investigator grant to CHL and post-doctoral fellowship to JV), Labex NUMEV (post-doctoral fellowship to JV), INSERM-ITMO Cancer project "LIONS" BIO2015-04. MT is a recipient of a CBS2-

Abstract

Gene expression is orchestrated by distinct regulatory regions to ensure a wide variety of cell types and functions. A challenge is to identify which regulatory regions are active, what are their associated features and how they work together in each cell type. Several approaches have tackled this problem by modeling gene expression based on epigenetic marks, with the ultimate goal of identifying driving regions and associated genomic variations that are clinically relevant in particular in precision medicine. However, these models rely on experimental data, which are limited to specific samples (even often to cell lines) and cannot be generated for all regulators and all patients. In addition, we show here that, although these approaches are accurate in predicting gene expression, inference of TF combinations from this type of models is not straightforward. Furthermore these methods are not designed to capture regulation instructions present at the sequence level, before the binding of regulators or the opening of the chromatin. Here, we probe sequence-level instructions for gene expression and develop a method to explain mRNA levels based solely on nucleotide features. Our method positions nucleotide composition as a critical component of gene expression. Moreover, our approach, able to rank regulatory regions according to their contribution, unveils a strong influence of the gene body sequence, in particular introns. We further provide evidence that the contribution of nucleotide content can be linked to co-regulations associated with genome 3D architecture and to associations of genes within topologically associated domains.

Author summary

Identifying a maximum of DNA determinants implicated in gene regulation will accelerate genetic analyses and precision medicine approaches by identifying key gene features. In that context decoding the sequence-level instructions for gene regulation is of prime importance. Among global efforts to achieve this objective, we propose a novel approach

I2S joint doctoral fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

able to explain gene expression in each patient sample using only DNA features. Our approach, which is as accurate as methods based on epigenetics data, reveals a strong influence of the nucleotide content of gene body sequences, in particular introns. In contrast to canonical regulations mediated by specific DNA motifs, our model unveils a contribution of global nucleotide content notably in co-regulations associated with genome 3D architecture and to associations of genes within topologically associated domains. Overall our study confirms and takes advantage of the existence of sequence-level instructions for gene expression, which lie in genomic regions largely underestimated in regulatory genomics but which appear to be linked to chromatin architecture.

Introduction

The diversity of cell types and cellular functions is defined by specific patterns of gene expression. The regulation of gene expression involves a plethora of DNA/RNA-binding proteins that bind specific motifs present in various DNA/RNA regulatory regions. At the DNA level, transcription factors (TFs) typically bind 6-8bp-long motifs present in promoter regions, which are close to transcription start site (TSS). TFs can also bind enhancer regions, which are distal to TSSs and often interspersed along considerable physical distance through the genome [1]. The current view is that DNA looping mediated by specific proteins and RNAs places enhancers in close proximity with target gene promoters (for review [2–5]). High-resolution chromatin conformation capture (Hi-C) technology identified contiguous genomic regions with high contact frequencies, referred to as topologically associated domains (TADs) [6]. Within a TAD, enhancers can work with many promoters and, on the other hand, promoters can contact more than one enhancer [5, 7]. Several large-scale data derived from high-throughput experiments (such as ChIP-seq [8], SELEX-seq [9], RNAcompete [10]) can be used to highlight TF/RBP binding preferences and build Position Weight Matrixes (PWMs) [11]. The human genome is thought to encode $\sim$2,000 TFs [12] and >1,500 RBPs [13]. It follows that gene regulation is achieved primarily by allowing the proper combination to occur i.e. enabling cell- and/or function-specific regulators (TFs or RBPs) to bind the proper sequences in the appropriate regulatory regions. In that context, epigenetics clearly plays a central role as it influences the binding of the regulators and ultimately gene expression [14]. Provided the variety of regulatory mechanisms, deciphering their combination requires mathematical/computational methods able to consider all possible combinations [15]. Several methods have recently been proposed to tackle this problem [16–19]. Although these models appear very efficient in predicting gene expression and identifying key regulators, they mostly rely on experimental data (ChIP-seq, methylation, DNase hypersensitivity), which are limited to specific samples (often to cell lines) and which cannot be generated for all TFs/RBPs and all cell types. These technological features impede from using this type of approaches in a clinical context in particular in precision medicine. In addition, we show here that, although these approaches are accurate, their biological interpretation can be misleading. Finally these methods are not designed to capture regulation instructions that may lie at the sequence-level before the binding of regulators or the opening of the chromatin. There is indeed a growing body of evidence suggesting that the DNA sequence *per se* contains information able to shape the epigenome and explain gene expression [20–25]. Several studies have shown that sequence variations affect histone modifications [21–23]. Specific DNA motifs can be associated with specific epigenetic marks and the presence of these motifs can predict the epigenome in a given cell type [24]. Quante and Bird proposed that proteins able to "read" domains of

relatively uniform DNA base composition may modulate the epigenome and ultimately gene expression [20]. In that view, modeling gene expression using only DNA sequences and a set of predefined DNA/RNA features (without considering experimental data others than expression data) would be feasible. In line with this proposal, Raghava and Han developed a Support Vector Machine (SVM)-based method to predict gene expression from amino acid and dipeptide composition in *Saccharomyces cerevisiae* [26].

Here, we built a global regression model per sample to explain the expression of the different genes using their nucleotide compositions as predictive variables. The idea beyond our approach is that the selected variables (defining the model) are specific to each sample. Hence the expression of a given gene may be predicted by different variables in different samples. This approach was tested on several independent datasets: 2,053 samples from The Cancer Genome Atlas (1,512 RNA-sequencing data and 582 microarrays) and 3 ENCODE cell lines (RNA sequencing). When restricted to DNA features of promoter regions our model showed accuracy similar to that of two independent methods based on experimental data [17, 19]. We confirmed the importance of nucleotide composition in predicting gene expression. Moreover the performance of our approach increases by combining the contribution of different types of regulatory regions. We thus showed that the gene body (introns, CDS and UTRs), as opposed to sequences located upstream (promoter) or downstream, had the most significant contribution in our model. We further provided evidence that the contribution of nucleotide composition in predicting gene expression is linked to co-regulations associated with genome architecture and TADs.

Materials and methods

Datasets, sequences and online resources

RNA-seq V2 level 3 processed data were downloaded from the TCGA Data Portal. Our training data set contained 241 samples randomly chosen from 12 different cancers (20 cancerous samples for each cancer except 21 for LAML). Our model was further evaluated on an additional set of 1,270 tumors from 14 cancer types. We also tested our model on 582 TCGA microarray data. The TCGA barcodes of the samples used in our study have been made available at http://www.univ-montp3.fr/miap/~lebre/IBCRegulatoryGenomics.

Isoform expression data (.rsem.isoforms.normalized_results files) were downloaded from the Broad TCGA GDAC (http://gdac.broadinstitute.org) using firehose_get. We collected data for 73599 isoforms in 225 samples of the 241 initially considered. All the genes and isoforms not detected (no read) in any of the considered samples were removed from the analyses. Expression data were log transformed.

All sequences were mapped to the hg38 human genome and the UCSC liftover tool was used when necessary. Gene TSS positions were extracted from GENCODEv24. UTR and CDS coordinates were extracted from ENSEMBL Biomart. To assign only one 5UTR sequence to one gene, we merged all annotated 5UTRs associated with the gene of interest using Bedtools merge [27] and further concatenated all sequences. The same procedure was used for 3UTRs and CDSs. Intron sequences are GENCODEv24 genes to which 5UTR, 3UTR and CDS sequences described above were substracted using Bedtools substract [27]. These sequences therefore corresponded to constitutive introns. The intron sequences were concatenated per gene. The downstream flanking region (DFR) was defined as the region spanning 1kb after GENCODE v24 gene end. Fasta files were generated using UCSC Table Browser or Bedtools getfasta [27].

TCGA isoform TSSs were retrieved from https://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/unc_hg19.bed and converted into hg38 coordinates with UCSC liftover.

For other regulatory regions associated to transcript isoforms (UTRs, CDS, introns and DFR), we used GENCODE v24 annotations.

Nucleotide composition

The nucleotide (n = 4) and dinucleotide (n = 16) percentages were computed from the different regulatory sequences where:

$$percentage(N, s) = \frac{\sharp N}{l}$$

is the percentage of nucleotide N in the regulatory sequence s, with N in $\{A, C, G, T\}$ and l the length of sequence s, and

$$percentage(NpM, s) = \frac{\sharp NpM}{l - 1}$$

is the NpM dinucleotide percentage in the regulatory sequence s, with N and M in $\{A, C, G, T\}$ and l the length of sequence s.

Motif scores

Motif scores in core promoters were computed using the method explained in [11] and Position Weight Matrix (PWM) available in JASPAR CORE 2016 database [28]. Let w be a motif and s a nucleic acid sequence. For all nucleotide N in $\{A, C, G, T\}$, we denoted by $P(N|w_j)$ the probability of nucleotide N in position j of motif w obtained from the PWM, and by $P(N)$ the prior probability of nucleotide N in all sequences.

The score of motif w at position i of sequence s is computed as follows:

$$score(w, s, i) = \sum_{j=0}^{|w|-1} \log \frac{P(s_{i+j}|w_j)}{P(s_{i+j})}$$

with $|w|$ the length of motif w, s_{i+j} the nucleotide at position $i + j$ in sequence s, The score of motif w for sequence s is computed as the maximal score that can be achieved at any position of s, i.e.:

$$score(w, s) = \max_{i=0}^{l-|w|} score(w, s, i),$$

with l the length of sequence s.

Models were also built on sum scores as:

$$scoreSum(w, s) = \sum_{i=0}^{l-|w|} score(w, s, i),$$

and further compared to models built on mean scores (S1 Fig). Taking mean or sum scores per region yielded similar results (Wilcoxon test p-value = 0.68).

DNAshape scores

DNA shape scores were computed using DNAshapeR [29]. Briefly, provided nucleotide sequences, DNAshapeR uses a sliding pentamer window to derive the structural features corresponding to minor groove width (MGW), helix twist (HelT), propeller twist (ProT) and Roll from all-atom Monte Carlo simulations [29]. Thus, for each DNA shape, a score is given to

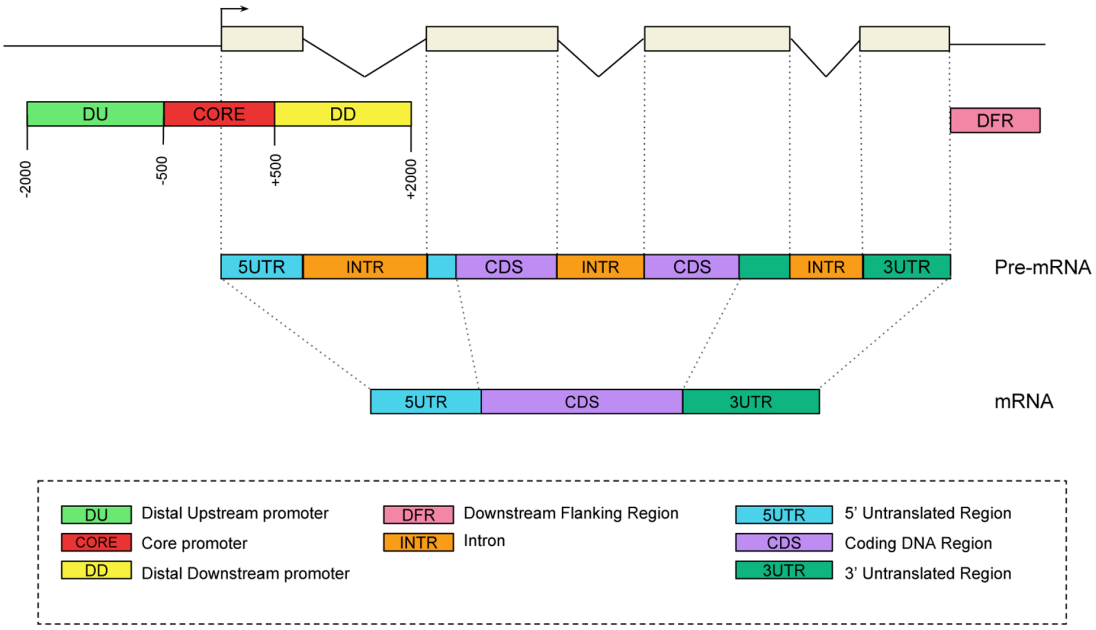


Fig 1. Genomic regions considered for gene expression prediction. An illustrative transcript is shown as example.

https://doi.org/10.1371/journal.pcbi.1005921.g001

each base of each sequence considered (DU, CORE and DD—see Fig 1). We then computed the mean of these scores for each sequence providing 12 additional variables per gene.

Enhancers

The coordinates of the enhancers mapped by FANTOM on the hg19 assembly [7] were converted into hg38 using UCSC liftover and further intersected with the different regulatory regions. We computed the density of enhancers per regulatory region (R) by dividing the sum, for all genes, of the intersection length of enhancers with gene $i(L_{enh_i})$ by the sum of the lengths of this regulatory region for all genes:

$$enhDensity_{(R)} = \frac{\sum_i (L_{enh_i} \text{ in } R_i)}{\sum_i length(R_i)}$$

Copy Number Variation (CNV)

Processed data were downloaded from the firehose Broad GDAC (https://gdac.broadinstitute.org/). We used the genome-wide SNP array data and the segment mean scores. In order to assign a CNV score to each gene, the coordinates (hg19) of the probes were intersected with that of GENCODE v19 genes using Bedtools intersect [27] and an overlap of 85% of the gene total length. The corresponding segment mean value was then assigned to the intersecting genes. In case no intersection was detected, the gene was assigned a score of 0. We next computed Spearman correlations between genes absolute error (lasso model) and genes absolute segment mean score for each of the 241 samples of the training set.

Expression quantitative trait loci and single nucleotide polymorphisms

The v6p GTex *cis*-eQTLs were downloaded from the GTex Portal (http://www.gtexportal.org/home/). The hg19 *cis*-eQTL coordinates were converted into hg38 using UCSC liftover and further intersected with the different regulatory regions. We restricted our analyses to *cis*-eQTLs impacting their own host gene. We computed the density of *cis*-eQTL per regulatory region (R) by dividing the sum, for all genes, of the number of *cis*-eQTLs of gene i ($eQTLs_i$) located in the considered region for gene i (R_i) by the sum of the lengths of this regulatory region for all genes:

$$eQTLdensity_{(R)} = \frac{\sum_i \#(eQTLs_i \text{ in } R_i)}{\sum_i length(R_i)}$$

Likewise we computed the density of SNPs in core promoters and introns by intersecting coordinates of these two regions (liftovered to hg19) with that of SNPs detected on chromosomes 1, 2 and 19 (ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b150_GRCh37p13/BED/):

$$SNPdensity_{(R)} = \frac{\sum_i \#(SNP_i \text{ in } R_i)}{\sum_i length(R_i)}$$

Methylation

Illumina Infinium Human DNA Methylation 450 level 3 data were downloaded from the Broad TCGA GDAC (http://gdac.broadinstitute.org) using firehose_get. The coordinates of the methylation sites (hg18) were converted into hg38 using the UCSC liftover and further intersected with that of the core promoters (hg38). For each gene, we computed the median of the beta values of the methylation sites present in the core promoter and further calculated the median of these values in 21 LAML and 17 READ samples with both RNA-seq and methylation data. We compared the overall methylation status of the core promoters in LAML and READ using a wilcoxon test.

Gini coefficient

We used 8,556 GTEx RNA-seq libraries (https://www.gtexportal.org/home/datasets) to compute the Gini coefficient for 16,134 genes on the 16,294 considered in our model. Gini coefficient measures statistical dispersion and can be used to measure gene ubiquity: value 0 represents genes expressed in all sam- ples while value 1 represents genes expressed in only one sample. To compute Gini coefficient we used `R` package `ineq`. We then computed, for the 241 samples, Spearman correlation between Gini coefficients and model gene absolute errors. Similar analyses were performed with 1,897 FANTOM 5 CAGE libraries to compute the Gini coefficients for 15,904 genes.

Functional enrichment

Gene functional enrichments were computed using the database for annotation, visualization and integrated discovery (DAVID) [30].

Linear regression with ℓ_1-norm penalty (Lasso)

We performed estimation of the linear regression model (1) via the lasso [31]. Given a linear regression with standardized predictors and centered response values, the lasso solves the

ℓ_1-penalized regression problem of finding the vector coefficient $\beta = \{\beta_i\}$ in order to minimize

$$Min\left(||y^c(g) - \sum_i \beta_i x^s_{i,g}||^2 + \lambda \sum_i |\beta_i|\right),$$

where $y^c(g)$ is the centered gene expression for all gene g, $x^s_{i,g}$ is the standardized DNA feature i for gene g and $\Sigma_i |\beta_i|$ is the ℓ_1-norm of the vector coefficient β. Parameter λ is the tuning parameter chosen by 10 fold cross validation. The higher the value of λ, the fewer the variables. This is equivalent to minimizing the sum of squares with a constraint of the form $\Sigma_i |\beta_i| \leq s$. Gene expression predictions are computed using coefficient β estimated with the value of λ that minimizes the mean square error. Lasso inference was performed using the function `cv.glmnet` from the `R` package `glmnet` [32]. The LASSO model was compared to two non parametric approaches: Regression trees (CART) [33] and Random forest [34]. S1 Table summarizes accuracy and computing time of each approach. Regression trees achieved significantly lower accuracy than the two other approaches (Wilcox test p-values $< 2e^{-16}$), while linear model and random forest yielded similar results (p-value 0.18). Moreover, computing time for linear model was much lower than that of random forest. These results emphasize the merits of linear model such as LASSO in their interpretability and efficiency.

Variable stability selection

We used the stability selection method developed by Meinshausen *et al.* [35], which is a classical selection method combined with lasso penalization. Consistently selected variables were identified as follows for each sample. First, the lasso inference is repeated 500 times where, for each iteration, (i) only 50% of the genes is used (uniformly sampled) and (ii) a random weight (uniformly sampled in [0.5;1]) is attributed to each predictive variable. Second, a variable is considered as stable if selected in more than 70% of the iterations, using the method proposed in [36] to set the value of lasso penalty λ. One of the advantage of this method is that the variable selection frequency is computed globally for all the variables by attributing a random weight to each variable at each iteration, thus taking into account the dependencies between the variables. This variable stability selection procedure was implemented using functions `stabpath` and `stabsel` from the `R` package `C060` for `glmnet` models [36].

Regression trees

Regression trees were implemented with the `rpart` package in `R` [32]. In order to avoid overfitting, trees were pruned based on a criterion chosen by cross validation to minimize mean square error. The minimum number of genes was set to 100 genes per leaf.

TAD enrichment

We considered TADs mapped in IMR90 cells [6] containing more than 10 genes (373 out of 2243 TADs with average number of genes = 14). The largest TAD had 76 associated genes. First, for each TAD and for each region considered, the percentage of each nucleotide and dinucleotide associated to the embedded genes were compared to that of all other genes using a Kolmogorov-Smirnov (KS) test. For a given dinucleotide (for example CpG), we applied KS tests to assess whether the CpG frequency distribution in genes in one specific TAD differs from the distribution in genes in other TADs. Correction for multiple tests was applied using the False Discovery Rate (FDR) < 0.05 [37] and the `R` function `p.adjust` [32]. Second, for each of the 967 groups of genes (identified by the regression trees, with mean error $<$ mean error of the 1st quartile), the over-representation of each TAD within each group was tested

using the R hypergeometric test function phyper [32]. Correction for multiple tests was applied using FDR< 0.05 [37].

Availability of data and materials

The matrices of predicted variables (log transformed RNA seq data) and predictive variables (nucleotide and dinucleotide percentages, motifs and DNA shape scores computed for all genes as described above) as well as the TCGA barcodes of the 241 samples used in our study have been made available at http://www.univ-montp3.fr/miap/~lebre/IBCRegulatoryGenomics.

Results

Mathematical approach to model gene expression

We built a global linear regression model to explain the expression of genes using DNA/RNA features associated with their regulatory regions (e.g. nucleotide composition, TF motifs, DNA shapes):

$$y(g) = a + \sum_i b_i x_{i,g} + e(g) \quad (1)$$

where $y(g)$ is the expression of gene g, $x_{i,g}$ is feature i for gene g, $e(g)$ is the residual error associated with gene g, a is the intercept and b_i is the regression coefficient associated with feature i.

The advantage of this approach is that it allows to unveil, into a single model, the most important regulatory features responsible for the observed gene expression. The relative contribution of each feature can thus be easily assessed. It is important to note that the model is specific to each sample. Hence the expression of a given gene may be predicted by different variables depending on the sample. Our computational approach was based on two steps. First, a linear regression model (1) was trained with a lasso penalty [31] to select sequence features relevant for predicting gene expression. Second, the performances of our model was evaluated by computing the mean square of the residual errors, and the correlation between the predicted and the observed expression for all genes. This was done in a 10 fold cross-validation procedure. Namely, in all experiments hereafter, the set of genes was randomly split in ten parts. Each part was alternatively used for the test (i.e. for comparing observed and predicted values) while the remaining genes were used to train the model. This ensures that the model used to predict the expression of a gene has not been trained with any information relative to this gene. Our approach was applied to a set of RNA sequencing data from TCGA. We randomly selected 241 gene expression data from 12 cancer types (see http://www.univ-montp3.fr/miap/~lebre/IBCRegulatoryGenomics for the barcode list). For each dataset (i.e sample), a regression model was learned and evaluated. See Materials and methods for a complete description of the data, the construction of the predictor variables and the inference procedure. We further evaluated our model on 3 independent ENCODE RNA-seq, 1,270 TCGA RNA-seq and 582 microarrays datasets (see below).

Contribution of the promoter nucleotide composition

We first evaluated the contribution of promoters, which are one of the most important regulatory sequences implicated in gene regulation [38]. We extracted DNA sequences encompassing ±2000 bases around all GENCODE v24 TSSs and looked at the percentage of dinucleotides along the sequences (S2 Fig). Based on these distributions, we segmented the promoter into three distinct regions: -2000/-500 (referred here to as distal upstream promoter, DU), -500/+500 (thereafter called core promoter though longer than the core promoter traditionally

considered) and +500/+2000 (distal downstream promoter, DD)(Fig 1). We computed the nucleotide (n = 4) and dinucleotide (n = 16) relative frequencies in the three distinct regions of each gene. For each sample, we trained one model using the 20 nucleotide/dinucleotide relative frequencies from each promoter segment separately, and from each combination of promoter segments. We observed that the core promoter had the strongest contribution compared to DU and DD (Fig 2A). Considering promoter as one unique sequence spanning -2000/+2000 around TSS achieved lower model accuracy than combining different promoter segments (Fig 2A). The highest accuracy was obtained combining all three promoter segments (Fig 2A).

Promoters are often centered around the 5' most upstream TSS (i.e. gene start). However genes can have multiple transcriptional start sites. The median number of alternative TSSs for the 19,393 genes listed in the TCGA RNA-seq V2 data is 5 and only 2,753 genes harbor a single TSS (S3 Fig). We therefore evaluated the performance of our model comparing different promoters centered around the first, second, third and last TSS (Fig 2B). In the absence of second TSS, we used the first TSS and likewise the second TSS in the absence of a third TSS. The last TSS represents the most downstream TSS in all cases. We found that our model achieved higher predictive accuracy with the promoters centered around the second TSS (Fig 2B), in agreement with [16]. As postulated by Cheng *et al.* [16] in the case of TFs, the nucleotide composition around the first TSS may be linked to the recruitment of chromatin remodelers and thereby prime the second TSS for gene expression. Dedicated experiments would be required to assess this point.

We noticed that incorporating the number of TSSs associated with each gene drastically increased the performance of our model (S4 Fig). Multiplying TSSs may represent a genuine mechanism to control gene expression level. On the other hand this effect may merely be due to the fact that the more a gene is expressed, the more its different isoforms will be detected (and hence more TSSs will be annotated). Because the number of known TSSs results from annotations deduced from experiments, we decided not to include this variable into our final model.

Contribution of specific features associated with promoters

Provided the importance of CpGs in promoter activity [38], we first compared our model with a model built only on promoter CpG content. We confirmed that CpG content had an important contribution in predicting gene expression (median R = 0.417, Fig 2C). However considering other dinucleotides achieved better model performances, indicating that dinucleotides other than CpG contribute to gene regulation. This is in agreement with results obtained by Nguyen *et al.*, who showed that CpG content is insufficient to encode promoter activity and that other features might be involved [39].

We integrated TF motifs considering Position Weight Matrix scores computed in the core promoter and observed a slight but significant increase of the regression performance (median r = 0.543 with motif scores vs. r = 0.502 without motif scores, Fig 2D). As DNA sequence is intrinsically linked to three-dimensional local structure of the DNA (DNA shape), we also computed, for each promoter segment (DU, CORE and DD), the mean scores of the four DNA shape features provided by DNAshapeR [29] (helix twist, minor groove width, propeller twist, and Roll), adding 12 variables to the model. Although the difference between models with and without DNA shapes is also significant, the increase in performance is more modest than when including TF motif scores (Fig 2D).

Our model suggested that nucleotide composition had a greater contribution in predicting gene expression compared to TF motifs and DNA shapes. This is in agreement with the

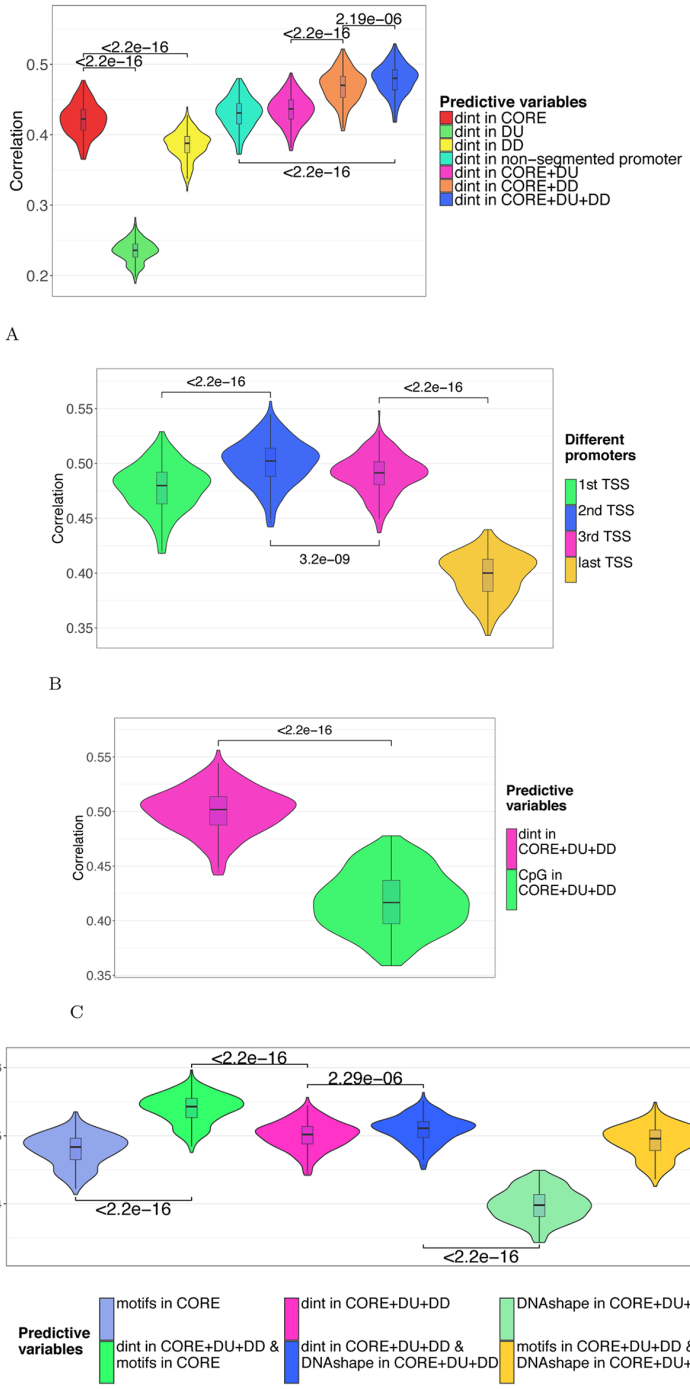


Fig 2. A: Contribution of the promoter segments. The model was built using 20 variables corresponding to the nucleotide (4) and dinucleotide (16) percentages computed in the CORE promoter (red), DU (green) or DD (yellow). These variables were then added in different combinations: CORE+DU (pink, 40 variables); CORE+DD (orange, 40 variables); CORE+DU+DD (light blue, 60 variables). Promoter segments were centered around the first most upstream TSS. For sake of comparison, the model was also built on 20 variables corresponding to the nucleotide and

dinucleotide compositions of the non segmented promoters (-2000/+2000 around the first most upstream TSS)(light blue). All different models were fitted on 19,393 genes for each of the 241 samples considered. The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients between observed and predicted gene expressions in a cross-validation procedure. The correlations obtained in all samples are shown as violin plots. **B: Prediction accuracy comparing alternative TSSs.** The model was built using the 60 nucleotide/dinucleotide percentages computed in the 3 promoter segments (CORE+DU+DD) centered around 1st, 2nd, 3rd and last TSSs (from left to right). **C: Contribution of CpG.** The model was built using the 60 nucleotide/dinucleotide or only the 3 CpG percentages computed in the 3 promoter segments (CORE+DU+DD) centered around the 2nd TSS. **D: Contribution of motifs and local DNA shapes.** The model was built using (i) 60 nucleotide/dinucleotide percentages computed in the 3 promoter segments (CORE+DU+DD) ("dint", pink),(ii) 471 JASPAR2016 PWM scores computed in the CORE segment ("motifs", light blue) and (iii) the 12 DNA shapes corresponding to the 4 known DNAshapes computed in CORE, DU and DD ("DNAshape", green). All sequences were centered around the 2nd TSS. These variables were further added in different combinations to build the models indicated: dint+motifs (531 variables, green), dint+DNAshapes (32 variables, dark blue), motifs+DNAshapes (483 variables, light green).

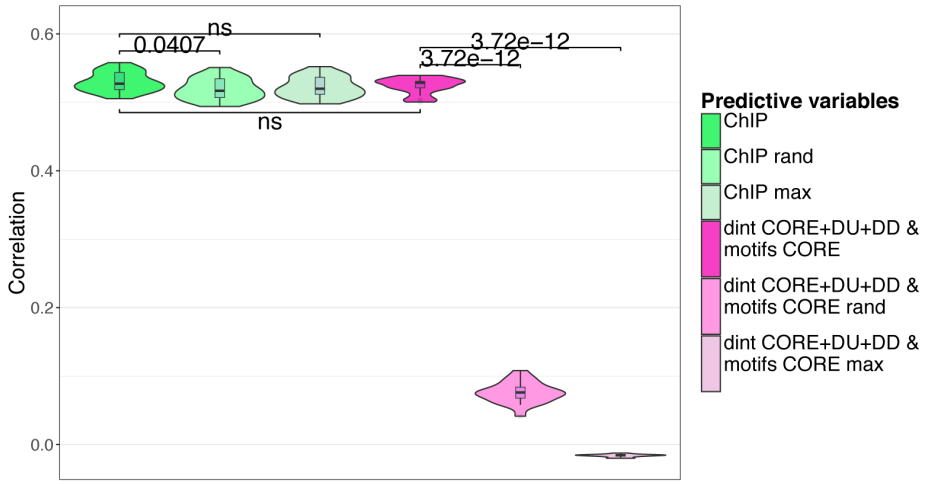
https://doi.org/10.1371/journal.pcbi.1005921.g002

findings revealing the influence of the nucleotide environment in TFBS recognition [40]. Note however that nucleotide composition, TF motifs and DNA shapes may be redundant variables. Besides, a linear model may not be optimal to efficiently capture the contributions of TF motifs and/or DNA shapes. The highest performance was achieved by combining nucleotide composition with TF motifs (Fig 2D). In the following analyses, the model was built on both dinucleotide composition and core promoter TF motifs.

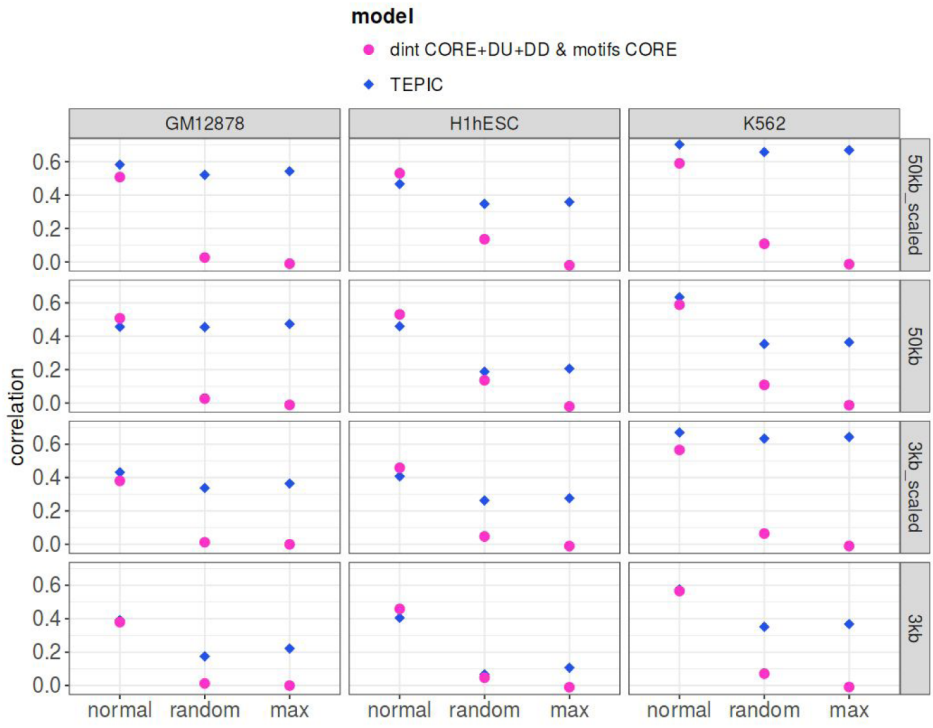
Comparison with models based on experimental data

The wealth of TF ChIP-seq, epigenetic and expression data has allowed the development of methods aimed at predicting gene expression based on differential binding of TFs and epigenetic marks [16–19]. We sought to compare our approach, which does not necessitate such cell-specific experimental data, to these methods. We first compared our results to that of Li *et al.* who used a regression approach called RACER to predict gene expression on the basis of experimental data, in particular TF ChIP-seq data and DNA methylation [17]. Note that, with this model, the contribution of TF regulation in predicting gene expression is higher than that of DNA methylation [17].

We computed the Spearman correlations between expressions observed in the subsets of LAMLs studied in [17] and expressions predicted by our model or by RACER (Fig 3A). For the sake of comparison, we used the RACER model built solely on ChIP-seq data, hereafter referred to as "ChIP-based model". RACER performance was assessed using the same cross-validation procedure we used for our method. Overall our model was as accurate as ChIP-based model (median correlation r = 0.529 with our model vs. median r = 0.527 with ChIP-based model (Fig 3A)). We then controlled the biological information retrieved by the two approaches by randomly permuting, for each gene, the values of the predictive variables (dinucleotide counts/motif scores in our model and ChIP-seq signals in the ChIP-based model). This creates a situation where the links between the combination of predictive variables and expression is broken, while preserving the score distribution of the variables associated with each gene. For example, genes associated with numerous ChIP-seq peaks will also have numerous ChIP-seq peaks in random data. In such situation, a regression model is expected to poorly perform. Surprisingly, the accuracy of ChIP-based model was not affected by the randomization process (median r = 0.517, Fig 3A) while that of our model was severely impaired (median r = 0.076, Fig 3A). We built another control model using a single predictive variable per gene corresponding to the maximum value of all predictive variables initially considered. Here again the ChIP-based model was not affected by this process (median r = 0.520, Fig 3A) while our model failed to accurately predict gene expression with this type of control variable (median r = -0.016, Fig 3A).



A



B

Fig 3. A: Comparison with model integrating TF-binding signals. The model was built using 531 variables corresponding to the 60 nucleotide/dinucleotide percentages and the 471 motif scores computed in the 3 promoter segments (CORE, DU, DD) centered around the 2nd TSS (pink). A model built on ChIP-seq data [17] was used for comparison (green). Both models were fitted on the same gene set (n = 16,298) for 21 LAML samples and assessed by cross-validation. The correlations obtained with ChIP-based RACER and our model were compared using Wilcoxon test but no significant difference was observed (p-value = 0.425). The two models were also built on randomized values of predictive variables (rand) and on the maximum value of all predictive variables (max). **B: Comparison with model integrating open-chromatin signals.** The linear model was built using the 531 variables (nucleotide/dinucleotide percentages and motif scores in CORE, DU and DD) and the expression data obtained in K562, hESC and GM12878 [19]. TEPIC was built as described in [19], within a 3 kb or a 50 kb window around TSSs. The scaled version of TEPIC

incorporates the abundance of open-chromatin peaks in the analyzed sequences. All types of TEPIC models were tested (3kb, 3kb-scaled, 50kb and 50kb-scaled) by cross-validation. In each case, our model was built on the set of genes considered by TEPIC. TEPIC uses 12 conditions making hard to compute Wilcoxon tests. A direct comparison showed that, in "normal" conditions (first column of each panel), our model and TEPIC give overall very similar results (our model being as accurate as TEPIC in 2 conditions and slightly better in 5 out of the 10 remaining conditions). Models were further built on randomized values of predictive variables (rand) and on the maximum value of all predictive variables (max). Overall, absence of effect of the randomization procedure suggests that RACER and TEPIC mainly capture the level of chromatin opening rather than the TF combinations responsible for gene expression.

https://doi.org/10.1371/journal.pcbi.1005921.g003

ChIP-seq data are probably the best way to measure the activity of a TF because binding of DNA reflects the output of RNA/protein expression as well as any appropriate post-translational modifications and subcellular localizations. However this type of data also reflects chromatin accessibility (i.e. most TFs bind accessible genomic regions) and TFs tend to form clusters on regulatory regions [41]. The binding of one TF in the promoter region is therefore likely accompanied by the binding of others. Hence, rather than inferring the TF combination responsible for gene expression, linear models based of ChIP-seq data predominantly captures the quantity of TFs (i.e. the opening of the chromatin) in the promoter region of each gene, which explains their good accuracy on randomized or maximized variables.

We indeed observed a similar bias in the results obtained by TEPIC [19], a regression method that predicts gene expression from PWM scores and open-chromatin data. Specifically, TEPIC computes a TF-affinity score for each gene and each PWM by summing up the TF affinities in all open-chromatin peaks (DNaseI-seq) within a close (3,000 bp) or large (50,000 bp) window around TSSs. This scoring takes into account the scores of PWMs in the open-chromatin peaks but is also influenced by the number of open-chromatin peaks in the analyzed sequences and the abundance of open-chromatin peaks ("scaled" version). As a result, genes with many open-chromatin peaks tend to get higher TF-affinity scores than genes with low number of open-chromatin peaks. We trained linear models on three cell-lines using either the four TEPIC affinity-scores or our variables and compared the results (Fig 3B). As for the ChIP-based models, we observed that our model was approximately as accurate as TEPIC score model, validating our approach with an independent dataset. Applying the random permutations on the TEPIC scores did not significantly impact the accuracy of the approach in most cases, especially for the scaled versions (Fig 3B). Hence, as for the ChIP-based model, the TEPIC score model seems to mainly capture the level of chromatin opening rather than the TF combinations responsible for gene expression. Conversely, our model solely built on DNA sequence features is not influenced by the chromatin accessibility and thus can yield relevant combinations of explanatory features (see the randomized control in Fig 3A and 3B). Note that the non-scaled version of TEPIC did show a loss of accuracy for cell-line H1-hESC (as well as a moderate loss for K562, but none for GM12878) when randomizing or maximizing the variables (Fig 3B). This result indicates that, although taking the abundance of open-chromatin peaks in the analyzed sequences does increase expression prediction accuracy, it might generate more irrelevant combinations of explanatory features than non-scaled versions.

Contribution of additional genomic regions

Additional genomic regions were integrated into our model. We first thought to consider enhancer sequences implicated in transcriptional regulation. We used the enhancer mapping made by the FANTOM5 project, which identified 38,554 human enhancers across 808 samples [7]. This mapping uses the CAGE technology, which captures the level of activity for both promoters and enhancers in the same samples. It is then possible to predict the potential target genes of the enhancers by correlating the activity levels of these regulatory regions over

hundreds of human samples [7]. However FANTOM5 enhancers are only assigned to 11,359 genes from the TCGA data, which correspond to the most expressed genes across different cancers (S5 Fig). Provided that the detection of enhancers relies on their activity, it is expected that enhancers are better characterized for the most frequently expressed genes. Because considering only the genes with annotated enhancers would considerably reduce the number of genes and including enhancers features only when available would introduce a strong bias in the performance of our model, we decided not to include these regulatory regions.

Second we analyzed the contribution of regions defined at the RNA level, namely 5'UTR, CDS, 3'UTR and introns, which can be responsible for post-transcriptional regulations [13, 17, 26, 42–50] (Fig 1). For all genes, we extracted all annotated 5'UTRs, 3'UTRs and CDSs, which were further merged and concatenated to a single 5'UTR, a single CDS, and a single 3'UTR per gene. Introns were defined as the remaining sequence (Fig 1). We also tested the potential contribution of the 1kb region located downstream the gene end, called thereafter Downstream Flanking Region (DFR, Fig 1). Our rationale was based on reports showing the presence of transient RNA downstream of polyadenylation sites [51], the potential presence of enhancers [7] and the existence of 5' to 3' gene looping [52].

We used a forward selection procedure by adding one region at a time: (i) all regions were tested separately and the region leading to the highest Spearman correlation between observed and predicted expression was selected as the 'first' seed region, (ii) each region not already in the model was added separately and the region yielding the best correlation was selected ('second region'), (iii) the procedure was repeated till all regions were included in the model. The correlations computed in a cross-validation procedure at each steps are indicated in S2 Table. As shown in Fig 4, the nucleotide composition of intronic sequences had the strongest contribution in the accuracy of our model, followed by UTRs (5' then 3') and CDS (Fig 4). The

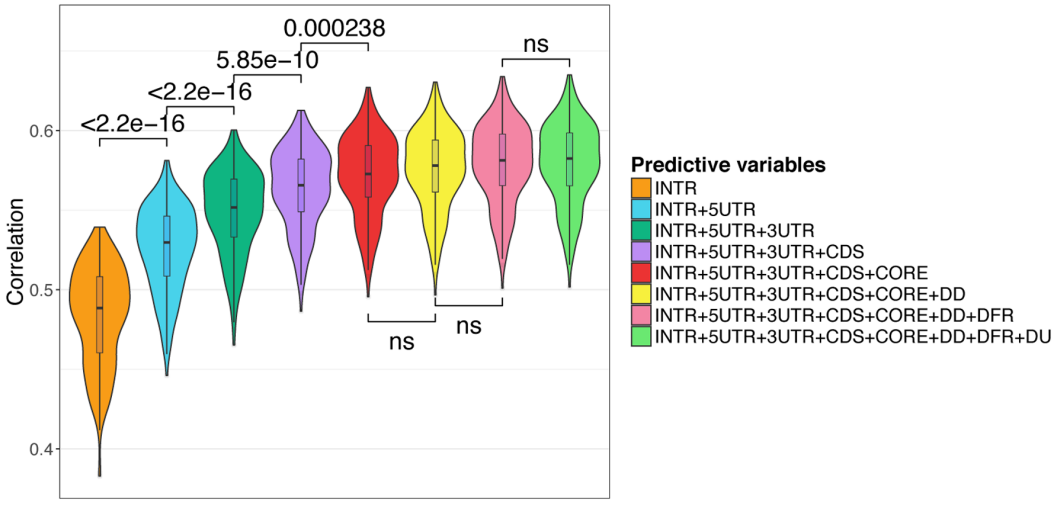


Fig 4. Contribution of additional genomic regions. Genomic regions were ranked according to their contribution in predicting gene expression. First, all regions were tested separately. Introns yielded the highest Spearman correlation between observed and predicted expressions (in a cross-validation procedure) and was selected as the 'first' seed region. Second, each region not already in the model was added separately. 5'UTR in association with introns yielded the best correlation and was therefore selected as the 'second' region. Third, the procedure was repeated till all regions were included in the model. The contribution of each region is then visualized starting from the most important (left) to the less important (right). Note that the distance between the second TSS and the first ATG is > 2000 bp for only 189 genes implying that 5'UTR and DD regions overlap. The correlations computed at each steps are indicated in (S2 Table). ns, non significant.

https://doi.org/10.1371/journal.pcbi.1005921.g004

nucleotide composition of core promoter moderately increased the prediction accuracy. In contrast the composition of regions flanking core promoter (DU and DD, Fig 1) as well as regions located downstream the end of gene (DFR, Fig 1) did not significantly improve the predictions of our model. Note that combining all regions improved the performance of our model compared to promoter alone (compare Figs 2B and 4).

We compared models built on ssDNA and dsDNA, and ssDNA-based models yielded better accuracy S6 Fig. We also compared models built on percentages of nucleotides (n = 4), dinucleotides (n = 16) and nucleotides+dinucleotides (n = 20). As shown S7A Fig, dinucleotides provided stronger prediction accuracy than nucleotides and the best accuracy was obtained combining both nucleotides and dinucleotides. We also built a model on trinucleotide percentage (n = 64) (S7A Fig). This model did yield better results than model built on nucleotide+dinucleotide. However, the correlation increase was not as important as that observed when adding dinucleotides to nucleotides. Besides, the model built on trinucleotides involves more variables and is computationally demanding. We compared models built on nucleotides+dinucleotides adding individually trinucleotide percentages of each region (i.e. 8 models built on nucleotides+dinucleotides in all regions + trinucleotides in one specific region) (S7B Fig). This analysis revealed that the correlation increase observed when incorporating trinucleotides was mostly due to the contribution of trinucleotides computed in introns, reinforcing our conclusions regarding the importance of sequence-level instructions located in this region.

Because RNA-associated regions (introns, UTRs, CDSs) had greater contribution to the prediction accuracy compared to DNA regions (promoters, DFR), we compared the accuracy of our model in predicting gene vs. transcript expression. We retrieved the normalized results for gene expression (RNAseqV2 rsem.genes.normalized_results) and the matched normalized expression signal of individual isoforms (RNAseqV2 rsem.isoforms.normalized_results) for 225 TCGA samples. Accordingly, we generated a set a predictive variables specific to each isoform (see Material and methods). We found that models built on isoforms are less accurate than models built on genes (median r = 0.35, S8 Fig and (S3 Table)). Focusing on the broad nucleotide composition may not be optimal to model isoform expression and to differentiate expression of one isoform from another. Yet another simple explanation could be that reconstructing and quantifying full-length mRNA transcripts is a difficult task, and no satisfying solution exists for now [53]. Consequently isoform as opposed to gene expression is more difficult to measure and thus to predict.

Additional validation of the model

In the above sections, our complete model, built on 160 variables corresponding to 4 nucleotide and 16 dinucleotide rates in 8 distinct regions (Fig 1), was trained with a data set containing 241 RNA-seq samples randomly chosen from 12 different cancers, and on 3 independent ENCODE RNA-seq datasets (see TEPIC comparison). We further evaluated our approach using two independent additional datasets: (a) a set of 1,270 RNA-seq samples collected from 14 cancer types and (b) a set of 582 microarray data. Overall, the RNA-seq and the microarray samples were collected from respectively 109 and 41 source sites and sequenced in 3 analysis centers. Similar accuracy was observed in all datasets (S9 and S10 Figs). Note that the correlations computed with microarray data were lower than that computed with RNA-seq data but involved lower number of genes (9,791 genes in microarrays vs. 16,294 in RNA-seq). For sake of comparison, we restricted RNA-seq data to the 9,791 microarray genes and we observed similar correlation (S10 Fig). Because our model was built on human reference genome, we also have computed the Spearman correlations between absolute values of CNV segment

mean scores and model prediction errors calculated for each gene in 241 samples corresponding to 12 cancer types. The median correlation was -0.014, arguing against the model performance being related to CNV-density (S11 Fig).

Selecting DNA features related to gene expression

We sought the main DNA features related to gene expression. The complete model built on all 8 regions (160 variables) selected $\sim$ 129 predictive variables per sample. We used the stability selection algorithm developed by Meinshausen *et al.* [35] to identify the variables that are consistently selected after data subsampling (see Materials and methods for a complete description of the procedure). This procedure selected a median of $\sim$ 16 variables per sample. The barplot in Fig 5A shows, for each variable, the proportion of samples in which the variable is selected with high consistency ($>$ 70% of the subsets).

We next determined whether stable variables exert a positive (activating) or a negative (inhibiting) effect on gene expression. For each sample, we fitted a linear regression model predicting gene expression using only the standardized variables that are stable for this sample. The activating/inhibiting effect of a variable is then indicated by the sign of its regression coefficient: <0 for a negative effect and >0 for a positive effect. The outcome of these analyses for all variables and all samples is shown Fig 5B. With the noticeable exception of CpG in the core promoter, all stable variables had an invariable positive (e.g. GpT in introns) or negative (e.g. CpA in DD and in 5UTR) contribution in gene expression prediction in all samples. In contrast, CpG in the core promoter had an alternating effect being positive in LAML and LGG for instance while negative in READ. It is also the only variable with a regression coefficient close to 0 (absolute value of median = 0.1, see S12 Fig), providing a partial explanation for the observed changes. As CpG methylation inhibits gene expression [38], we also investigated potential differences in core promoter methylation in LAML (positive contribution of CpG_CORE) and READ (negative contribution of CpG_CORE). We used the Illumina Infinium Human DNA Methylation 450 made available by TCGA and focused on the estimated methylation level (beta values) of the sites intersecting with the core promoter. We noticed that core promoters in LAML were overall more methylated (median = 0.85) than in READ (median = 0.69, wilcoxon test p-value $<$ 2.2e-16), opposite to the sign of CpG coefficient in LAML (positive contribution of CpG_CORE) and READ (negative contribution of CpG_CORE). This argued against a contribution of methylation in the alternating effect of CpG_CORE.

We observed that the accuracy of our model varied between cancer types (S9 Fig). In order to characterize well predicted genes in each sample, we used a regression tree [54] to classify genes according to the prediction accuracy of our model (i.e. absolute error). The nucleotide and dinucleotide compositions of the various considered regions were used as classifiers. This approach identified groups of genes with similar (di)nucleotide composition in the regulatory regions considered and for which our model showed similar accuracy (S13 Fig). Implicitly, it identified the variables associated with a better or a poorer prediction. We applied this approach to the 241 linear models. The number of groups built by a regression tree differs from one sample to another (average number = 14). The resulting 3,680 groups can be visualized in the heatmap depicted in Fig 6, wherein each column represents a sample and each line corresponds to a group of genes identified by a regression tree. This analysis showed that our model is not equally accurate in predicting the expression of all genes but mainly fits certain classes of genes (bottom rows of the heatmap, Fig 6) with specific genomic features (S13 Fig). Note that the groups well predicted in all cancers presumably correspond to highly and ubiquitously expressed housekeeping genes: groups with low prediction error in all samples and

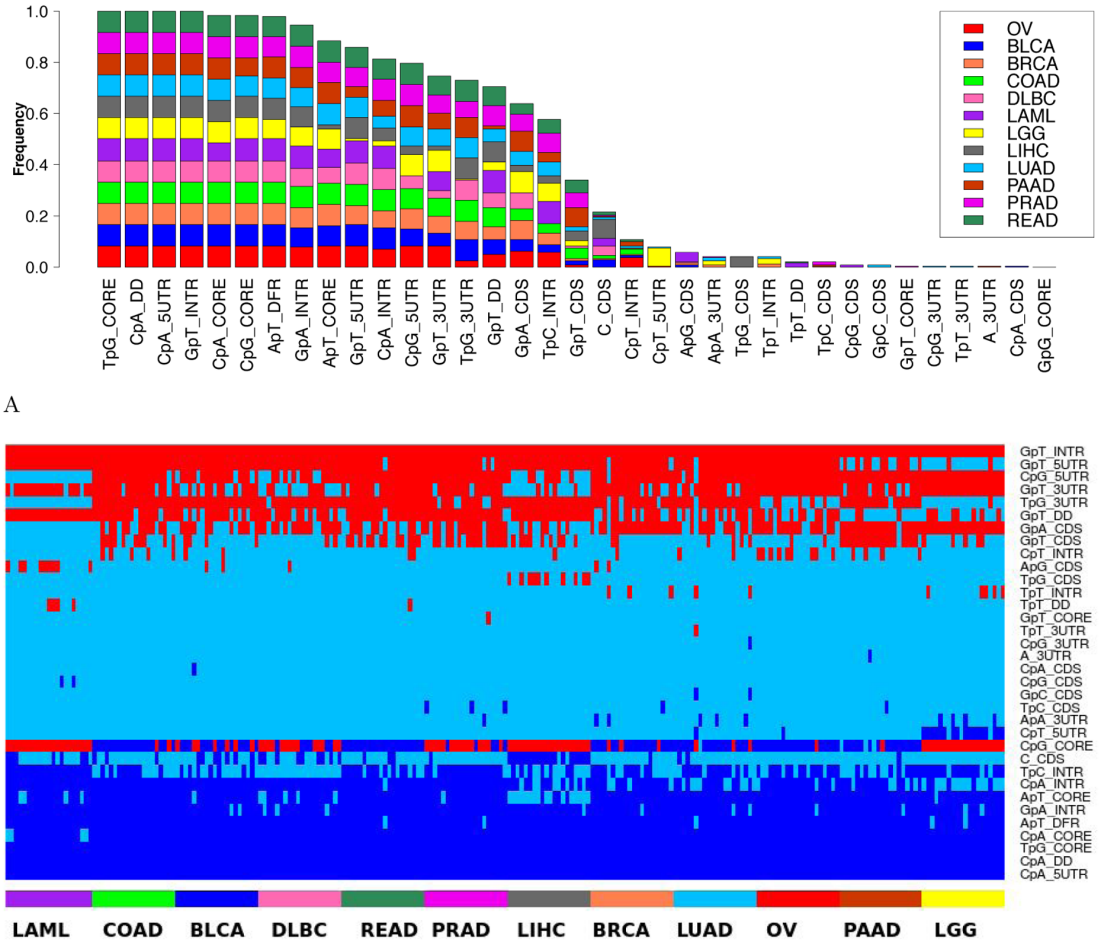


Fig 5. A: Consistently selected variables among 12 types of cancer. For each variable, the fraction of samples in which the variable is considered as stable (i. e. selected in more than 70% of the subsets after subsampling) is shown. Each color refers to a specific type of cancer. Only variables consistently selected in at least one sample are shown (out of the 160 variables). See Materials and methods for stable variable selection procedure and cancer acronyms. **B: Biological effect of the stable variables**. For each of the 241 samples (columns), a linear model was fitted using the variables (rows) stable for this sample only. The sign of the contribution of each variable in each sample is represented as follows: red for positive contribution, dark blue for negative contribution and sky blue refers to variables not selected (i.e. not stably selected for the considered sample). Only the variables stable in at least one sample are represented. Cancers and samples from the same cancer types are ranked by decreasing mean error of the linear model.

https://doi.org/10.1371/journal.pcbi.1005921.g005

cancer types (see S13 Fig for an example group of 996 genes identified by a regression tree learned in one PRAD sample) are functionally enriched for general and widespread biological processes (S4 Table). In contrast, groups well predicted in only certain cancers were associated to specific biological function. For instance, a regression tree learned on one PAAD sample identified a group of 1,531 genes, which has low prediction error in LGG and PAAD samples but high error in LAML, LIHC and DLBC samples (Fig 6 and S13 Fig). Functional annotation of this group showed that, in contrast to the group described above (S13 Fig and S4 Table), this group is also linked to specific biological processes (S5 Table).

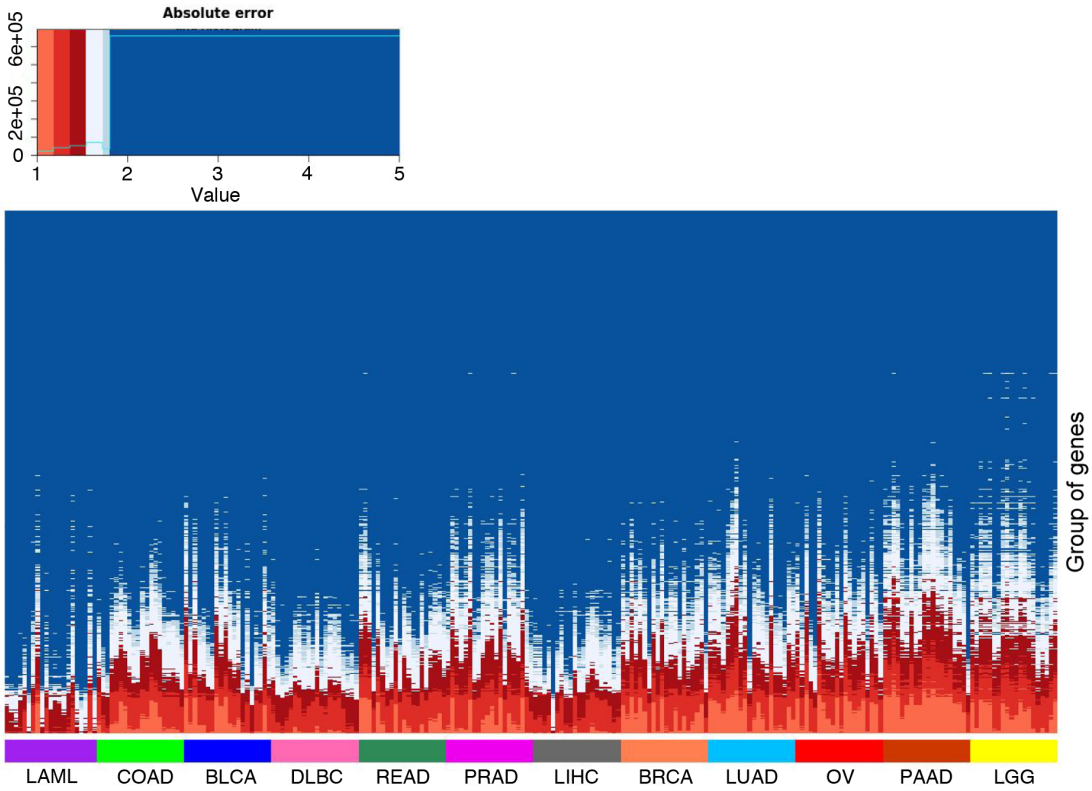


Fig 6. Gene classification according to prediction accuracy. Columns represent the various samples gathered by cancer type. Samples from the same cancer type are ranked by decreasing mean squared prediction error. Lines represent the 3,680 groups of gene obtained with the regression trees (one tree for each of the 241 samples) ranked by decreasing mean squared prediction error. Groups gathering the top 25% well predicted genes (error $< \sim 1.77$) are indicated in red and light blue.

https://doi.org/10.1371/journal.pcbi.1005921.g006

We further computed Gini coefficient for 16,134 genes using 8,556 GTEx libraries [55]. Gini coefficient measures statistical dispersion which can be used to measure gene expression ubiquity: value 0 represents genes expressed in all samples, while value 1 represents genes expressed in only one sample. We observed that the correlations obtained between Gini coefficient and model errors in each TCGA sample ranged from 0.22 to 0.36. We also compared model errors associated to first and last quartiles of the Gini coefficient distribution using a Wilcoxon test for each of the 241 samples. The test was invariably significant with maximum p-value = $2.881e^{-7}$. Likewise analyses were performed with 1,897 FANTOM CAGE libraries [56] considering 15,904 genes. In that case, correlation between models errors and Gini coefficients ranged from 0.25 to 0.4. Overall these analyses suggested that our model better predicts expression of highly and ubiquitously expressed genes. We do not exclude that, when predicting tissue-specific genes, ChIP-seq data collected from the same tissue may add explanatory power to the sequence model. Note, however, that the model performances vary between cancer and cell types implying that part of cell-specific genes are also well predicted by the model (S9 Fig).

Relationships between selected nucleotide composition and genome architecture

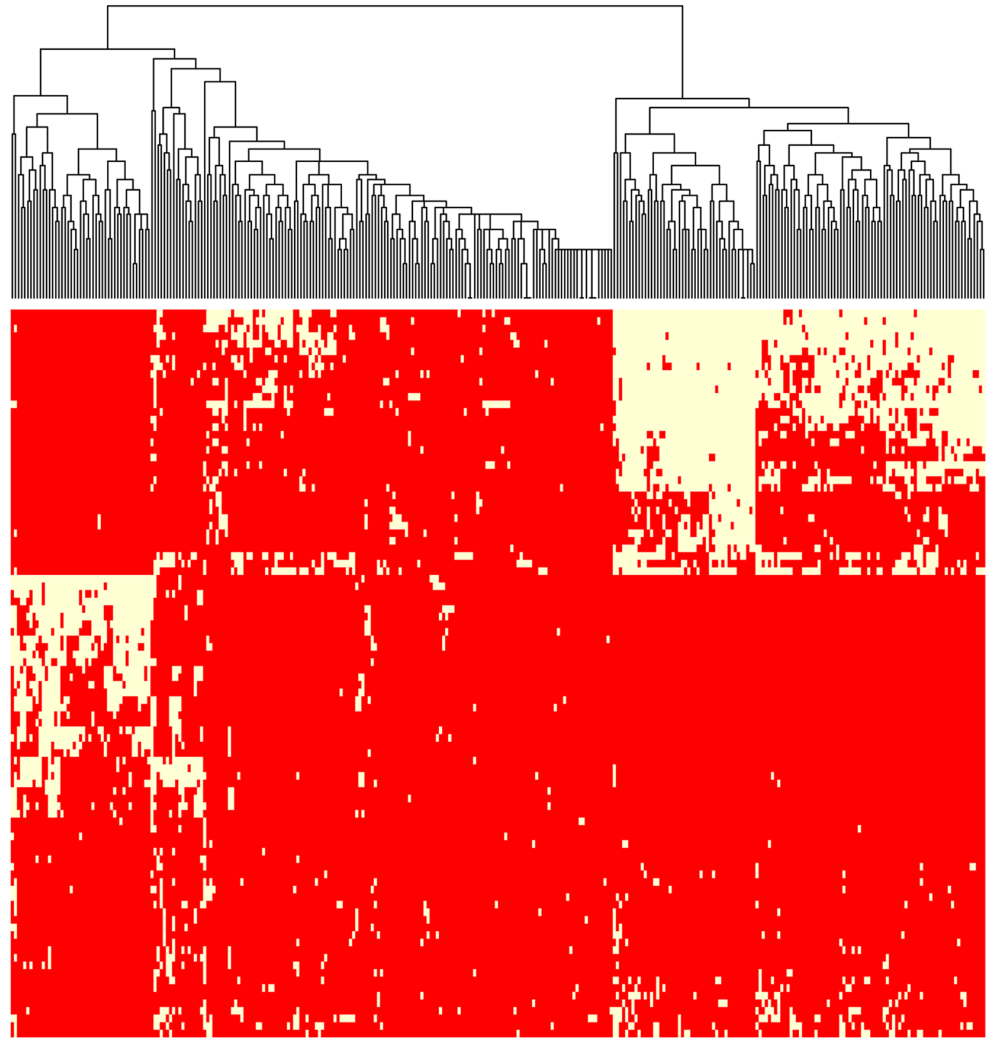
We probed the regulatory activities of the selected regions. We first determined whether introns contained specific regulatory sequence code by assessing the presence of *cis* expression

quantitative trait loci (*cis*-eQTLs). Zhou *et al.* indeed showed that the effect of eQTL SNPs can be predicted from a regulatory sequence code learned from genomic sequences [25]. These findings also implied that *cis*-eQTLs preferentially affect DNA sequences at precise locations (e.g. TF binding sites) rather than global nucleotide composition (i.e. nucleotide/dinucleotide percentages used as variables in our model). We used the v6p GTEx release to compute the average frequencies of *cis*-eQTLs present in the considered genomic regions and directly linked to their host genes (S6 Table). We noticed that introns contained the smallest density of *cis*-eQTLs (10 times less than any other regions), while containing comparable amount of SNPs (S7 Table). This result argued against the presence of a regulatory sequence code similar to that observed in promoters for instance [25], despite the presence of enhancers (S8 Table. These results rather unveiled the existence of another layer of intron-mediated regulation, which involves global nucleotide compositions of larger DNA regions. We then asked whether the groups of genes identified by the regression trees (Fig 6) correspond to specific TADs. Genes within the same TAD tend to be coordinately expressed [57, 58]. TADs with similar chromatin states tend to associate to form two genomic compartments called A and B: A contains transcriptionally active regions while B corresponds to transcriptionally inactive regions [59]. The driving forces behind this compartmentalization and the transitions between compartments observed in different cell types are not fully understood, but chromatin composition and transcription are supposed to play key roles [5]. Jabbari and Bernardi showed that nucleotide composition along the genome (notably isochores) can help define TADs [60]. As intronic sequences represent $\sim$ 50% of the human genome (1,512,685,844 bp out of 3,137,161,264 according to ENSEMBL merged intron coordinates), the nucleotide composition of introns likely resemble that of neighbor genes and more globally that of the corresponding TAD. We used the 373 TADs containing more than 10 genes mapped in IMR90 cells [6]. For each TAD and each (di)nucleotide, we used a Kolmogorov-Smirnov test to compare the (di)nucleotide distribution of the embedded genes with that of all other genes. We used a Benjamini-Hochberg multiple testing correction to control the False Discovery Rate (FDR), which was fixed at 0.05 (see Materials and methods section). We found that 324 TADs out of 373 ($\sim$87%) are characterized by at least one specific nucleotide signature (Fig 7A). In addition, our results clearly showed the existence of distinct classes of TADs related to GC content (GC-rich, GC-poor and intermediate GC content) (Fig 7A), in agreement with [60]. We next considered the 967 groups of genes defined in Fig 6 whose expression is accurately predicted by our model (i.e. groups with mean error $<$ mean error of the 1st quartile). We thus focused our analyses on genes for which we did learn some regulatory features. We evaluated the enrichment for specific TADs in each group (considering only TADs containing more than 10 genes) using an hypergeometric test (Fig 7B). We found that 60% of these groups were enriched for at least one TAD (p-value $<$ 0.05). Hence, several groups of genes identified by the regression trees (Fig 6) do correspond to specific TADs (Fig 7B). We concluded that our model, primarily based on intronic sequences, select gene nucleotide compositions that better distinguish active TADs.

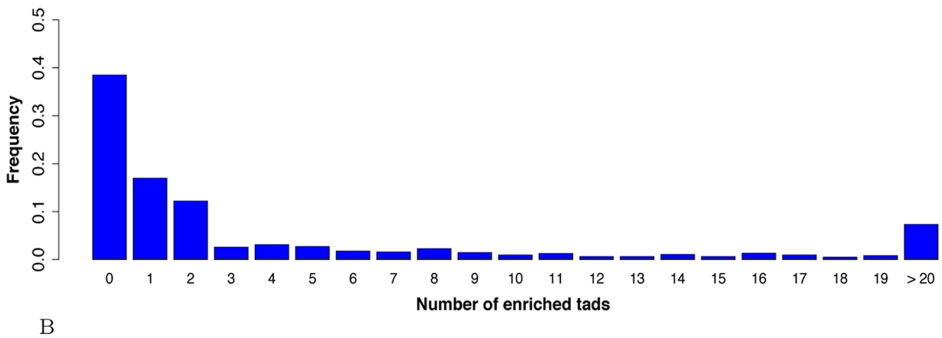
Discussion

In this study, we corroborate the hypothesis that DNA sequence contains information able to explain gene expression [20–25]. We built a global regression model to predict, in any given sample, the expression of the different genes using only nucleotide compositions as predictive variables. Overall our model provided a framework to study gene regulation, in particular the influence of regulatory regions and their associated nucleotide composition.

A surprising result of our study is that sequence-level information is highly predictive of gene expression and in some occasions comparable to reference ChIP-seq data alone [17, 19].



A



B

Fig 7. A: Nucleotide compositions of resident genes distinguish TADs. For each TAD and for each region considered, the percentage of each nucleotide and dinucleotide associated to the embedded genes were compared to that of all other genes using a Kolmogorov-Smirnov test. Red indicates FDR-corrected p-value ≥ 0.05 and yellow FDR-corrected p-value < 0.05. TAD clustering was made using this binary information. Only TADs with at least one p-value < 0.05 are shown (i.e. 87% of the TADs containing at least 10 genes). y-axis from top to bottom: G_INTR, GpC_INTR, CpC_INTR, CpC_3UTR,

GpC_3UTR, G_3UTR, GpC_CDS, CpC_CDS, G_CDS, G_DFR, CpC_DFR, GpC_DFR, CpG_INTR, CpG_3UTR, CpG_CDS, CpG_DFR, G_DU, GpC_DD, CpG, DU, CpG_DD, GpC_DU, CpC_DU, CpC_DD, G_DD, GpC_5UTR, CpG_5UTR, G_5UTR, GpC_CORE, CpG_CORE, CpC_CORE, G_CORE, CpC_5UTR, CpT_3UTR, CpT_CDS, CpT_INTR, ApT_INTR, TpA_INTR, A_INTR, ApA_INTR, TpA_3UTR, ApT_3UTR, A_3UTR, ApA_3UTR, ApA_CDS, A_CDS, ApT_CDS, TpA_CDS, A_DD, ApA_DD, ApT_DD, TpA_DD, TpA_DU, ApT_DU, ApA_DU, A_DU, TpA_DFR, ApT_DFR, A_DFR, ApA_DFR, ApA_CORE, A_CORE, ApT_CORE, TpA_CORE, ApA_5UTR, ApT_5UTR, A_5UTR, TpA_5UTR, ApC_DFR, ApC_DD, ApC_DU, TpC_DU, TpC_DFR, ApC_CORE, CpA_DU, CpA_DFR, CpA_CDS, ApC_CDS, ApC_3UTR, TpC_CDS, TpC_CORE, CpT_5UTR, TpC_5UTR, CpT_CORE, TpC_DD, CpA_CORE, ApC_5UTR, CpA_5UTR, ApC_INTR, CpA_DD, CpT_DFR, CpT_DD, CpT_DU, TpC_3UTR, TpC_INTR, CpA_INTR, CpA_3UTR. **B: TAD enrichment within groups of genes whose expression is accurately predicted by our model.** The enrichment for each TAD (containing more than 10 genes) in each gene group accurately predicted by our model (i.e. groups with mean error < mean errors of the 1st quartile) was evaluated using an hypergeometric test. The fraction of groups with enriched TADs (p-value < 0.05) is represented.

https://doi.org/10.1371/journal.pcbi.1005921.g007

The similar accuracy of models built on real and randomly permuted experimental data indicated that, though the experimental data are biologically relevant, their interpretation through a linear model, in particular inference of TF combinations, is not straightforward as randomization of experimental data did not show the expected loss of accuracy (Fig 3). An interesting perspective would be to devise a strategy to infer TF combinations from experimental data without being influenced by the opening of the chromatin.

The accuracy of our model confirmed that DNA sequence *per se* and basic information like dinucleotide frequencies have very high predictive power. It remains to determine the exact nature of these sequence-level instructions. Interestingly, nucleotide environment contributes to prediction of TF binding sites and motifs bound by a TF have a unique sequence environment that resembles the motif itself [40]. Hence, the potential of the nucleotide content to predict gene expression may be related to the presence of regulatory motifs and TFBSs. However, we showed that the gene body (introns, CDS and UTRs), as opposed to sequences located upstream (promoter) or downstream (DFR), had the most significant contribution in our model. Moreover, *cis*-eQTL frequencies argue against the presence of a regulatory sequence code in introns similar to that observed in promoters, suggesting the existence of another layer of regulation implicating the nucleotide composition of large DNA regions.

Gene nucleotide compositions vary across the genome and can even help define TAD boundaries [60]. In line with [60], we showed that genes located within the same TAD share similar nucleotide compositions, which provides a nucleotide signature for their TADs (Fig 7A). Our model aimed at predicting gene expression, and therefore intimately linked to TAD compartmentalization, appeared to capture these signatures. Several studies have already demonstrated the existence of sequence-level instructions able to determine genomic interactions. Using an SVM-based approach, Nikumbh *et al* demonstrated that sequence features can determine long-range chromosomal interactions [61]. Similar results were obtained by Singh *et al.* using deep learning-based models [62]. Using biophysical approaches, Kornyshev *et al.* showed that sequence homology influences physical attractive forces between DNA fragments [63]. It would be interesting to determine whether the nucleotide signatures identified by our model are directly implicated in DNA folding and 3D genome architecture.

Finally, although sequence-level instructions are—almost—identical in all cells of an individual, their usage must be cell-type specific to allow proper A/B compartimentalization of TADs, gene expression and ultimately diversity of cell functions. At this stage, the mechanisms driving this cell-type specific selection of nucleotide compositions remain to be characterized.

Supporting information

S1 Fig. Comparison of models built on maximum or sum PWM motif scores. The model was built (i) using 60 nucleotide/dinucleotide percentages computed in the 3 promoter

segments (CORE+DU+DD) and 471 JASPAR2016 PWM maximum scores computed in the CORE segment (pink) or (ii) using 60 nucleotide/dinucleotide percentages computed in the 3 promoter segments (CORE+DU+DD) and 471 JASPAR2016 PWM sum scores computed in the CORE segment (green). All sequences were centered around the 2nd TSS and the 2 models were fitted on 16,294 genes for each of the 241 samples.
(PDF)

S2 Fig. Dinucleotide local distribution around GENCODEv24 TSSs. Dinucleotide percentages (y-axis) along 140,604 DNA regions centered around GENCODE v24 TSSs ±2000 bp (the distance to TSS is shown in the x-axis). Dinucleotide combinations are represented as first nucleotide on left and second nucleotide on top. The promoter segmentation used in this study (Fig 1) is indicated with vertical dashed lines at -500 bp and 500 bp from the TSS.
(PDF)

S3 Fig. Number of TSSs by gene. We considered 19,393 TCGA genes listed in TCGA and the TSSs annotated by GENCODE v24.
(PDF)

S4 Fig. Contribution in the model of the TSS number. The model is built using 20 variables corresponding to the nucleotide (4) and dinucleotide (16) percentages computed in the CORE promoter (red), DU (green) or DD (yellow) centered around the second TSS as predictive variables (green). Linear models are also built on the number of isoforms (dark pink) and the number of TSSs (dark blue). Finally models are built using the combinations of variables indicated. All different models were fitted on 19,393 genes for each of the 241 samples considered. The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients between observed and predicted gene expressions. The correlations obtained in all samples are shown as violin plots. These two last plots underscored the importance of these two variables in predicting gene expression.
(PDF)

S5 Fig. Gene expression distribution and FANTOM5 enhancer association. The 19,393 genes listed in one LAML sample (TCGA.AB.2939.03A.01T.0740.13_LAML) (pink) and a subset of 11,359 genes with assigned FANTOM enhancers (green) were considered. The median expression of genes with assigned enhancers is greater than that of all genes (wilcoxon test p-value $<$ 2.2e-16)
(PDF)

S6 Fig. Accuracies of models built on dsDNA or ssDNA. A: Models were built using nucleotide and dinucleotide percentages computed on dsDNA (2 nucleotides + 8 dinucleotides; green violin) or on ssDNA (4 nucleotides + 16 dinucleotides; purple violin) in all the regulatory regions (CORE, DU, DD, 5UTR, CDS, 3UTR, INTR, DFR). The 2 models were fitted on 16,294 genes for each of the 241 samples. The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients. **B:** Same analyses focusing on each of the indicated regions.
(PDF)

S7 Fig. Model accuracy with different set of nucleotide predictive variables. A: Models were built using different set of variables including nucleotide (4 x 8 regions), dinucleotide (16 x 8 regions) and/or trinucleotide (64 x 8 regions) percentages computed in all the regulatory regions (CORE, DU, DD, 5UTR, CDS, 3UTR, INTR, DFR). All different models were fitted on 16,280 genes for each of the 241 samples considered. The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients. **B:** Models were built using

nucleotide (4 x 8 regions) and dinucleotide (16 x 8 regions) percentages computed in all the regulatory regions and trinucleotide (64) percentages computed in each of the indicated region separately.
(PDF)

S8 Fig. Forward selection procedure with models built on isoform expressions. The procedure is identical to that described in Fig 4 but models were built on isoform-specific variables and correlations were computed between observed and predicted isoform expression, not gene expression.
(PDF)

S9 Fig. Model accuracy in different cancer types. The model with 160 variables (20 (di)nucleotide rates in 8 regions) was built on 16,294 genes in 241 samples corresponding to the initial training set corresponding to 12 cancer types (**A**) and in an additional set of 1,270 samples corresponding to 14 different cancer types (**B**). The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients between observed and predicted gene expressions. The correlations obtained in all samples of each data sets are shown as violin plots in **A** (training set) and **B** (additional set). The color code indicates the cancer types. The horizontal dashed lines indicates the median correlation (**A**, 0.582; **B**, 0.577).
(PDF)

S10 Fig. Comparison on models built on RNA-seq or microarray data. The model with 160 variables (20 (di)nucleotide rates in 8 regions) was built on 9,791 genes in 582 samples with matched RNA-seq and microarray data. The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients between observed and predicted gene expressions. The correlations obtained in all samples with RNA-seq- or microarray-built models are shown as violin plots.
(PDF)

S11 Fig. Spearman correlations between CNV segment mean score and model prediction error. CNV absolute segment mean scores were computed for each as explained in Materials and Methods section. Model prediction absolute error for each gene are given by our predictive model using nucleotide and dinuclotide percentages computed in all the regulatory regions. Models were fitted on 16,294 genes for each of the 234 on 241 samples having CNV TCGA data available. The median correlation for the 234 samples is -0.014.
(PDF)

S12 Fig. Absolute values of the regression coefficients. A linear regression model was built, for each sample, on standardized stable variables only. The boxplots show absolute values of the corresponding coefficients in all samples for each variable considered. Color code as in Fig 5. CpG in the core promoter is highlighted in white. Purple line represents the median of CpG_CORE coefficients.
(PDF)

S13 Fig. Example of regression trees learned on two linear models. A: Regression tree leading to a group of genes well predicted in all samples. This tree has been learned on the sample TCGA.FC.A5OB.01A.11R.A29R.07_PRAD using all nucleotide composition in all regions. The red path defines a group of 996 genes which has low Lasso error in all samples and cancer types. This group was used for functional annotation (S4 Table). **B: Regression tree leading to a group of genes well predicted in LGG and PPAD samples.** This tree has been learned on the sample TCGA.IB.7646.01A.11R.2156.07_PAAD using all nucleotide composition in all

regions. The red path defines a group of 1,531 genes which has low Lasso error in LGG and PAAD samples but high error in LAML, LIHC and DLBC samples. This group was used for functional annotation (S5 Table).
(PDF)

S1 Table. Model comparison. Each model is fitted for each tumor, using all the variables over all regions (160 variables among 8 regulatory regions). First and second columns are median correlation and mean square error over all the tumors. The third column represents mean computing time per tumor (in minutes) on a standard laptop.
(PDF)

S2 Table. Contributions of additional genomic regions. Genomic regions were ranked according to their contribution in predicting gene expression. First, all regions were tested separately. Introns yielded the highest Spearman correlation between observed and predicted expressions and was selected as the 'first' seed region. Second, each region not already in the model was added separately. 5UTR in association with introns yielded the best correlation and was therefore selected as the 'second' region. Third, the procedure was repeated till all regions were included in the model. The contribution of each region is then visualized starting from the most important (left) to the less important (right). The correlations computed at each steps are indicated.
(PDF)

S3 Table. Correlations between observed and predicted isoform expression. The procedure is identical to that described in S2 Table but models were built on isoform-specific variables and correlations were computed between observed and predicted isoform expression, not gene expression.
(PDF)

S4 Table. Functional enrichment of a group of genes well predicted in all samples. The group of 996 genes is obtained by fitting a regression tree on the sample TCGA.FC.A5OB.01A.11R.A29R.07_PRAD using all the nucleotide composition in all regions. These genes are well predicted (mean error $<$ 1st quartile) for all samples of different type cancers. This group of genes was further annotated using the DAVID functional annotation tool. Only the top 5 biological processes indicated by DAVID is shown. The GO term yielded by this analysis corresponded to general and widespread biological processes indicating that these genes likely corresponded to housekeeping genes.
(PDF)

S5 Table. Functional enrichment of a group of genes well predicted in LGG and PAAD. The group of 1,531 genes is obtained by fitting a regression tree on the sample TCGA.IB.7646.01A.11R.2156.07_PAAD using all the nucleotide composition in all regions. These genes are well predicted (mean error $<$ 1st quartile) for all LGG and PAAD samples but not that of LAML, DBLC and LIHC. This group of genes was further annotated using the DAVID functional annotation tool. Only the top 5 biological processes indicated by DAVID is shown. The GO term "Nervous system development" indicates that these genes can be involved in specific biological processes.
(PDF)

S6 Table. Frequencies of *cis*-eQTLs in the genomic regions considered. We computed the density of *cis*-eQTL per regulatory region by dividing the sum of *cis*-eQTLs intersecting with the region considered for all genes by the sum of the lengths of the same regulatory region of

all genes. see Material and methods for details.
(PDF)

S7 Table. Frequencies of SNPs in CORE and INTRON regions. We computed the density of SNPs per regulatory region by dividing the sum of SNPs intersecting with the region considered for all genes by the sum of the lengths of the same regulatory region of all genes. We only considered SNPs detected on chromosomes 1, 2 and 19. see Material and methods for details.
(PDF)

S8 Table. Intersection between enhancers and the genomic regions considered. We computed the density of enhancers per regulatory region by dividing the total length of the intersection between the enhancers and the region considered for all genes by the sum of the lengths of the same regulatory region of all genes. see Material and methods for details.
(PDF)

Acknowledgments

We thank Mohamed Elati, Mathieu Lajoie, Anthony Mathelier and Cédric Notredame for insightful discussions and suggestions. We also thank Yue Li, Zhaolei Zhang, Florian Schmidt and Marcel H. Schulz for sharing data. We are indebted to the researchers around the globe who generated experimental data and made them freely available. C-H.L. is grateful to Marc Piechaczyk, Edouard Bertrand, Anthony Mathelier and Wyeth W. Wasserman for continued support.

Author Contributions

Conceptualization: Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

Formal analysis: Chloé Bessière, May Taha, Florent Petitprez, Jimmy Vandel.

Funding acquisition: Jean-Michel Marin, Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

Investigation: Chloé Bessière, May Taha, Florent Petitprez, Jimmy Vandel, Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

Methodology: Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

Project administration: Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

Supervision: Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

Validation: Chloé Bessière, May Taha.

Writing – original draft: Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

Writing – review & editing: Chloé Bessière, May Taha, Florent Petitprez, Jimmy Vandel, Jean-Michel Marin, Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

References

1. Andersson R, Sandelin A, Danko CG. A unified architecture of transcriptional regulatory elements. Trends in genetics: TIG. 2015; 31(8):426–433. https://doi.org/10.1016/j.tig.2015.05.007 PMID: 26073855
2. Babu D, Fullwood MJ. 3D genome organization in health and disease: emerging opportunities in cancer translational medicine. Nucleus (Austin, Tex). 2015; 6(5):382–393.

3. Ea V, Baudement MO, Lesne A, Forné T. Contribution of Topological Domains and Loop Formation to 3D Chromatin Organization. Genes. 2015; 6(3):734–750. https://doi.org/10.3390/genes6030734 PMID: 26226004
4. Gonzalez-Sandoval A, Gasser SM. On TADs and LADs: Spatial Control Over Gene Expression. Trends Genet. 2016;. https://doi.org/10.1016/j.tig.2016.05.004 PMID: 27312344
5. Merkenschlager M, Nora EP. CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. Annu Rev Genomics Hum Genet. 2016; 17:17–43. https://doi.org/10.1146/annurev-genom-083115-022339 PMID: 27089971
6. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485(7398):376–380. https://doi.org/10.1038/nature11082 PMID: 22495300
7. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014; 507(7493):455–461. https://doi.org/10.1038/nature12787 PMID: 24670763
8. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007; 316(5830):1497–1502. https://doi.org/10.1126/science.1141319 PMID: 17540862
9. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, et al. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell. 2011; 147(6):1270–1282. https://doi.org/10.1016/j.cell.2011.10.053 PMID: 22153072
10. Ray D, Kazan H, Chan ET, Pena Castillo L, Chaudhry S, Talukder S, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol. 2009; 27(7):667–670. https://doi.org/10.1038/nbt.1550 PMID: 19561594
11. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004; 5(4):276–287. https://doi.org/10.1038/nrg1315 PMID: 15131651
12. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell. 2010; 140(5):744–752. https://doi.org/10.1016/j.cell.2010.01.044 PMID: 20211142
13. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet. 2014; 15(12):829–845. https://doi.org/10.1038/nrg3813 PMID: 25365966
14. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57–74. https://doi.org/10.1038/nature11247
15. Lundberg SM, Tu WB, Raught B, Penn LZ, Hoffman MM, Lee SI. ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data. Genome Biol. 2016; 17:82. https://doi.org/10.1186/s13059-016-0925-0 PMID: 27139377
16. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res. 2012; 22(9):1658–1667. https://doi.org/10.1101/gr.136838.111 PMID: 22955978
17. Li Y, Liang M, Zhang Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. PLoS Comput Biol. 2014; 10(10):e1003908. https://doi.org/10.1371/journal.pcbi.1003908 PMID: 25340776
18. Jiang P, Freedman ML, Liu JS, Liu XS. Inference of transcriptional regulation in cancers. Proc Natl Acad Sci USA. 2015; 112(25):7731–7736. https://doi.org/10.1073/pnas.1424272112 PMID: 26056275
19. Schmidt F, Gasparoni N, Gasparoni G, Gianmoena K, Cadenas C, Polansky JK, et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. Nucleic Acids Res. 2017; 45(1):54–66. https://doi.org/10.1093/nar/gkw1061 PMID: 27899623
20. Quante T, Bird A. Do short, frequent DNA sequence motifs mould the epigenome? Nat Rev Mol Cell Biol. 2016; 17(4):257–262. PMID: 26837845
21. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of genetic variants that affect histone modifications in human cells. Science. 2013; 342(6159):747–749. https://doi.org/10.1126/science.1242429 PMID: 24136359
22. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science. 2013; 342(6159):744–747. https://doi.org/10.1126/science.1242463 PMID: 24136355
23. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, et al. Extensive variation in chromatin states across humans. Science. 2013; 342(6159):750–752. https://doi.org/10.1126/science.1242510 PMID: 24136358

24. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nat Methods. 2015; 12(3):265–272. https://doi.org/10.1038/nmeth.3065 PMID: 25240437

25. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015; 12(10):931–934. https://doi.org/10.1038/nmeth.3547 PMID: 26301843

26. Raghava GP, Han JH. Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. BMC Bioinformatics. 2005; 6:59. https://doi.org/10.1186/1471-2105-6-59 PMID: 15773999

27. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014; 47:1–34.

28. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016; 44(D1):D110–115. https://doi.org/10.1093/nar/gkv1176 PMID: 26531826

29. Chiu TP, Comoglio F, Zhou T, Yang L, Paro R, Rohs R. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics. 2016; 32(8):1211–1213. https://doi.org/10.1093/bioinformatics/btv735 PMID: 26668005

30. Jiao X, Sherman BT, Huang daW, Stephens R, Baseler MW, Lane HC, et al. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics. 2012; 28(13):1805–1806. https://doi.org/10.1093/bioinformatics/bts251 PMID: 22543366

31. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996; p. 267–288.

32. R Core Team. R: A Language and Environment for Statistical Computing; 2013. Available from: http://www.R-project.org/.

33. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks; 1984.

34. Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

35. Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010; 72(4):417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

36. Sill M, Hielscher T, Becker N, Zucknick M, et al. c060: Extended inference with lasso and elastic-net regularized Cox and generalized linear models. Journal of Statistical Software. 2015; 62(5).

37. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995; p. 289–300.

38. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet. 2012; 13(4):233–245. PMID: 22392219

39. Nguyen TA, Jones RD, Snavely A, Pfenning A, Kirchner R, Hemberg M, et al. High-throughput functional comparison of promoter and enhancer activities. Genome Res. 2016;. https://doi.org/10.1101/gr.204834.116

40. Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. Genome Res. 2015; 25(9):1268–1280. https://doi.org/10.1101/gr.184671.114 PMID: 26160164

41. Diamanti K, Umer HM, Kruczyk M, Dąbrowski MJ, Cavalli M, Wadelius C, et al. Maps of context-dependent putative regulatory regions and genomic signal interactions. Nucleic Acids Res. 2016;. https://doi.org/10.1093/nar/gkw800 PMID: 27625394

42. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013; 499(7457):172–177. https://doi.org/10.1038/nature12311 PMID: 23846655

43. Li X, Quon G, Lipshitz HD, Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA. 2010; 16(6):1096–1107. https://doi.org/10.1261/rna.2017210 PMID: 20418358

44. Auweter SD, Oberstrass FC, Allain FH. Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic Acids Res. 2006; 34(17):4943–4959.

45. Liu C, Mallick B, Long D, Rennie WA, Wolenc A, Carmack CS, et al. CLIP-based prediction of mammalian microRNA binding sites. Nucleic Acids Res. 2013; 41(14):e138. https://doi.org/10.1093/nar/gkt435 PMID: 23703212

46. Boel G, Letso R, Neely H, Price WN, Wong KH, Su M, et al. Codon influence on protein expression in E. coli correlates with mRNA levels. Nature. 2016; 529(7586):358–363. https://doi.org/10.1038/nature16509 PMID: 26760206

47. Bazzini AA, Del Viso F, Moreno-Mateos MA, Johnstone TG, Vejnar CE, Qin Y, et al. Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. EMBO J. 2016;. https://doi.org/10.15252/embj.201694699

48. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, et al. Codon optimality is a major determinant of mRNA stability. Cell. 2015; 160(6):1111–1124. https://doi.org/10.1016/j.cell.2015.02.029 PMID: 25768907

49. Chorev M, Carmel L. The function of introns. Front Genet. 2012; 3:55. https://doi.org/10.3389/fgene.2012.00055 PMID: 22518112

50. Rose AB. Intron-mediated regulation of gene expression. Curr Top Microbiol Immunol. 2008; 326:277–290. PMID: 18630758

51. Schwalb B, Michel M, Zacher B, Fruhauf K, Demel C, Tresch A, et al. TT-seq maps the human transient transcriptome. Science. 2016; 352(6290):1225–1228. https://doi.org/10.1126/science.aad9841 PMID: 27257258

52. Bunting KL, Soong TD, Singh R, Jiang Y, Beguelin W, Poloway DW, et al. Multi-tiered Reorganization of the Genome during B Cell Affinity Maturation Anchored by a Germinal Center-Specific Locus Control Region. Immunity. 2016; 45(3):497–512. https://doi.org/10.1016/j.immuni.2016.08.012 PMID: 27637145

53. Hayer KE, Pizarro A, Lahens NF, Hogenesch JB, Grant GR. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. Bioinformatics. 2015; 31(24):3938–3945. https://doi.org/10.1093/bioinformatics/btv488 PMID: 26338770

54. Breiman L, et al. Classification and Regression Trees. New York: Chapman & Hall; 1984.

55. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015; 348(6235):660–665. https://doi.org/10.1126/science.aaa0355 PMID: 25954002

56. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, et al. A promoter-level mammalian expression atlas. Nature. 2014; 507(7493):462–470. https://doi.org/10.1038/nature13182 PMID: 24670764

57. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012; 485(7398):381–385. https://doi.org/10.1038/nature11049 PMID: 22495304

58. Fanucchi S, Shibayama Y, Burd S, Weinberg MS, Mhlanga MM. Chromosomal contact permits transcription between coregulated genes. Cell. 2013; 155(3):606–620. https://doi.org/10.1016/j.cell.2013.09.051 PMID: 24243018

59. Lieberman-Aiden E, Berkum NLv, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009; 326(5950):289–293. https://doi.org/10.1126/science.1181369 PMID: 19815776

60. Jabbari K, Bernardi G. An Isochore Framework Underlies Chromatin Architecture. PLoS ONE. 2017; 12(1):e0168023. https://doi.org/10.1371/journal.pone.0168023 PMID: 28060840

61. Nikumbh S, Pfeifer N. Genetic sequence-based prediction of long-range chromatin interactions suggests a potential role of short tandem repeat sequences in genome organization. BMC Bioinformatics. 2017; 18(1):218. https://doi.org/10.1186/s12859-017-1624-x PMID: 28420341

62. Singh S, Yang Y, Poczos B, Ma J. Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks. BioRxiv. 2016;

63. Kornyshev AA, Leikin S. Sequence recognition in the pairing of DNA duplexes. Phys Rev Lett. 2001; 86(16):3666–3669. https://doi.org/10.1103/PhysRevLett.86.3666 PMID: 11328049

Annexe E

Dondelinger, F., Lèbre, S. and Husmeier, D. (2013)

Mach Learn (2013) 90:191–230
DOI 10.1007/s10994-012-5311-x

Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure

Frank Dondelinger · Sophie Lèbre · Dirk Husmeier

Received: 16 September 2011 / Revised: 19 March 2012 / Accepted: 15 June 2012 /
Published online: 18 July 2012

Abstract The proper functioning of any living cell relies on complex networks of gene regulation. These regulatory interactions are not static but respond to changes in the environment and evolve during the life cycle of an organism. A challenging objective in computational systems biology is to infer these time-varying gene regulatory networks from typically short time series of transcriptional profiles. While homogeneous models, like conventional dynamic Bayesian networks, lack the flexibility to succeed in this task, fully flexible models suffer from inflated inference uncertainty due to the limited amount of available data. In the present paper we explore a semi-flexible model based on a piecewise homogeneous dynamic Bayesian network regularized by gene-specific inter-segment information sharing. We explore different choices of prior distribution and information coupling and evaluate their performance on synthetic data. We apply our method to gene expression time series obtained during the life cycle of *Drosophila melanogaster*, and compare the predicted segmentation with other state-of-the-art techniques. We conclude our evaluation with an ap-

Editor: James Cussens.

Software: R code for all models described in this paper is available from http://www.bioss.ac.uk/students/frankd.html, and will be made available as an R package on the Comprehensive R Archive Network (CRAN) in the near future.

F. Dondelinger · D. Husmeier
Biomathematics and Statistics Scotland, JCMB, Edinburgh EH9 3JZ, UK

F. Dondelinger
Institute for Adaptive and Neural Computation, The University of Edinburgh, Edinburgh EH8 9AB, UK
e-mail: frankd@bioss.ac.uk

S. Lèbre
LSIIT, UMR 7005, Université de Strasbourg, 67412 Illkirch, France
e-mail: slebre@unistra.fr

D. Husmeier (✉)
School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QW, UK
e-mail: dirk.husmeier@glasgow.ac.uk

plication to synthetic biology, where the objective is to predict an in vivo regulatory network of five genes in *Saccharomyces cerevisiae* subjected to a changing environment.

Keywords Dynamic Bayesian networks · Hierarchical Bayesian models · Multiple changepoint processes · Reversible jump Markov chain Monte Carlo · Gene expression time series · Systems and synthetic biology

1 Introduction

One of the challenging problems in the field of systems biology is the inference of gene regulatory networks from high-throughput transcriptomic profiles, as obtained e.g. with microarrays or next generation sequencing. While protein interactions can be measured directly with various high-throughput assays (e.g. yeast-2-hybrid or phage display), gene regulatory interactions involve several intermediate steps related to the formation, activation and complex formation of transcription factors (e.g. via phosphorylation or dimerization). These processes are not observable at the transcriptional level. For that reason the inference of interactions has to be based on indirect noisy measurements of mRNA concentrations (a proxy for gene activity), rendering the problem of regulatory network reconstruction more difficult than for proteins. Various statistical techniques aim to perform network inference on this data, and the reconstructed regulation networks can reveal how the genes and the proteins they code for interact. However, many of the regulatory interactions in the cell vary in time. During the development and growth of an organism, some genes and pathways are more active during the early stages, but show practically no activity during the later stages, or vice-versa. *Drosophila melanogaster*, for instance, goes through several developmental stages, from embryo to larva to pupa to adult. Genes involved in wing muscle development would naturally fulfill different roles during the embryonal phase, when no wings are present, than they do in the adult fly, when the wings have fully developed. Another instance in which the gene regulatory network varies in time is in reaction to an environmental trigger, such as the type of growth substrate. Such a trigger can enhance or prevent the interactions of certain genes, which in turn can have repercussions for the whole gene network.

We are therefore presented with the problem of inferring a regulatory network from a series of discrete measurements or observations in time, where the structure of the network is subject to potential change. Moreover, we may not always know at which stage structural changes are likely to occur, as the underlying processes may be time-delayed or dependent on unobservable external factors. To extend conventional reverse engineering methods, which only aim to infer a single immutable regulatory network, our work builds on recent research in combining dynamic Bayesian networks (DBNs) with multiple changepoint processes (Robinson and Hartemink 2009, 2010; Grzegorczyk and Husmeier 2009, 2011; Lèbre 2007; Lèbre et al. 2010; Kolar et al. 2009). Below, we will briefly review the state of the art and the shortcomings of existing methods that we aim to address.

The standard assumption underlying DBNs is that time-series have been generated from a homogeneous Markov process. This assumption is too restrictive, as discussed above, and can potentially lead to erroneous conclusions. While there have been various efforts to relax the homogeneity assumption for undirected graphical models (Talih and Hengartner 2005; Xuan and Murphy 2007), relaxing this restriction in DBNs is a more recent research topic (Robinson and Hartemink 2009, 2010; Grzegorczyk and Husmeier 2009, 2011; Ahmed and Xing 2009; Lèbre 2007; Lèbre et al. 2010; Kolar et al. 2009). At present, none of the proposed methods is without its limitations, leaving room for further methodological

innovation. The method proposed in Ahmed and Xing (2009) and Kolar et al. (2009) is non-Bayesian. This requires certain regularization parameters to be optimized "externally", by applying information criteria (like AIC or BIC), cross-validation or bootstrapping. The first approach is suboptimal, the latter approaches are computationally expensive.[1] In the present paper we therefore follow the Bayesian paradigm, as in Robinson and Hartemink (2009, 2010), Grzegorczyk and Husmeier (2009, 2011), Lèbre (2007) and Lèbre et al. (2010). These approaches also have their limitations. The method proposed in Grzegorczyk and Husmeier (2009, 2011) assumes a fixed network structure and only allows the interaction parameters to vary with time. This assumption is too rigid when looking at processes where changes in the overall regulatory network structure are expected, e.g. in morphogenesis or embryogenesis. The method proposed in Robinson and Hartemink (2009, 2010) requires a discretization of the data, which incurs an inevitable information loss. These limitations are addressed in Lèbre (2007) and Lèbre et al. (2010), where the authors propose a method for continuous data that allows network structures associated with different nodes to change with time in different ways. However, this high flexibility causes potential problems when applied to time series with a low number of measurements, as typically available from systems biology, leading to overfitting or inflated inference uncertainty.

The objective of the present paper is to propose a novel model that addresses the methodological shortcomings of the three Bayesian methods mentioned above, and to demonstrate its viability by application to gene expression time series from *Drosophila melanogaster* and *Saccharomyces cerevisiae*. Unlike Robinson and Hartemink (2009, 2010), our model is continuous and therefore avoids the information loss inherent in a discretization of the data. We further improve on the model in Robinson and Hartemink (2009, 2010) by allowing for different penalties for changing edges and non-edges in the network, and by allowing different nodes in the networks to have different penalty terms. Unlike Grzegorczyk and Husmeier (2009, 2011), our model allows the network structure to change among segments, leading to greater model flexibility. As an improvement on Lèbre (2007) and Lèbre et al. (2010), our model introduces information sharing among time series segments, which provides an essential regularization effect. We have applied the model to reconstruct two regulatory networks: a network of genes involved in wing muscle development during the life cycle of *Drosophila melanogaster* (Arbeitman et al. 2002), and an engineered network from synthetic biology, consisting of five genes in *Saccharomyces cerevisiae* (Cantone et al. 2009).

The present paper follows on from two earlier conference papers of ours (Dondelinger et al. 2010; Husmeier et al. 2010). In Dondelinger et al. (2010), we compared two different information coupling paradigms: global information coupling and sequential information coupling. Global information coupling is appropriate when there is no natural sequential order of the time series segments, such as for segments derived from different experimental conditions. Sequential information sharing, which we investigated in more detail in Husmeier et al. (2010) and in the present paper, is appropriate for modelling a temporal developmental process, such as those related to morphogenesis, where changes to the network structure happen sequentially.

The present paper extends Dondelinger et al. (2010) and Husmeier et al. (2010) in several respects. Firstly, restricted by a strict page limit, our earlier papers were rather terse. The present paper provides a more comprehensive exposition of the methodology, which is self-contained. Secondly, we have explored different versions of information coupling (hard versus soft) and functional forms of the prior (exponential versus binomial). In Husmeier

[1] See Larget and Simon (1999) for a demonstration of the higher computational costs of bootstrapping over Bayesian approaches based on MCMC.

et al. (2010), not all combinations of strength versus functional form were investigated, and we have completed these combinations in our present work. Thirdly, we have improved the MCMC scheme. In our earlier work, a standard Metropolis-Hastings-Green (RJMCMC) sampler was employed. In the present work we have identified several scenarios where this sampler is bound to fail, and we propose a new type of MCMC proposal move. We show that these new moves avoid the convergence problems encountered with the original sampler, leading to a substantial improvement in mixing. Fourthly, the Bayesian hierarchical models that we propose depend on various hyperparameters. As opposed to our earlier work, we have investigated the influence of the higher level hyperparameters. To this end, we have first carried out a set of simulation studies for the proposed model. To substantiate our findings, we have then additionally carried out semi-analytical investigations for a simplified scenario, in which the computation of the marginal likelihood is tractable (see Sect. 5.2). Fifthly, we have rerun all our earlier simulations to understand the effect of model choice, unconfounded by MCMC mixing problems, and we have improved the interpretation of the results for the real-world problems.

We note that while we were extending our earlier work of Husmeier et al. (2010), a somewhat related paper has been published: Wang et al. (2011). While methodologically similar, there is an important difference in the application and inference, though. The objective of Wang et al. (2011) is online parameter estimation via particle filtering, with applications e.g. in tracking. This is a different scenario from most systems biology applications, where an interaction structure is typically learnt off-line after completion of a series of high-throughput experiments. Unlike Wang et al. (2011), our work thus follows other applications of DBNs in systems biology (Robinson and Hartemink 2009, 2010; Grzegorczyk and Husmeier 2009, 2011; Lèbre 2007; Lèbre et al. 2010; Kolar et al. 2009) and aims to infer the model structure by marginalizing out the parameters in closed form. To paraphrase this: while inference in Wang et al. (2011) is based on a filter, inference in our work is based on a smoother.

Our paper is organized as follows. Section 2 reviews the non-homogeneous DBN on which our work is based. Section 3 describes the methodological innovation of Bayesian regularization via information coupling. Section 4 describes the implementation of our method and the setup of the simulation studies. Section 5 discusses results obtained on synthetic data, with an investigation of the influence of the hyperparameters. Section 6 describes and interprets two real-world applications, related to morphogenesis in *Drosophila melanogaster* and synthetic biology in *Saccharomyces cerevisiae*. The paper concludes in Sect. 7 with a general discussion and summary.

2 Background: non-homogeneous DBNs

This section summarizes the auto regressive time-varying DBN proposed in Lèbre (2007) and Lèbre et al. (2010). A similar model was proposed in Punskaya et al. (2002). The idea is to combine the Bayesian regression model of Andrieu and Doucet (1999) with multiple changepoint processes and pursue Bayesian inference with reversible jump Markov chain Monte Carlo (RJMCMC) (Green 1995). We call this method TVDBN (Time-Varying Dynamic Bayesian Network).

The model is based on the first-order Markov assumption. This assumption is not critical, though, and a generalization to higher orders, as pursued in Punskaya et al. (2002), is straightforward. The value that a node in the graph takes on at time t is determined by the values that the node's parents (i.e. potential regulators, see below) take on at the previous

time point, $t-1$. More specifically, the conditional probability of the observation associated with a node at a given time point is a conditional Gaussian distribution, where the conditional mean is a linear weighted sum of the parent values at the previous time point, and the interaction parameters and parent sets depend on the time series segment. The latter dependence adds extra flexibility to the model and thereby relaxes the homogeneity assumption. The interaction parameters, the variance parameters, the number of potential parents, the location of changepoints demarcating the time series segments, and the number of changepoints are given (conjugate) prior distributions in a hierarchical Bayesian model. For inference, all these quantities are sampled from the posterior distribution with RJMCMC. Note that a complete specification of all node-parent configurations determines the structure of a regulatory network: each node receives incoming directed edges from each node in its parent set. In what follows, we will refer to nodes as genes and to the network as a gene regulatory network. The method is not restricted to molecular systems biology, though.

2.1 Graph

Let p be the number of observed genes, and let $\boldsymbol{x} = (x_i(t))_{1\le i\le p, 1\le t\le N}$ be the expression values measured at N time points. $\boldsymbol{G}^h$ represents a directed graph, i.e. the network defined by a set of directed edges among the p genes. $\boldsymbol{G}_i^h$ is the subnetwork associated with target gene i, determined by the set of its parents, i.e. the nodes with a directed edge feeding into gene i; these are the potential regulators of the target gene. The meaning of the superscript h is explained in the next section.

2.2 Multiple changepoint process

The set of regulatory relationships among the genes, defined by $\boldsymbol{G}^h$, may vary across time, which we model with a multiple changepoint process. For each target gene i, an unknown number k_i of changepoints define k_i+1 non-overlapping segments. Segment $h = 1, .., k_i+1$ starts at changepoint ξ_i^{h-1} and stops before ξ_i^h, where $\boldsymbol{\xi}_i = (\xi_i^0, \ldots, \xi_i^{h-1}, \xi_i^h, \ldots, \xi_i^{k_i+1})$ with $\xi_i^{h-1} < \xi_i^h$. To delimit the bounds, two pseudo-changepoints are introduced: $\xi_i^0 = 2$ and $\xi_i^{k_i+1} = N+1$. Thus vector $\boldsymbol{\xi}_i$ has length $|\boldsymbol{\xi}_i| = k_i+2$. The set of changepoints is denoted by $\boldsymbol{\xi} = (\boldsymbol{\xi}_i)_{1\le i\le p}$. This changepoint process induces a partition of the time series, $\boldsymbol{x}_i^h = (x_i(t))_{\xi_i^{h-1}\le t<\xi_i^h}$, with different network structures $\boldsymbol{G}_i^h$ associated with the different segments $h \in \{1, \ldots, k_i+1\}$. Identifiability is satisfied by ordering the changepoints based on their position in the time series. We define $\boldsymbol{G}_i = \{\boldsymbol{G}_i^h\}_{1\le h\le k_i+1}$ and $\boldsymbol{G} = \{\boldsymbol{G}_i\}_{1\le i\le p}$.

2.3 Regression model

For each gene i, the random variable $X_i(t)$ refers to the expression of gene i at time t. Within any segment h, the expression of gene i depends on the p gene expression values measured at the previous time point through a regression model defined by (a) a set of s_i^h parents denoted by $\boldsymbol{G}_i^h = \{j_1, \ldots, j_{s_i^h}\} \subseteq \{1, \ldots, p\}$, $|\boldsymbol{G}_i^h| = s_i^h$, and (b) a set of parameters $(\boldsymbol{a}_i^h,\ \sigma_i^h)$ where $\boldsymbol{a}_i^h = (a_{ij}^h)_{0\le j\le p}$, $a_{ij}^h \in \mathbb{R}$ and $\sigma_i^h > 0$. For all $j \neq 0$, $a_{ij}^h = 0$ if $j \notin \boldsymbol{G}_i^h$. For each gene i, for each time point t in segment h ($\xi_i^{h-1} \le t < \xi_i^h$), the random variable $X_i(t)$ depends on the p variables $\{X_j(t-1)\}_{1\le j\le p}$ according to

$$X_i(t) = a_{i0}^h + \sum\nolimits_{j\in\boldsymbol{G}_i^h} a_{ij}^h\ X_j(t-1) + \varepsilon_i^h(t) \tag{1}$$

where the noise $\varepsilon_i^h(t)$ is assumed to be Gaussian with mean 0 and variance $(\sigma_i^h)^2$, $\varepsilon_i^h(t) \sim N(0, (\sigma_i^h)^2)$. We define $\boldsymbol{a}_i = (\boldsymbol{a}_i^h)_{1\le h\le k_i+1}$, $\boldsymbol{a} = (\boldsymbol{a}_i)_{0\le i\le p}$, $\boldsymbol{\sigma_i^2} = (\sigma_i^h)^2_{1\le h\le k_i+1}$ and $\boldsymbol{\sigma^2} = (\boldsymbol{\sigma_i^2})_{0\le i\le p}$.

2.4 Prior

The $k_i + 1$ segments are delimited by k_i changepoints, where k_i is distributed a priori as a truncated Poisson random variable with mean λ and maximum $\overline{k} = N - 2$:

$$P(k_i|\lambda) \propto \frac{\lambda^{k_i}}{k_i!}\mathbb{1}_{\{k_i\le\overline{k}\}}; \qquad P(\boldsymbol{k}|\lambda) = \prod_{i=1}^{p} P(k_i|\lambda) \tag{2}$$

where $\boldsymbol{k} = (k_1, \ldots, k_p)$. Conditional on k_i changepoints, the changepoint position vector $\boldsymbol{\xi}_i = (\xi_i^0, \xi_i^1, \ldots, \xi_i^{k_i+1})$ takes non-overlapping integer values, which we take to be uniformly distributed a priori. There are $(N-2)$ possible positions for the k_i changepoints, thus vector $\boldsymbol{\xi}_i$ has prior density:

$$P(\boldsymbol{\xi}_i|k_i) = 1\Big/\binom{N-2}{k_i} = \frac{k_i!(N-2-k_i)!}{(N-2)!} \tag{3}$$

For each gene i, for each segment h, the number s_i^h of parents for node i follows a truncated Poisson distribution with mean Λ and maximum $\overline{s} = 5$:

$$P\big(s_i^h\big|\Lambda\big) \propto \frac{\Lambda^{s_i^h}}{s_i^h!}\mathbb{1}_{\{s_i^h\le\overline{s}\}} \tag{4}$$

Conditional on s_i^h, the prior for the parent set $\boldsymbol{G}_i^h$ is a uniform distribution over all parent sets with cardinality s_i^h,

$$P\big(\boldsymbol{G}_i^h\big|s_i^h\big) = 1\Big/\binom{p}{s_i^h} = \frac{s_i^h!(p-s_i^h)!}{p!} \tag{5}$$

The overall prior on the network structures is given by marginalization:

$$P\big(\boldsymbol{G}_i^h\big|\Lambda\big) = \sum\nolimits_{s_i^h=0}^{\overline{s}} P\big(\boldsymbol{G}_i^h\big|s_i^h\big)P\big(s_i^h\big|\Lambda\big) \tag{6}$$

Conditional on the parent set $\boldsymbol{G}_i^h$ of size s_i^h, the $s_i^h + 1$ regression coefficients form a subset of $\boldsymbol{a}_i^h$ denoted by $\boldsymbol{a}_{\boldsymbol{G}_i^h} = (a_{i0}^h, (a_{ij}^h)_{j\in\boldsymbol{G}_i^h})$. They are assumed zero-mean multivariate Gaussian with covariance matrix $(\sigma_i^h)^2\boldsymbol{\Sigma}_{\boldsymbol{G}_i^h}$,

$$P\big(\boldsymbol{a}_i^h\big|\boldsymbol{G}_i^h, \sigma_i^h\big) = \big|2\pi\big(\sigma_i^h\big)^2\boldsymbol{\Sigma}_{\boldsymbol{G}_i^h}\big|^{-\frac{1}{2}}\exp\left(-\frac{\boldsymbol{a}_{\boldsymbol{G}_i^h}^{\dagger}\boldsymbol{\Sigma}_{\boldsymbol{G}_i^h}^{-1}\boldsymbol{a}_{\boldsymbol{G}_i^h}}{2(\sigma_i^h)^2}\right) \tag{7}$$

where $|.|$ denotes the determinant of a matrix, the symbol $\dagger$ denotes matrix transposition, $\boldsymbol{\Sigma}_{\boldsymbol{G}_i^h} = \delta^{-2}\boldsymbol{D}_{\boldsymbol{G}_i^h}^{\dagger}\boldsymbol{D}_{\boldsymbol{G}_i^h}$ and $\boldsymbol{D}_{\boldsymbol{G}_i^h}$ is the $(\xi_i^h - \xi_i^{h-1}) \times (s_i^h + 1)$ matrix whose first column is a vector of 1's (for the constant in model (1)) and each $(j+1)^{th}$ column contains the observed values $(x_j(t))_{\xi_i^{h-1}-1\le t<\xi_i^h-1}$ for each factor gene j in $\boldsymbol{G}_i^h$. This so-called g-prior was also

used in Andrieu and Doucet (1999) and is motivated in Zellner (1986). Finally, the conjugate prior for the variance $(\sigma_i^h)^2$ is the inverse gamma distribution, $P((\sigma_i^h)^2) = \mathcal{IG}(\upsilon_0, \gamma_0)$. Following Lèbre (2007) and Lèbre et al. (2010), we set the hyper-hyperparameters for shape, $\upsilon_0 = 0.5$, and scale, $\gamma_0 = 0.05$, to fixed values that give a vague distribution. The terms λ and Λ can be interpreted as the expected number of changepoints and parents, respectively, and δ^2 is the expected signal-to-noise ratio. These hyperparameters are drawn from vague conjugate hyperpriors, which are in the (inverse) gamma distribution family:

$$P(\Lambda) = P(\lambda) = \mathcal{G}a(0.5, 1) = \Lambda^{-0.5} \frac{\exp(-\Lambda)}{\Gamma(0.5)} \tag{8}$$

and

$$P\left(\delta^2\right) = \mathcal{IG}(2, 0.2) = \delta^{-6} \frac{0.04 \exp(-\frac{0.2}{\delta^2})}{\Gamma(2)} \tag{9}$$

2.5 Posterior

Equation (1) implies that

$$\begin{aligned} &P\left(\boldsymbol{x}_i^h \middle| \boldsymbol{G}_i^h, \boldsymbol{a}_i^h, \sigma_i^h\right) \\ &\quad = \left(\sqrt{2\pi}\sigma_i^h\right)^{-\text{length}(\boldsymbol{x}_i^h)} \exp\left(-\frac{(\boldsymbol{x}_i^h - \boldsymbol{D}_{\boldsymbol{G}_i^h}\boldsymbol{a}_{\boldsymbol{G}_i^h})^\dagger (\boldsymbol{x}_i^h - \boldsymbol{D}_{\boldsymbol{G}_i^h}\boldsymbol{a}_{\boldsymbol{G}_i^h})}{2(\sigma_i^h)^2}\right) \end{aligned} \tag{10}$$

where $\text{length}(\boldsymbol{x}_i^h)$ is the length of the time series segment h. From Bayes' theorem, the posterior is given by the following equation, where all prior distributions have been defined above:

$$\begin{aligned} &P\left(\boldsymbol{k}, \boldsymbol{\xi}, \boldsymbol{G}, \boldsymbol{a}, \boldsymbol{\sigma^2}, \lambda, \Lambda, \delta^2 \middle| \boldsymbol{x}\right) \\ &\quad \propto P\left(\delta^2\right) P(\lambda) P(\Lambda) \prod_{i=1}^{p} P(k_i|\lambda) P(\boldsymbol{\xi}_i|k_i) \prod_{h=1}^{k_i} P\left(\boldsymbol{G}_i^h \middle| \Lambda\right) \\ &\qquad \times P\left(\left[\sigma_i^h\right]^2\right) P\left(\boldsymbol{a}_i^h \middle| \boldsymbol{G}_i^h, \left[\sigma_i^h\right]^2, \delta^2\right) P\left(\boldsymbol{x}_i^h \middle| \boldsymbol{G}_i^h, \boldsymbol{a}_i^h, \left[\sigma_i^h\right]^2\right) \end{aligned} \tag{11}$$

An attractive feature of the chosen model is that the integration over the parameters $\boldsymbol{a}$ and $\boldsymbol{\sigma^2}$ in the posterior distribution of Eq. (11) is analytically tractable:

$$\begin{aligned} &P\left(\boldsymbol{k}, \boldsymbol{\xi}, \boldsymbol{G}, \lambda, \Lambda, \delta^2 \middle| \boldsymbol{x}\right) \\ &\quad = \int\int P\left(\boldsymbol{k}, \boldsymbol{\xi}, \boldsymbol{G}, \boldsymbol{a}, \boldsymbol{\sigma^2}, \lambda, \Lambda, \delta^2 \middle| \boldsymbol{x}\right) d\boldsymbol{a} d\boldsymbol{\sigma^2} \\ &\quad \propto P\left(\delta^2\right) P(\lambda) P(\Lambda) \prod_{i=1}^{p} \int\int P\left(k_i, \boldsymbol{\xi}_i, \boldsymbol{G}_i, \boldsymbol{a}_i, \boldsymbol{\sigma}_i^{\boldsymbol{2}}, \boldsymbol{x}_i \middle| \lambda, \Lambda, \delta^2\right) d\boldsymbol{a}_i d\boldsymbol{\sigma}_i^{\boldsymbol{2}} \end{aligned} \tag{12}$$

For each gene i, the joint distribution for k_i, $\boldsymbol{\xi}_i$, $\boldsymbol{G}_i$, $\boldsymbol{a}_i$, $\boldsymbol{\sigma}_i^{\boldsymbol{2}}$, $\boldsymbol{x}_i$ conditional on hyperparameters λ, Λ, δ^2, is integrated over the parameters $\boldsymbol{a}_i$ (normal distribution) and $\boldsymbol{\sigma}_i^{\boldsymbol{2}}$ (inverse gamma distribution). Solving this integral (for details see Lèbre et al. 2010), the following

expression is obtained:

$$\int\int P\big(k_i, \boldsymbol{\xi}_i, \boldsymbol{G}_i, \boldsymbol{a}_i, \boldsymbol{\sigma}_i^2, \boldsymbol{x}_i \big| \lambda, \Lambda, \delta^2\big) d\boldsymbol{a}_i d\boldsymbol{\sigma}_i^2$$

$$= C_\lambda \lambda^{k_i} \frac{(N-2-k_i)!}{(N-2)!} \prod_{h=1}^{k_i+1} \left\{ \frac{(p-s_i^h)!}{p!} C_\Lambda \Lambda^{s_i^h} P\big(\boldsymbol{x}_i^h \big| \boldsymbol{G}_i^h, \delta^2\big) \right\} \tag{13}$$

where C_λ, C_Λ are the normalization constants required by the truncation of the Poisson distribution (2) and (4) and where

$$P\big(\boldsymbol{x}_i^h \big| \boldsymbol{G}_i^h, \delta^2\big) = \big(\delta^2+1\big)^{-\frac{s_i^h+1}{2}} \frac{(\frac{\gamma_0}{2})^{\upsilon_0/2}}{\Gamma(\frac{\upsilon_0}{2})} \Gamma\left(\frac{\upsilon_0 + \text{length}(\boldsymbol{x}_i^h)}{2}\right)$$

$$\times \left(\frac{\gamma_0 + (\boldsymbol{x}_i^h)^\dagger \boldsymbol{P}_i^h \boldsymbol{x}_i^h}{2}\right)^{-\frac{\upsilon_0+\text{length}(\boldsymbol{x}_i^h)}{2}} \tag{14}$$

where the matrices $\boldsymbol{P}_i^h$ and $\boldsymbol{M}_i^h$ are defined as follows, with $\boldsymbol{I}$ referring to the identity matrix of size length $(\boldsymbol{x}_i^h)$:

$$\boldsymbol{P}_i^h = \boldsymbol{I} - \boldsymbol{D}_{\boldsymbol{G}_i^h} \boldsymbol{M}_i^h \boldsymbol{D}_{\boldsymbol{G}_i^h}^\dagger, \tag{15}$$

$$\boldsymbol{M}_i^h = \frac{\delta^2}{\delta^2+1} \big(\boldsymbol{D}_{\boldsymbol{G}_i^h}^\dagger \boldsymbol{D}_{\boldsymbol{G}_i^h}\big)^{-1} \tag{16}$$

The number of changepoints $\boldsymbol{k}$ and their location, $\boldsymbol{\xi}$, the network structure $\boldsymbol{G}$ and the hyperparameters λ, Λ and δ^2 can be sampled from the posterior distribution $P(\boldsymbol{k}, \boldsymbol{\xi}, \boldsymbol{G}, \lambda, \Lambda, \delta^2 | \boldsymbol{x})$ with a reversible jump MCMC (Green 1995) scheme detailed in the next subsection.

2.6 RJMCMC scheme

Four different update moves are proposed: birth of a new changepoint (B); death (removal) of an existing changepoint (D); shift of a changepoint to a different time-point (S); and update of the network topology within the segments (N). These moves occur with probabilities b_{k_i} for B, d_{k_i} for D, u_{k_i} for S and v_{k_i} for N, depending only on the current number of changepoints k_i and satisfying $b_{k_i} + d_{k_i} + u_{k_i} + v_{k_i} = 1$. The changepoint birth and death moves represent changes from, respectively, k_i to $k_i + 1$ segments and k_i to $k_i - 1$ segments. In order to preserve the restriction on the number of changepoints, some probabilities are set to 0: $d_0 = u_0 = 0$ and $b_{\overline{k}} = 0$. Otherwise, following Green (1995), these probabilities are chosen as follows,

$$b_{k_i} = c \min\left\{1, \frac{P(k_i+1|\lambda)}{P(k_i|\lambda)}\right\}, \qquad d_{k_i+1} = c \min\left\{1, \frac{P(k_i|\lambda)}{P(k_i+1|\lambda)}\right\} \tag{17}$$

where $P(k_i|\lambda)$ is the prior distribution for the number of changepoints defined in Eq. (2) and the constant c is chosen to be smaller than $1/4$ so that network structure updates and changepoint position shifts are proposed more frequently than births and deaths of changepoints. This improves mixing and convergence with respect to changepoint positions and network structures within the different segments. Shifting of a changepoint is proposed with

probability $u_{k_i} = (1 - b_{k_i} - d_{k_i+1})/3$, and updating of the network structure within each segment is proposed with probability $v_{k_i} = 1 - (b_{k_i} + d_{k_i} + u_{k_i})$.

Following Green (1995), the RJMCMC acceptance probability of a changepoint birth is equal to $\min\{1, R\}$ where the acceptance ratio R reads as follows:

$$R = (\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio}) \times (\text{Jacobian}) \tag{18}$$

The product of the likelihood and the prior ratio is the posterior ratio which is derived from Eq. (12). The computation of the proposal ratio and the Jacobian depends on the choice for the various moves designed to sample the time-varying network distribution. We briefly describe below the chosen moves and their associated acceptance ratio. A complete description of the computation of the acceptance ratio for each move can be found in Lèbre et al. (2010).

Let $\boldsymbol{\xi}_i$ be the current changepoint vector containing k_i changepoints. For a changepoint birth move, a new changepoint position $\xi^\star$ is sampled uniformly from the available positions. The new changepoint is within an existing segment $h^\star$ of the target gene i, $\xi_i^{h^\star-1} < \xi^\star < \xi_i^{h^\star}$. Let us denote by $h_L^\star$ and $h_R^\star$ the segments to the left and to the right of the new changepoint respectively and by $\boldsymbol{x}_i^{h^\star} = (\boldsymbol{x}_i^{h_L^\star}, \boldsymbol{x}_i^{h_R^\star})$ the observed values for gene i in those segments. One of $h_L^\star$ and $h_R^\star$ is chosen with equal probability. That segment retains the current network topology $\boldsymbol{G}_i^{h^\star}$ of segment $h^\star$, and an entirely new topology is sampled from the prior defined in Eq. (6) for the other segment. Let us denote by $s^\star$ the number of edges of the new topology. The Jacobian is equal to 1 and the prior ratio is computed from the probability of choosing a new changepoint position and a new network structure for the new segment. Then the birth of the proposed changepoint is accepted with probability $A(\boldsymbol{\xi}_i^+|\boldsymbol{\xi}_i) = \min\{1, R(\boldsymbol{\xi}_i^+|\boldsymbol{\xi}_i)\}$, with

$$\begin{aligned} R\big(\boldsymbol{\xi}_i^+\big|\boldsymbol{\xi}_i\big) = {} & \frac{1}{(\delta^2+1)^{(s^\star+1)/2}} \frac{(\frac{\gamma_0}{2})^{\upsilon_0/2}}{\Gamma(\frac{\upsilon_0}{2})} \frac{\Gamma_{h_L^\star}\Gamma_{h_R^\star}}{\Gamma_{h^\star}} \left(\frac{\upsilon_0 + (\boldsymbol{x}_i^{h^\star})^\dagger \boldsymbol{P}_i^{h^\star} \boldsymbol{x}_i^{h^\star}}{2}\right)^{\frac{1}{2}(\upsilon_0+\xi_i^{h^\star}-\xi_i^{h^\star-1})} \\ & \times \left(\frac{\upsilon_0 + (\boldsymbol{x}_i^{h_L^\star})^\dagger \boldsymbol{P}_i^{h_L^\star} \boldsymbol{x}_i^{h_L^\star}}{2}\right)^{-\frac{1}{2}(\upsilon_0+\xi_i^{h_L^\star}-\xi_i^{h_L^\star-1})} \\ & \times \left(\frac{\upsilon_0 + (\boldsymbol{x}_i^{h_R^\star})^\dagger \boldsymbol{P}_i^{h_R^\star} \boldsymbol{x}_i^{h_R^\star}}{2}\right)^{-\frac{1}{2}(\upsilon_0+\xi_i^{h_R^\star}-\xi_i^{h_R^\star-1})} \end{aligned} \tag{19}$$

For details see Lèbre et al. (2010). Here $\boldsymbol{\xi}_i^+$ refers to the proposed changepoint vector after adding the new changepoint $\xi^\star$ to the current vector $\boldsymbol{\xi}_i$ and for all h in $\{1, \ldots, k_{i+1}\}$, $\Gamma_h = \Gamma(\frac{\upsilon_0+\xi_i^h-\xi_i^{h-1}}{2})$, and all other quantities are defined in Sect. 2.5.

For a changepoint death move, an existing changepoint in the current configuration is selected uniformly at random. The two segments adjacent to this changepoint are proposed to be merged into one segment, which will conserve the network structure of one of the two segments (selected with equal probability). Let us denote by $\boldsymbol{\xi}_i^-$ the proposed changepoint vector after removing the selected changepoint from the current vector $\boldsymbol{\xi}_i$. The acceptance ratio of the changepoint death move is equal to the inverse of the changepoint birth acceptance ratio $R(\boldsymbol{\xi}_i|\boldsymbol{\xi}_i^-)$ for proposing a change from $\boldsymbol{\xi}_i^-$ to $\boldsymbol{\xi}_i$, given in Eq. (19). Therefore the acceptance probability of a changepoint death move is,

$$A\big(\boldsymbol{\xi}_i^-\big|\boldsymbol{\xi}_i\big) = \min\big\{1, \big(R\big(\boldsymbol{\xi}_i\big|\boldsymbol{\xi}_i^-\big)\big)^{-1}\big\} \tag{20}$$

Proposed shifts in changepoint positions are accepted using a standard Metropolis-Hastings step (Hastings 1970) where a change is accepted with probability $\min\{1, R\}$ where $R =$ (posterior ratio) $\times$ (proposal ratio). The new changepoint vector $\tilde{\boldsymbol{\xi}}_i$ is obtained by replacing ξ_i^h with $\tilde{\xi}_i^h$ such that the absolute value $|\xi_i^h - \tilde{\xi}_i^h| = 1$. The posterior ratio is obtained from Eq. (12). Let us denote by $\mathcal{Q}(\tilde{\boldsymbol{\xi}}_i|\boldsymbol{\xi}_i)$ the probability of shifting changepoint ξ_i^h to $\tilde{\xi}_i^h$ in the current changepoint vector $\boldsymbol{\xi}^i$ (and reciprocally for $\mathcal{Q}(\boldsymbol{\xi}_i|\tilde{\boldsymbol{\xi}}_i)$), then the changepoint shift is accepted with probability $A(\tilde{\boldsymbol{\xi}}_i|\boldsymbol{\xi}_i) = \min\{1, R(\tilde{\boldsymbol{\xi}}_i|\boldsymbol{\xi}_i)\}$ where,

$$R(\tilde{\boldsymbol{\xi}}_i|\boldsymbol{\xi}_i) = \left(\frac{(\gamma_0 + (\tilde{\boldsymbol{x}}_i^h)^\dagger \tilde{\boldsymbol{P}}_i^h \tilde{\boldsymbol{x}}_i^h)^{(\upsilon_0+\tilde{\xi}_i^h-\xi_i^{h-1})}(\gamma_0 + (\tilde{\boldsymbol{x}}_i^{h+1})^\dagger \tilde{\boldsymbol{P}}_i^{h+1} \tilde{\boldsymbol{x}}_i^{h+1})^{(\upsilon_0+\xi_i^{h+1}-\tilde{\xi}_i^h)}}{(\gamma_0 + (\boldsymbol{x}_i^h)^\dagger \boldsymbol{P}_i^h \boldsymbol{x}_i^h)^{(\upsilon_0+\xi_i^h-\xi_i^{h-1})}(\gamma_0 + (\boldsymbol{x}_i^{h+1})^\dagger \boldsymbol{P}_i^{h+1} \boldsymbol{x}_i^{h+1})^{(\upsilon_0+\xi_i^{h+1}-\xi_i^h)}}\right)^{1/2}$$
$$\times \frac{\Gamma(\frac{\upsilon_0+\tilde{\xi}_i^h-\xi_i^{h-1}}{2})\Gamma(\frac{\upsilon_0+\xi_i^{h+1}-\tilde{\xi}_i^h}{2})}{\Gamma(\frac{\upsilon_0+\xi_i^h-\xi_i^{h-1}}{2})\Gamma(\frac{\upsilon_0+\xi_i^{h+1}-\xi_i^h}{2})} \frac{\mathcal{Q}(\boldsymbol{\xi}_i|\tilde{\boldsymbol{\xi}}_i)}{\mathcal{Q}(\tilde{\boldsymbol{\xi}}_i|\boldsymbol{\xi}_i)}, \tag{21}$$

where $\tilde{\boldsymbol{x}}_i^h$ and $\tilde{\boldsymbol{x}}_i^{h+1}$ refer to the expression levels for gene i observed in phase h and $h+1$ of the new changepoint vector $\tilde{\boldsymbol{\xi}}_i$, and $\tilde{\boldsymbol{P}}_i^h$ and $\tilde{\boldsymbol{P}}_i^{h+1}$ are the projection matrices built from $\tilde{\boldsymbol{x}}_i^h$ and $\tilde{\boldsymbol{x}}_i^{h+1}$ as defined in Eq. (15), and all other quantities are as defined in Sect. 2.5. See Lèbre et al. (2010) for the derivation of this equation.

Finally, network structure updates within segments invoke a second RJMCMC scheme, which was adapted from the model selection approach of Andrieu and Doucet (1999). When such a move is chosen, for each segment successively, we consider either the birth or death of an edge. For an edge birth move, a new edge is selected uniformly at random from the set of possible edges. For an edge death move, an edge to be removed is selected uniformly at random from the set of existing edges. The edge birth and death moves represent changes from s_i^h to $s_i^h + 1$ or $s_i^h - 1$ parents in the regression model. The probabilities of choosing these moves, $b_{s_i^h}$ and $d_{s_i^h}$ respectively, are defined as follows,

$$b_{s_i^h} = C_{s_i^h} \min\left\{1, \frac{P_{\bar{s}}(s_i^h + 1)}{P_{\bar{s}}(s_i^h)}\right\} \quad \text{and} \quad d_{s_i^h} = C_{s_i^h} \min\left\{1, \frac{P_{\bar{s}}(s_i^h - 1)}{P_{\bar{s}}(s_i^h)}\right\} \tag{22}$$

where $C_{s_i^h}$ is a normalization constant dependent on s_i^h, and set to ensure that $b_{s_i^h} + d_{s_i^h} = 1$. Additionally, we define $b_0 = 1$, $d_0 = 0$, $b_{\bar{s}} = 0$ and $d_{\bar{s}} = 1$. The acceptance ratio $R(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)$ for the new set of $\tilde{s}_i^h$ parents $\tilde{\boldsymbol{G}}_i^h$ (which corresponds to $\boldsymbol{G}_i^h$ with a parent added or removed) is computed according to Eq. (18). Using Eqs. (4) and (5), the edge birth prior ratio becomes

$$R_{prior} = \frac{P(\tilde{\boldsymbol{G}}_i^h|\tilde{s}_i^h)}{P(\boldsymbol{G}_i^h|s_i^h)} \frac{P(\tilde{s}_i^h|\Lambda)}{P(s_i^h|\Lambda)} \tag{23}$$

and the proposal ratio becomes

$$R_{proposal} = \frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)} \tag{24}$$

where $\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)$ is the proposal probability of parent set $\tilde{\boldsymbol{G}}_i^h$ given parent set $\boldsymbol{G}_i^h$, which is defined as follows:

$$\mathcal{Q}\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) = b_{|\boldsymbol{G}_i^h|}\delta\big(\big|\tilde{\boldsymbol{G}}_i^h\big|, \big|\boldsymbol{G}_i^h\big| + 1\big)\mathcal{Q}^+\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) + d_{|\boldsymbol{G}_i^h|}\delta\big(\big|\tilde{\boldsymbol{G}}_i^h\big|, \big|\boldsymbol{G}_i^h\big| - 1\big)\mathcal{Q}^-\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) \tag{25}$$

with $\delta(x, y)$ being the Kronecker delta function. $\mathcal{Q}^+(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h) = 1/(p - |\tilde{\boldsymbol{G}}_i^h|)$ is the proposal probability of an edge birth move, and $\mathcal{Q}^-(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h) = 1/|\tilde{\boldsymbol{G}}_i^h|$ is the proposal probability of an edge death move. The Jacobian equals 1. Then using Eq. (14) for the likelihood ratio, the Metropolis-Hastings acceptance ratio for an edge move becomes

$$R\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) = \frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)} \frac{P(\tilde{s}_i^h|\Lambda)}{P(s_i^h|\Lambda)} \frac{P(\tilde{\boldsymbol{G}}_i^h|\tilde{s}_i^h)}{P(\boldsymbol{G}_i^h|s_i^h)} \frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h, \delta^2)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h, \delta^2)} \tag{26}$$

Note that the prior ratio and the proposal ratio cancel out, and hence the edge move acceptance ratio is equal to the likelihood ratio, that is,

$$R\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) = \frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h, \delta^2)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h, \delta^2)} \tag{27}$$

Finally, the probability of accepting an edge move is,

$$A\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) = \min\big\{1, R\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big)\big\} \tag{28}$$

The sampling scheme for updating the hyperparameters δ^2, λ and Λ is described in Lèbre (2007) and Lèbre et al. (2010). Together the four moves B, D, S and N allow the generation of samples from probability distributions defined on unions of spaces of different dimensions for both the number of changepoints k_i and the number of parents s_i^h within each segment h for gene i.

3 Model improvement: information coupling between segments

Allowing the network structure to change between segments leads to a highly flexible model. However, this approach faces a conceptual and a practical problem. The *practical* problem is potential model over-flexibility. If subsequent changepoints are close together, network structures have to be inferred from short time series segments. This will almost inevitably lead to overfitting (in a maximum likelihood context) or inflated inference uncertainty (in a Bayesian context). The *conceptual* problem is the underlying assumption that structures associated with different segments are a priori independent. While this may be true in some circumstances (e.g. if a drug treatment leads to a drastic, rather than gradual, change), in most cases this assumption is not realistic. For instance, for the evolution of a gene regulatory network during embryogenesis, we would assume that the network evolves gradually and that networks associated with adjacent time intervals are a priori similar.

To address these problems, we propose four methods of information sharing among time series segments, as illustrated in Figs. 1 and 2. The first method is based on hard information coupling between the nodes, using the exponential distribution proposed in Werhli and Husmeier (2008). The second scheme uses the same exponential distribution, but replaces the hard by a soft information coupling scheme. The third and fourth scheme are also based on hard and soft information coupling, respectively, but use a binomial distribution with a conjugate beta prior.

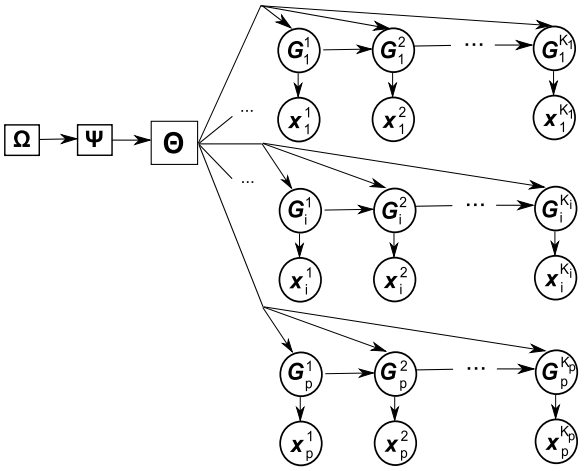


Fig. 1 Hierarchical Bayesian model for inter-segment and hard inter-node information coupling. Hard coupling among nodes i is achieved by a common hyperparameter $\boldsymbol{\Theta}$ regulating the strength of the coupling between structures associated with adjacent segments, $\boldsymbol{G}_i^h$ and $\boldsymbol{G}_i^{h+1}$. This corresponds to the models in Sect. 3.2, with $\boldsymbol{\Theta} = \{\beta\}$, $\boldsymbol{\Psi} = [0, 20]$, and no $\boldsymbol{\Omega}$, and Sect. 3.4, with $\boldsymbol{\Theta} = \{a, b\}$, $\boldsymbol{\Psi} = \{\alpha, \overline{\alpha}, \gamma, \overline{\gamma}\}$, and $\boldsymbol{\Omega} = \{1, 2, \ldots, 100\}$

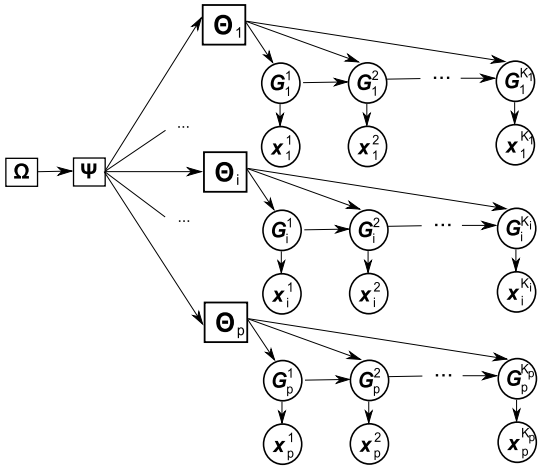


Fig. 2 Hierarchical Bayesian model for inter-segment and soft inter-node information coupling. Soft coupling among nodes i is achieved by node-specific hyperparameters $\boldsymbol{\Theta}_i$ regulating the strength of the coupling between structures associated with adjacent segments, $\boldsymbol{G}_i^h$ and $\boldsymbol{G}_i^{h+1}$, coupled via level-2 hyperparameters $\boldsymbol{\Psi}$. This corresponds to the model in Sect. 3.3, with $\boldsymbol{\Theta}_i = \{\beta_i\}$, $\boldsymbol{\Psi} = \kappa$, and $\boldsymbol{\Omega} = \lambda_\kappa = 10$, and Sect. 3.5, with $\boldsymbol{\Theta}_i = \{a_i, b_i\}$, $\boldsymbol{\Psi} = \{\alpha, \overline{\alpha}, \gamma, \overline{\gamma}\}$, and $\boldsymbol{\Omega} = \{1, 2, \ldots, 100\}$

3.1 Hard versus soft information coupling of nodes

As noted above, we propose to share information about the network structure among the different time series segments that result from the changepoint process. The strength of these couplings is governed by the hyperparameters associated with the information sharing prior. We represent these hyperparameters collectively by $\boldsymbol{\Theta}$. However, another level of coupling is possible, coupling genes (nodes in the network) rather than time series segments.

Recall from Sect. 2 that each node in the network is associated with a random variable $X_i(t)$ that represents the gene expression level of gene i at time t. Under the regression model in Eq. (1), the regulators for gene i are independent of the structure of the rest of

the network. Once we bring in information sharing, however, there is a set of hyperparameters that could conceivably be shared among different nodes; namely $\boldsymbol{\Theta}$. We address this by proposing two different ways of sharing $\boldsymbol{\Theta}$: Hard coupling, where the information sharing prior has the same hyperparameters $\boldsymbol{\Theta}$ for all nodes (with hyperprior having level-2 hyperparameters $\boldsymbol{\Psi}$); and soft coupling, where the information sharing prior has node-specific hyperparameters $\boldsymbol{\Theta}_i$, with common level-2 hyperparameters $\boldsymbol{\Psi}$. In both cases we have a prior on $\boldsymbol{\Psi}$ with level-3 hyperparameters $\boldsymbol{\Omega}$. See Figs. 1 and 2 for an illustration of hard versus soft information coupling of nodes.

In the following sub-sections, we will describe the different information sharing schemes in more detail.

3.2 Hard information coupling based on an exponential prior

Denote by $K_i := k_i + 1$ the total number of partitions in the time series associated with node i, and recall that each time series segment $\boldsymbol{x}_i^h$ is associated with a separate subnetwork $\boldsymbol{G}_i^h$, $1 \leq h \leq K_i$. We modify the prior from Eq. (6) by imposing a prior distribution $P(\boldsymbol{G}_i^h | \boldsymbol{G}_i^{h-1}, \beta)$ on the structures, and the joint probability distribution factorizes according to a Markovian dependence:

$$\begin{aligned} &P\big(\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^{K_i}, \boldsymbol{G}_i^1, \ldots, \boldsymbol{G}_i^{K_i}, \beta\big) \\ &\quad = P\big(\boldsymbol{x}_i^1 \big| \boldsymbol{G}_i^1\big) P\big(\boldsymbol{G}_i^1\big) P(\beta) \prod_{h=2}^{K_i} P\big(\boldsymbol{x}_i^h \big| \boldsymbol{G}_i^h\big) P\big(\boldsymbol{G}_i^h \big| \boldsymbol{G}_i^{h-1}, \beta\big) \end{aligned} \tag{29}$$

Similar to Werhli and Husmeier (2008) we define

$$P\big(\boldsymbol{G}_i^h \big| \boldsymbol{G}_i^{h-1}, \beta\big) = \frac{\exp(-\beta |\boldsymbol{G}_i^h - \boldsymbol{G}_i^{h-1}|)}{Z(\beta, \boldsymbol{G}_i^{h-1})} \tag{30}$$

for $h \geq 2$, where β is a hyperparameter that defines the strength of the coupling between $\boldsymbol{G}_i^h$ and $\boldsymbol{G}_i^{h-1}$, and $|.|$ denotes the Hamming distance. For $h = 1$, $P(\boldsymbol{G}_i^h)$ is given by (6). The denominator $Z(\beta, \boldsymbol{G}_i^{h-1})$ in (30) is a normalizing constant, also known as the partition function: $Z(\beta, \boldsymbol{G}_i^{h-1}) = \sum_{\boldsymbol{G}_i^h \in \mathbb{G}} e^{-\beta |\boldsymbol{G}_i^h - \boldsymbol{G}_i^{h-1}|}$ where $\mathbb{G}$ is the set of all valid subnetwork structures. If we ignore any fan-in restriction that might have been imposed a priori (via $\overline{s}$ in Eq. (4)), then the expression for the partition function can be simplified: $Z(\beta, \boldsymbol{G}_i^{h-1}) \approx \prod_{j=1}^{p} Z_j(\beta, e_{ij}^{h-1})$, where e_{ij}^h is a binary variable indicating the presence or absence of a directed edge from node j to node i in time series segment h, and $Z_j(\beta, e_{ij}^{h-1}) = \sum_{e_{ij}^h=0}^{1} e^{-\beta |e_{ij}^h - e_{ij}^{h-1}|} = 1 + e^{-\beta}$. Note that this expression no longer depends on $\boldsymbol{G}_i^{h-1}$, and hence

$$Z\big(\beta, \boldsymbol{G}_i^{h-1}\big) = Z(\beta) = \big(1 + e^{-\beta}\big)^p \tag{31}$$

Inserting this expression into (30) gives:

$$P\big(\boldsymbol{G}_i^h \big| \boldsymbol{G}_i^{h-1}, \beta\big) = \frac{\exp(-\beta |\boldsymbol{G}_i^h - \boldsymbol{G}_i^{h-1}|)}{(1 + e^{-\beta})^p} \tag{32}$$

It is straightforward to integrate the proposed model into the RJMCMC scheme of Lèbre (2007) and Lèbre et al. (2010), which we have summarized in Sect. 2.6. When proposing a

new network structure $\boldsymbol{G}_i^h \to \tilde{\boldsymbol{G}}_i^h$ for segment h, the prior probability ratio in Eq. (23) has to be replaced by $\frac{P(\boldsymbol{G}_i^{h+1}|\tilde{\boldsymbol{G}}_i^h,\beta)P(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^{h-1},\beta)}{P(\boldsymbol{G}_i^{h+1}|\boldsymbol{G}_i^h,\beta)P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1},\beta)}$, leading to the acceptance probability

$$A\big(\tilde{\boldsymbol{G}}_i^h\big|\boldsymbol{G}_i^h\big) = \min\left\{\frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h)P(\boldsymbol{G}_i^{h+1}|\tilde{\boldsymbol{G}}_i^h,\beta)P(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^{h-1},\beta)\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)P(\boldsymbol{G}_i^{h+1}|\boldsymbol{G}_i^h,\beta)P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1},\beta)\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)},1\right\} \tag{33}$$

This equation is equivalent to Eq. (28), with the prior probabilities in Eq. (23) replaced by those in Eq. (32). Note that $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)$ is short for $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)$ which is defined in Eq. (14) and the proposal ratio $\frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)}$ is defined in Eqs. (24) and (25). An additional MCMC step is introduced for sampling the hyperparameter β from the posterior distribution. For a proposal move $\beta \to \tilde{\beta}$ with symmetric proposal probability $\mathcal{Q}(\tilde{\beta}|\beta) = \mathcal{Q}(\beta|\tilde{\beta})$ we get the following acceptance probability:

$$A(\tilde{\beta}|\beta) = \min\left\{\frac{P(\tilde{\beta})}{P(\beta)}\prod_{i=1}^{p}\prod_{h=2}^{K_i}\frac{\exp(-\tilde{\beta}|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}{\exp(-\beta|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}\frac{(1+e^{-\beta})^p}{(1+e^{-\tilde{\beta}})^p},1\right\} \tag{34}$$

where in our study the hyperprior $P(\beta)$ was chosen as the uniform distribution on the interval $[0, 20]$.

3.3 Soft information coupling based on an exponential prior

We modify the model defined in (29) by making the hyperparameter β, which defines the prior coupling strength between structures associated with adjacent segments, node-dependent: $\beta \to \beta_i$, and

$$P\big(\boldsymbol{x}_i^1,\ldots,\boldsymbol{x}_i^{K_i},\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^K,\beta_i\big) = P\big(\boldsymbol{x}_i^1\big|\boldsymbol{G}_i^1\big)P\big(\boldsymbol{G}_i^1\big)\prod_{h=2}^{K_i}P\big(\boldsymbol{x}_i^h\big|\boldsymbol{G}_i^h\big)P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1},\beta_i\big)P(\beta_i) \tag{35}$$

with

$$P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1},\beta_i\big) = \frac{\exp(-\beta_i|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}{Z(\beta_i,\boldsymbol{G}_i^{h-1})} = \frac{\exp(-\beta_i|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}{(1+e^{-\beta_i})^p} \tag{36}$$

where by analogy with the previous section, $Z(\beta_i,\boldsymbol{G}_i^{h-1}) \approx (1+e^{-\beta_i})^p$. To introduce soft information coupling between the subnetworks, we choose a hierarchical structure for the prior distribution on the hyperparameters β_i. At the first level, the hyperparameters are given a common gamma prior:

$$P(\beta_i) = P(\beta_i|\kappa,\rho) = \beta_i^{\kappa-1}\frac{\exp(-\beta_i/\rho)}{\rho^\kappa\Gamma(\kappa)} \tag{37}$$

with shape parameter $\kappa > 0$ and scale parameter $\rho > 0$. Recall that the gamma distribution has mean $\mu = \kappa\rho$ and variance $\sigma^2 = \kappa\rho^2$. We elect to set the scale parameter $\rho = 0.1$ fixed. The shape parameter κ is given a vague exponential prior:

$$P(\kappa|\lambda_\kappa) = \lambda_\kappa\exp(-\kappa/\lambda_\kappa) \tag{38}$$

with $\lambda_\kappa = 10$ to reflect our prior ignorance. This choice of prior has the following motivation. The coupling strength between the substructures is defined by the coefficient of variation

$\sigma/\mu = 1/\sqrt{\kappa}$, with smaller coefficients corresponding to stronger coupling strengths, and a zero coefficient ($\kappa \to \infty$) reducing to the hard coupling scheme discussed in the previous section. By inferring the shape parameter κ from the data, starting from a vague yet proper prior distribution, we determine if the coupling strength should be strong or weak.

It is straightforward to adapt the RJMCMC scheme of the previous section. When proposing a new network structure $\boldsymbol{G}_i^h \to \tilde{\boldsymbol{G}}_i^h$ for segment h, the prior probability ratio in Eq. (23) has to be replaced by the ratio $\frac{P(\boldsymbol{G}_i^{h+1}|\tilde{\boldsymbol{G}}_i^h,\beta_i)P(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^{h-1},\beta_i)}{P(\boldsymbol{G}_i^{h+1}|\boldsymbol{G}_i^h,\beta_i)P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1},\beta_i)}$, leading to the equivalent of the acceptance probability in Eq. (28):

$$A\big(\tilde{\boldsymbol{G}}_i^h\big|\boldsymbol{G}_i^h\big) = \min\left\{\frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h)P(\boldsymbol{G}_i^{h+1}|\tilde{\boldsymbol{G}}_i^h,\beta_i)P(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^{h-1},\beta_i)\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)P(\boldsymbol{G}_i^{h+1}|\boldsymbol{G}_i^h,\beta_i)P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1},\beta_i)\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)},1\right\} \tag{39}$$

Note that $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)$ is short for $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)$ which is defined in Eq. (14) and the proposal ratio $\frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)}$ defined in Eqs. (24) and (25). When proposing new hyperparameters $\tilde{\beta}_i$ from a symmetric proposal distribution $\mathcal{Q}(\tilde{\beta}_i|\beta_i) = \mathcal{Q}(\beta_i|\tilde{\beta}_i)$ we get the following acceptance probability:

$$A(\tilde{\beta}_i|\beta_i) = \min\left\{\frac{P(\tilde{\beta}_i|\rho,\kappa)}{P(\beta_i|\rho,\kappa)}\prod_{h=2}^{K_i}\frac{\exp(-\tilde{\beta}_i|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}{\exp(-\beta_i|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}\left(\frac{1+e^{-\beta_i}}{1+e^{-\tilde{\beta}_i}}\right)^p,1\right\} \tag{40}$$

An additional sampling step is needed for the shape parameter κ of the level-2 hyperprior. Drawing a new shape parameter $\tilde{\kappa}$ from a symmetric proposal distribution $\mathcal{Q}(\tilde{\kappa}|\kappa)$, the acceptance probability is given by

$$A(\tilde{\kappa}|\kappa) = \min\left\{\frac{\exp(-\tilde{\kappa}/\lambda_\kappa)}{\exp(-\kappa/\lambda_\kappa)}\prod_{i=1}^{p}\frac{P(\beta_i|\tilde{\kappa},\rho)}{P(\beta_i|\kappa,\rho)},1\right\} \tag{41}$$

3.4 Hard information coupling based on a binomial prior

An alternative way of information sharing among segments and nodes is by using a binomial prior:

$$P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1},a,b\big) = a^{N_1^1[h,i]}(1-a)^{N_1^0[h,i]}b^{N_0^0[h,i]}(1-b)^{N_0^1[h,i]} \tag{42}$$

where we have defined the following sufficient statistics: $N_1^1[h,i]$ is the number of edges in $\boldsymbol{G}_i^{h-1}$ that are matched by an edge in $\boldsymbol{G}_i^h$, $N_1^0[h,i]$ is the number of edges in $\boldsymbol{G}_i^{h-1}$ for which there is no edge in $\boldsymbol{G}_i^h$, $N_0^1[h,i]$ is the number of edges in $\boldsymbol{G}_i^h$ for which there is no edge in $\boldsymbol{G}_i^{h-1}$, and $N_0^0[h,i]$ is the number of coinciding non-edges in $\boldsymbol{G}_i^{h-1}$ and $\boldsymbol{G}_i^h$. Since the hyperparameters are shared, the joint distribution can be expressed as:

$$P\big(\{\boldsymbol{G}_i^h\}|a,b\big) = \prod_{i=1}^{p}P\big(\boldsymbol{G}_i^1\big)\prod_{h=2}^{K_i}P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1},a,b\big) = a^{N_1^1}(1-a)^{N_1^0}b^{N_0^0}(1-b)^{N_0^1}\prod_{i=1}^{p}P\big(\boldsymbol{G}_i^1\big) \tag{43}$$

where we have defined $N_k^l = \sum_{i=1}^{p}\sum_{h=2}^{K_i}N_k^l[h,i]$, and the right-hand side follows from Eq. (42). The conjugate prior for the hyperparameters a,b is a beta distribution,

$$P(a,b|\alpha,\overline{\alpha},\gamma,\overline{\gamma}) \propto a^{(\alpha-1)}(1-a)^{(\overline{\alpha}-1)}b^{(\gamma-1)}(1-b)^{(\overline{\gamma}-1)} \tag{44}$$

which using Bayes' rule leads to the (beta) posterior distribution:

$$P\big(a,b\big|\alpha,\overline{\alpha},\gamma,\overline{\gamma},\big\{\boldsymbol{G}_i^h\big\}\big)\propto a^{(\alpha+N_1^1-1)}(1-a)^{(\overline{\alpha}+N_1^0-1)}b^{(\gamma+N_0^0-1)}(1-b)^{(\overline{\gamma}+N_0^1-1)} \tag{45}$$

This allows the hyperparameters to be integrated out in closed form:

$$\begin{aligned}&P\big(\big\{\boldsymbol{G}_i^h\big\}\big|\alpha,\overline{\alpha},\gamma,\overline{\gamma}\big)\\&\quad=\int\int P\big(\big\{\boldsymbol{G}_i^h\big\}\big|a,b\big)P(a,b|\alpha,\overline{\alpha},\gamma,\overline{\gamma})da\,db\\&\quad\propto\frac{\Gamma(\alpha+\overline{\alpha})}{\Gamma(\alpha)\Gamma(\overline{\alpha})}\frac{\Gamma(N_1^1+\alpha)\Gamma(N_1^0+\overline{\alpha})}{\Gamma(N_1^1+\alpha+N_1^0+\overline{\alpha})}\frac{\Gamma(\gamma+\overline{\gamma})}{\Gamma(\gamma)\Gamma(\overline{\gamma})}\frac{\Gamma(N_0^0+\gamma)\Gamma(N_0^1+\overline{\gamma})}{\Gamma(N_0^0+\gamma+N_0^1+\overline{\gamma})}\end{aligned} \tag{46}$$

The level-2 hyperparameters $\alpha,\overline{\alpha},\gamma,\overline{\gamma}$, which can be interpreted as fictitious prior observations due to the conjugacy of the prior, are given a discrete uniform hyperprior over $\{1,2,\ldots,100\}$. The MCMC scheme of Sect. 2.6 has to be modified as follows. When proposing a new network structure for node i and segment h, $\boldsymbol{G}_i^h\rightarrow\tilde{\boldsymbol{G}}_i^h$, the structures $\boldsymbol{G}_i^h$ and $\tilde{\boldsymbol{G}}_i^h$ enter the prior probability ratio in Eq. (23) via the expression $P(\{\boldsymbol{G}_i^h\}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})$. The prior probability ratio becomes $\frac{P(\{\boldsymbol{G}_i^1,\ldots,\tilde{\boldsymbol{G}}_i^h,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\alpha,\overline{\alpha},\gamma,\overline{\gamma})}{P(\{\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^h,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\alpha,\overline{\alpha},\gamma,\overline{\gamma})}$, leading to the acceptance probability

$$A\big(\tilde{\boldsymbol{G}}_i^h\big|\boldsymbol{G}_i^h\big)=\min\left\{\frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h)P(\{\boldsymbol{G}_i^1,\ldots,\tilde{\boldsymbol{G}}_i^h,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\alpha,\overline{\alpha},\gamma,\overline{\gamma})\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)P(\{\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^h,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\alpha,\overline{\alpha},\gamma,\overline{\gamma})\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)},1\right\} \tag{47}$$

This equation is equivalent to Eq. (28), with the prior probabilities in Eq. (23) replaced by those in Eq. (46). Note that $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)$ is short for $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)$ which is defined in Eq. (14) and the proposal ratio $\frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)}$ defined in Eqs. (24) and (25). From Fig. 1, it becomes clear that as a consequence of integrating out the hyperparameters, all network structures become interdependent, and information about the structures is contained in the sufficient statistics N_1^1,N_1^0,N_0^1,N_0^0. A new proposal move for the level-2 hyperparameters is added to the existing RJMCMC scheme of Sect. 2.6. New values for the level-2 hyperparameters α are proposed from a uniform distribution over the support of $P(\alpha)$. For a move $\alpha\rightarrow\tilde{\alpha}$, the acceptance probability is:

$$A(\tilde{\alpha}|\alpha)=\min\left\{\frac{P(\{\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\tilde{\alpha},\overline{\alpha},\gamma,\overline{\gamma})}{P(\{\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\alpha,\overline{\alpha},\gamma,\overline{\gamma})},1\right\} \tag{48}$$

and similarly for $\overline{\alpha}$, γ and $\overline{\gamma}$.

3.5 Soft information coupling based on a binomial prior

We can relax the information sharing scheme from a hard to a soft coupling by introducing node-specific hyperparameters a_i,b_i that are softly coupled via a common level-2 hyperprior, $P(a_i,b_i|\alpha,\overline{\alpha},\gamma,\overline{\gamma})\propto a_i^{(\alpha-1)}(1-a_i)^{(\overline{\alpha}-1)}b_i^{(\gamma-1)}(1-b_i)^{(\overline{\gamma}-1)}$ as illustrated in Fig. 2:

$$P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1},a_i,b_i\big)=(a_i)^{N_1^1[h,i]}(1-a_i)^{N_1^0[h,i]}(b_i)^{N_0^0[h,i]}(1-b_i)^{N_0^1[h,i]} \tag{49}$$

This leads to a straightforward modification of Eq. (43)—replacing a, b by a_i, b_i—from which we get as an equivalent to (46), using the definition $N_k^l[i] = \sum_{h=2}^{K_i} N_k^l[h,i]$:

$$P\left(\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^{K_i}\middle|\alpha,\overline{\alpha},\gamma,\overline{\gamma}\right) \propto \frac{\Gamma(\alpha+\overline{\alpha})}{\Gamma(\alpha)\Gamma(\overline{\alpha})}\frac{\Gamma(N_1^1[i]+\alpha)\Gamma(N_1^0[i]+\overline{\alpha})}{\Gamma(N_1^1[i]+\alpha+N_1^0[i]+\overline{\alpha})} \times \frac{\Gamma(\gamma+\overline{\gamma})}{\Gamma(\gamma)\Gamma(\overline{\gamma})}\frac{\Gamma(N_0^0[i]+\gamma)\Gamma(N_0^1[i]+\overline{\gamma})}{\Gamma(N_0^0[i]+\gamma+N_0^1[i]+\overline{\gamma})} \tag{50}$$

As in Sect. 3.4, we extend the RJMCMC scheme from Sect. 2.6 so that when proposing a new network structure, $\boldsymbol{G}_i^h \to \tilde{\boldsymbol{G}}_i^h$, the prior probability ratio in Eq. (23) has to be replaced by: $\frac{P(\boldsymbol{G}_i^1,\ldots,\tilde{\boldsymbol{G}}_i^h,\ldots,\boldsymbol{G}_i^{K_i}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})}{P(\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^h,\ldots,\boldsymbol{G}_i^{K_i}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})}$, leading to the equivalent of the acceptance probability in Eq. (28):

$$A\left(\tilde{\boldsymbol{G}}_i^h\middle|\boldsymbol{G}_i^h\right) = \min\left\{\frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h)P(\boldsymbol{G}_i^1,\ldots,\tilde{\boldsymbol{G}}_i^h,\ldots,\boldsymbol{G}_i^{K_i}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)P(\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^h,\ldots,\boldsymbol{G}_i^{K_i}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)},1\right\} \tag{51}$$

Note that $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)$ is short for $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)$ which is defined in Eq. (14) and the proposal ratio $\frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)}$ defined in Eqs. (24) and (25). In addition, we have to add a new level-2 hyperparameter update move: when proposing a level-2 hyperparameter $\alpha \to \tilde{\alpha}$, where the prior and proposal probabilities are the same as in Sect. 3.4, the acceptance probability becomes:

$$A(\tilde{\alpha}|\alpha) = \min\left\{\prod_{i=1}^{p}\frac{P(\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^{K_i}|\tilde{\alpha},\overline{\alpha},\gamma,\overline{\gamma})}{P(\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^{K_i}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})},1\right\} \tag{52}$$

and similarly for $\overline{\alpha}$, γ and $\overline{\gamma}$.

3.6 Improved MCMC scheme

The various information sharing priors that we have introduced in the previous Sects 3.2, 3.3, 3.4, 3.5 share the characteristic that they encourage the networks of all segments to be similar to each other.[2] When applying the MCMC scheme from Lèbre et al. (2010), summarized in Sect. 2.6, adapted to our prior as discussed above, this can lead to the following curious effect. On simulated data where the network structure is the same for all segments we found that the network reconstruction accuracy deteriorated when we increased the coupling strength between the structures. The results will be presented below, in Sect. 5 and Fig. 4. These findings appear counter-intuitive, given that increasing the coupling strength brings the prior more in line with the truth (the perfect prior would have infinitely strong coupling). However, it is easily seen that increasing the coupling strength adversely affects the mixing of the Markov chains. Consider a set of identical network structures which, at an initial stage of the MCMC simulations, are all poor at explaining the data. We now visit a

[2]Note that the binomial information sharing prior (Sects. 3.4 and 3.5) can in principle encourage either similarity or dissimilarity depending on the hyperparameters a and b. As discussed in Sect. 5, we had originally envisaged setting the level-2 hyperparameters $\overline{\alpha}$ and $\overline{\gamma}$ equal to 1 to enforce similarity, but Fig. 8 demonstrates that this constraint is too restrictive.

segment and propose a modification of the network structure associated with it. This modification introduces a mismatch between the structures and is, hence, discouraged by the prior. For strong coupling this discouragement might outweigh the gain in the likelihood that would result from a better structure. The structures thus remain identical, which in turn will tend to increase the coupling strength. The MCMC simulation thus gets trapped in a suboptimal state of the configuration space (local optimum).

To deal with this problem, we have implemented an alternative MCMC scheme where changes are applied to multiple segments. The new moves will propose changes to the network structure in more than one segment, and we will hence refer to them as multi-segment moves. Note that the moves for proposing new changepoint configurations are unaffected by these modifications. The multi-segment moves are presented as target-node specific (i.e. they presuppose a choice of target node i). However, they can be generalized for inference over the whole network by simply picking a target node at random. Given a node, the proposal move consists of two steps: (1) Pick one of p possible parents for the target node i. (2) For each segment h of the K_i segments, flip the edge status (changing an edge to a non-edge or vice-versa) between the parent node and the target node with probability q. In our simulations, we set $q = \frac{1}{2}$ so that flipping the edge status and conserving it are equally likely outcomes. It is straightforward to adapt this parameter during the burn-in phase. This means that the probability of proposing a new set of structures $\tilde{\boldsymbol{G}}_i$ given the set of network structures $\boldsymbol{G}_i$ using the multi-segment move is:

$$\mathcal{Q}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i) = \frac{1}{p2^{K_i}} \tag{53}$$

where $\boldsymbol{G}_i = \{\boldsymbol{G}_i^h\}_{1\leq h\leq K_i}$ as before.

We now derive the acceptance ratio for multi-segment moves. We define $R_{prior}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)$ to be the ratio of the prior probabilities of the original set $\boldsymbol{G}_i$ and the proposed set $\tilde{\boldsymbol{G}}_i$. Let $R_{likelihood}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h) = \frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h,\delta^2)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)}$ be the likelihood ratio of the original and proposed network structures for segment h and target node i, where the likelihood $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)$ is defined in Eq. (14) of Sect. 2.6. Note that the changes introduced by multi-segment moves are equivalent to a sequence of add and remove edge moves applied to individual segments, so that this ratio remains unchanged. Then the acceptance ratio for multi-segment moves can be expressed as:

$$R(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i) = R_{prior}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)R_{proposal}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)\prod_{h=1}^{K_i} R_{likelihood}\big(\tilde{\boldsymbol{G}}_i^h\big|\boldsymbol{G}_i^h\big) \tag{54}$$

where $R_{prior}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i) = \frac{P(\tilde{\boldsymbol{G}}_i)}{P(\boldsymbol{G}_i)}$. The form of $P(\boldsymbol{G}_i)$ depends on our choice of prior. If segments are independent, then $P(\boldsymbol{G}_i) = \prod_{h=1}^{K_i} P(\boldsymbol{G}_i^h)$, where $P(\boldsymbol{G}_i^h)$ is the prior from Eq. (6), with a Poisson distribution on the number of parents. If we want to use information sharing between segments, then the prior for segment h depends on segment $h-1$, so that $P(\boldsymbol{G}_i) = P(\boldsymbol{G}_i^1)\prod_{h=2}^{K_i} P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1})$, where $P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1})$ could be any of the information sharing priors introduced in Sect. 3. Finally, $R_{proposal}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)$ is the Hastings ratio:

$$R_{proposal}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i) = \frac{\mathcal{Q}(\boldsymbol{G}_i|\tilde{\boldsymbol{G}}_i)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)} \tag{55}$$

where $\mathcal{Q}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)$ is defined in Eq. (53). Since the proposal probability $\mathcal{Q}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)$ is independent of the set of network structures $\boldsymbol{G}_i$, the multi-segment moves are symmetric, and we obtain that $R_{proposal}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i) = 1$.

We have explored an alternative proposal scheme consisting of two moves: (1) a move proposing network structures where an edge has been set identical in all segments, and (2) the move described above, which corresponds to a random perturbation of an edge. However, we found that including the first kind of proposal move adversely affected mixing and convergence in simulations where the true network structure presented differences among segments. These network structures are less likely to be proposed when both moves are included. Details can be found in Dondelinger (2012).

4 Implementation and simulations

We have implemented our model in R, based on code from Lèbre (2007) and Lèbre et al. (2010). The network structure, the changepoints and the hyperparameters are sampled from the posterior distribution using RJMCMC as described in Sects. 2.6 and 3.6. We ran the MCMC chains until we were satisfied that convergence was reached. Then we sampled 1000 network and changepoint configurations in intervals of 200 RJMCMC steps. By marginalization and under the assumption of convergence, this represents a sample from the posterior distribution in Eq. (12). By further marginalization, we get the posterior probabilities of all gene regulatory interactions, which defines a ranking of the interactions in terms of posterior confidence. We use the potential scale reduction factor (PSRF) (Gelman and Rubin 1992), computed from the within-chain and between-chain variances of marginal edge posterior probabilities, as a convergence diagnostic. The usual threshold for sufficient convergence lies at PSRF ≤ 1.1. In our simulations, we extended the burn-in phase until a value of PSRF ≤ 1.05 was reached.

For the study on simulated data, and the synthetic biology data, the true interaction network is known. Therefore, varying the threshold on this ranking allows us to construct the Receiver Operating Characteristic (ROC) curve (plotting the sensitivity or recall[3] against the complementary specificity[4]) and the precision-recall (PR) curve (plotting the precision[5] against the recall), and to assess the network reconstruction accuracy in terms of the areas under these graphs (AUROC and AUPRC, respectively); see Davis and Goadrich (2006). These two measures are widely used in the systems biology literature to quantify the overall network reconstruction accuracy (Prill et al. 2010), with larger values indicating a better prediction performance overall.

5 Evaluation on simulated data

5.1 Comparative evaluation of network reconstruction and hyperparameter inference

The purpose of the simulation study is two-fold. Firstly, we want to carry out a comparative evaluation of the proposed Bayesian regularization schemes for a controlled scenario in

[3]The *sensitivity* or *recall* denotes the fraction of true interaction that have been recovered.

[4]The *specificity* denotes the fraction of spurious interactions that have been successfully avoided.

[5]The *precision* is the fraction of predicted interactions that are correct.

which the true network structure is known. Secondly, we want to assess the Bayesian inference scheme and test the viability of the proposed MCMC samplers. To focus on the task of network reconstruction, we keep the changepoints fixed at their true values. The inference of the changepoints will be investigated later, on the real gene expression time series (see Fig. 12).

The simulation set-up we chose was as follows. We randomly generated 10 networks with 10 nodes each. A Poisson distribution with mean $\lambda_{parents} = 3$ was used to determine the number of parents for each node. We simulated changes in the network structure by producing 4 different network segments, where a Poisson distribution with mean $\lambda_{changes} \in \{0.25, 0.5, 1\}$ was used to determine the number of changes per node. The changes were then applied uniformly at random to edges and non-edges in the previous segment. For each segment h, we generated a time series of length 15 using a linear regression model:

$$\boldsymbol{x}(t) = \boldsymbol{W}^h \boldsymbol{x}(t-1) + \boldsymbol{\epsilon} \tag{56}$$

where $\boldsymbol{x}(t)$ is the 10×1 vector of observations at time t and $\boldsymbol{W}^h = \{w_{ij}^h\}$ is the 10×10 matrix of segment-specific regression weights for each edge. We chose the regression weights such that $w_{ij}^h = 0$ if there is no edge between node i and node j in the network structure for segment h, and $w_{ij}^h \sim N(0, 1)$ otherwise. We added Gaussian observation noise $\epsilon_i \sim N(0, 1)$ independently for each observation of node i.

First, we consider the scenario of homogeneous time series in which the regulatory network structure does not change (although the regression coefficients associated with each edge may change between segments). This is the situation in which the proposed Bayesian regularization scheme should achieve the strongest boost in the network reconstruction accuracy. We indeed found this conjecture confirmed in our simulations, as demonstrated in Fig. 3 (0 % changes). We would also assume that high values of the hyperparameter β should lead to the best network reconstruction accuracy, as this corresponds to the tightest tying between adjacent structures. However, repeating the MCMC simulations initially did not confirm this conjecture; see Figs. 4(c) and 4(d). As discussed in Sect. 3.6, the observed mismatch was a consequence of poor mixing and convergence for large hyperparameter values, which is endemic to the naive extension of the MCMC sampler from Lèbre et al. (2010). Repeating the simulations with the novel MCMC scheme proposed in Sect. 3.6 leads to the graphs of Figs. 4(a) and 4(b). Here, the network reconstruction accuracy no longer deteriorates with increasing hyperparameters, indicating that the mixing and convergence problems have been averted.

Another question we investigated is whether the sampled values of the hyperparameters concur with those that optimize the network reconstruction accuracy. While the hyperparameter β of the exponential prior does indeed tend to higher values, the situation is different for the hyperparameters a and b of the binomial prior. The top panels in Fig. 5 show the network reconstruction accuracy in terms of AUROC and AUPRC scores for several fixed values of the hyperparameters a and b. As expected, the peak performance is reached for the highest values, as no mismatch between the structures implies that tight coupling is consistent with the data. The centre panels of Fig. 5 show the posterior distribution of the hyperparameters that was obtained with the conventional MCMC proposal scheme adapted from Lèbre et al. (2010) and described in Sect. 2.6. There is an obvious mismatch between the high-posterior probability region and the region of hyperparameters that optimize the network reconstruction. This provides more evidence that the sampler adapted for segment coupling from Lèbre et al. (2010) suffers from mixing and convergence problems. The bottom panels of Fig. 5 show the marginal posterior distributions of the hyperparameters inferred in the

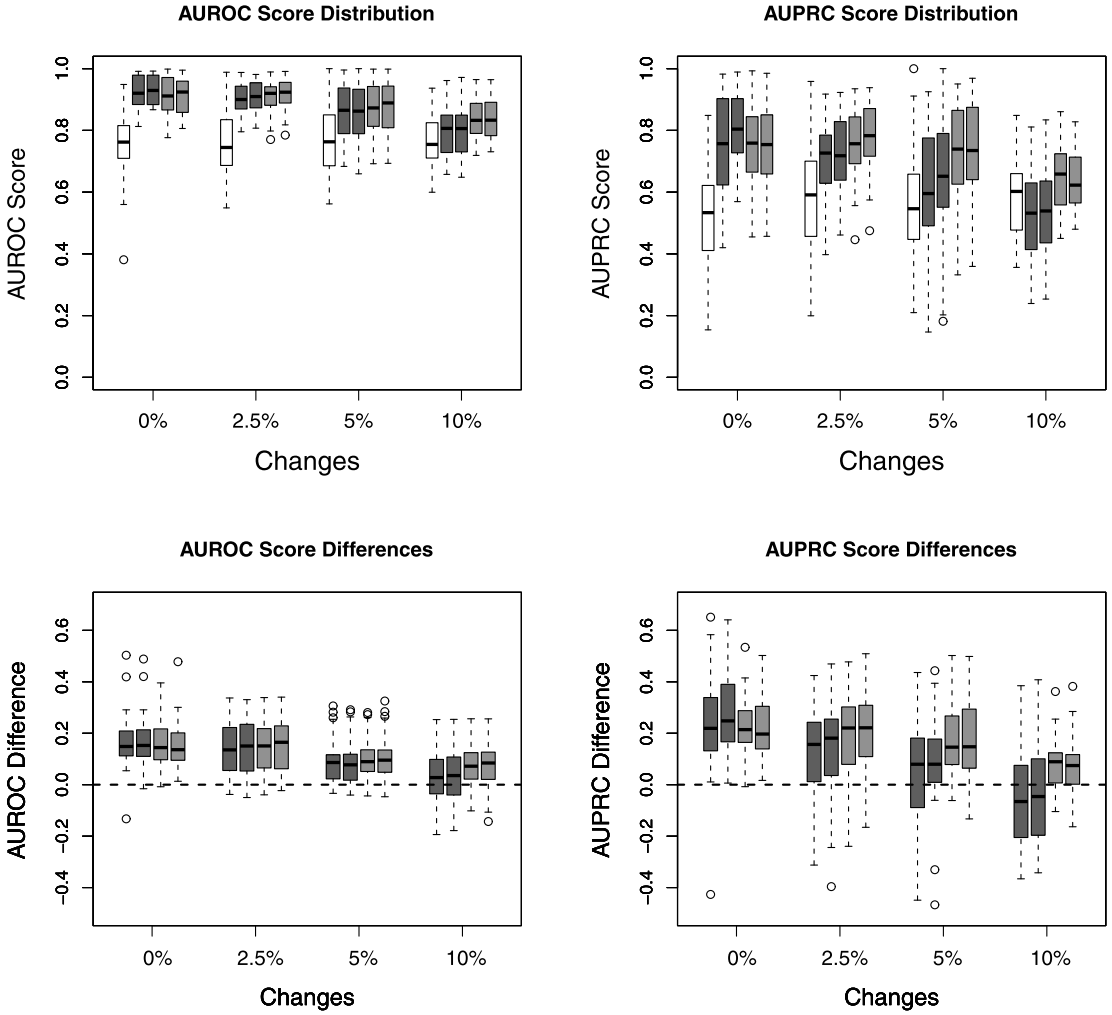


Fig. 3 Evaluation of AUROC and AUPRC network reconstruction scores for the five methods, TVDBN-0 (*white*), TVDBN-Exp-hard (*dark grey, left*), TVDBN-Exp-soft (*dark grey, right*), TVDBN-Bino-hard (*light grey, left*), TVDBN-Bino-soft (*light grey, right*). *Top row*: The *boxplots* show the distributions of the reconstruction scores. *Bottom row*: The *boxplots* show the difference of the AUROC and AUPRC reconstruction scores to TVDBN-0; larger differences indicate better performance with information sharing. All simulations were repeated for 10 independent data sets with 4 network segments each. Structure changes were applied to the segments sequentially, changing between 0–10 % of the edges with each new segment. A paired t-test shows that for 0 % changes, the difference to TVDBN-0 was significant for all methods ($p < 0.05$). For >0 % changes, the difference to TVDBN-0 was significant ($p < 0.05$) except for the difference in AUPRC scores for TVDBN-Exp-hard for 5 % changes ($p = 0.08$) and TVDBN-Exp-hard and TVDBN-Exp-soft for 10 % changes ($p = \{0.07, 0.18\}$). In all *plots*, the *horizontal bar of the boxplot* shows the median, the *box margins* show the 25th and 75th percentiles, the *whiskers* indicate data within 2 times the interquartile range, and *circles* are outliers. See Table 1 for hyperparameter settings

MCMC simulations with the novel multi-segment proposal move introduced in Sect. 3.6. It is seen that, unlike the centre panels in Fig. 5, and as a consequence of the different proposal scheme, the high posterior probability region now concurs with the region of maximum network reconstruction accuracy. This agreement suggests that the novel MCMC sampler leads to a significant improvement in mixing and convergence, in corroboration of our conjecture in Sect. 3.6.

Table 1 List of different information sharing (IS) priors for the TVDBN (Time-Varying Dynamic Bayesian Network), the equation where they were defined, and the most common hyperparameter settings that were used, or hyperparameter ranges if they are inferred. Only the highest level hyperparameters in the Bayesian hierarchy are shown

Name	Prior	Section	Equation	Hyperparameters
TVDBN-0	Poisson (No IS)	2.4	4, 6	$\Lambda = 3$
TVDBN-Exp-hard	Exponential Hard IS	3.2	30	$\beta \in [0, 20]$
TVDBN-Exp-soft	Exponential Soft IS	3.3	38	$\lambda_\kappa = 10$
TVDBN-Bino-hard	Binomial Hard IS	3.4	45, 46	$\alpha, \overline{\alpha}, \gamma, \overline{\gamma} \in \{1, 2, \ldots, 100\}$
TVDBN-Bino-soft	Binomial Soft IS	3.5	50	$\alpha, \overline{\alpha}, \gamma, \overline{\gamma} \in \{1, 2, \ldots, 100\}$

Next, we turn our attention to varying network structures. We varied the percentage of edges that change from segment to segment between 2.5 % to 10 %.[6] A significant improvement in the network reconstruction accuracy can be achieved over the unregularized method, as shown in the bottom panels of Fig. 3. However, the magnitude of the improvement in the scores decreases as the number of changes between adjacent segments increases. This is plausible: as we introduce more structural changes between adjacent networks, we would expect to gain less benefit from information sharing. We note that the degradation in performance seems to be stronger for the exponential prior than for the binomial prior.

We investigated whether the inferred hyperparameters coincide with the optimal reconstruction performance for the case where 10 % of the edges in the network change between adjacent segments. There are two effects to be traded off. Hyperparameter values that are too low will not bring about any improvement over the uncoupled unregularized scenario. Hyperparameter values that are too high will not allow the network structure to change with time. We would therefore expect to find some optimal finite range of hyperparameter values, $0 < \beta < \infty$ and $0 < a, b < 1$. This has in fact been borne out in our simulations. Figure 6 shows the network reconstruction accuracy in terms of AUROC and AUPRC scores for different values of the hyperparameters a, b. The best network reconstruction accuracy is obtained when b, which governs consistency among non-interactions, is high ($\geq$0.9), while a, which controls agreement among interactions, is reduced to a range around its uninformative setting $a \approx 0.5$. The bottom panel of Fig. 6 shows that the inferred posterior distribution is consistent with these ranges, and that the Bayesian inference scheme thus optimizes the network reconstruction accuracy. A slightly different picture emerges for the exponential prior, though. Figures 7(a)–7(b) show the AUROC and AUPRC scores for different values of β, indicating a clear peak in the network reconstruction accuracy for finite $0 < \beta < \infty$. This peak does not coincide with the high posterior probability range of β, as shown in Fig. 7(c). Only when increasing the data set size by a factor of 4 does the Bayesian inference scheme succeed in optimizing the network reconstruction accuracy in the sense that the high posterior probability region now coincides with the range of the highest AUROC/AUPRC scores. The obvious question to ask is whether this trend is another artifact of poor MCMC convergence/mixing. To this end we have devised a simplified model for which the posterior distribution can be computed in closed form. Our analysis, which we present in Sect. 5.2, re-

[6] Because our simulation was set up so that we had on average 3 regulatory interactions per node, this corresponds to a change of between 8.25 % and 33 % of the original interactions.

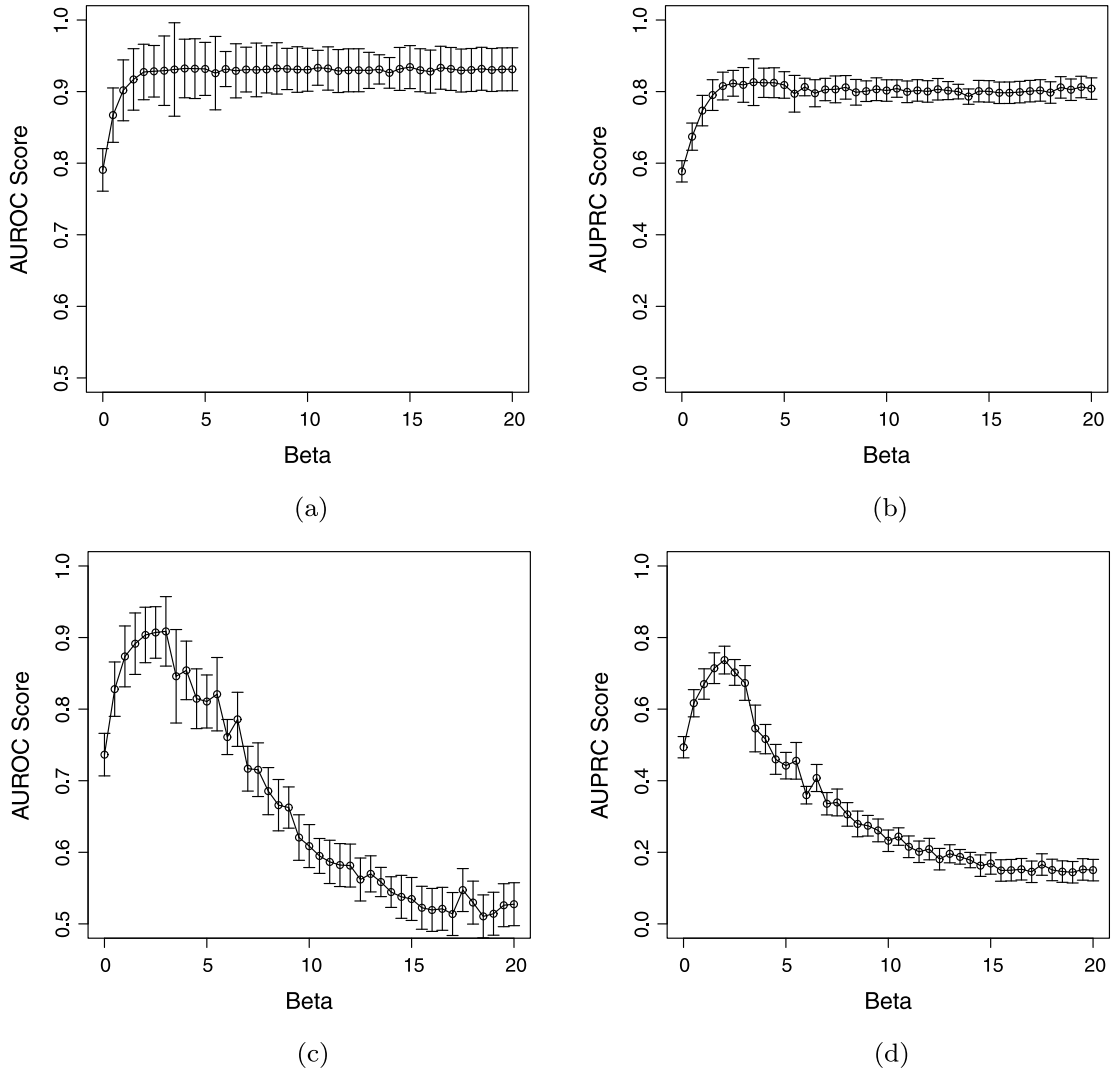


Fig. 4 Results for the exponential prior with hard coupling on the simulated data without mismatch among the structures. Panel (**a**) shows the AUROC scores for different values of the hyperparameter β. Panel (**b**) shows a corresponding plot for the AUPRC scores. The simulations were repeated on 10 independent data instantiations of time series length 60. The *error bars* show the standard error. The results were obtained with the novel MCMC sampler, described in Sect. 3.6. Panels (**c**) and (**d**) show the results from corresponding simulations with the old MCMC sampler adapted from Lèbre et al. (2010) and described in Sect. 2.6. The reconstruction performance deteriorates with larger values of the hyperparameter, as a consequence of poor MCMC mixing and convergence

produces the results from this simulation study, suggesting that the suboptimal performance of the Bayesian inference scheme is intrinsic to the chosen form of the prior.[7]

Returning to the binomial prior, we finally investigated the influence of the level-2 hyperparameters $\alpha, \overline{\alpha}, \gamma$, and $\overline{\gamma}$. Recall that owing to the conjugacy of the prior, these values can be interpreted as fictitious prior observation counts. Our initial idea was to keep the mismatch hyperparameters fixed at $\overline{\alpha} = \overline{\gamma} = 1$, while putting a vague uniform distribution over

[7]We note that the results for the exponential prior seem to be at odds with those reported in Husmeier et al. (2010). The reason is that in Husmeier et al. (2010) we had selected, by a fluke, a more restrictive prior on the hyperparameter: $\beta \in [0, 5]$. As our discussion in Sect. 5.2 shows, this setting boosts the network reconstruction performance.

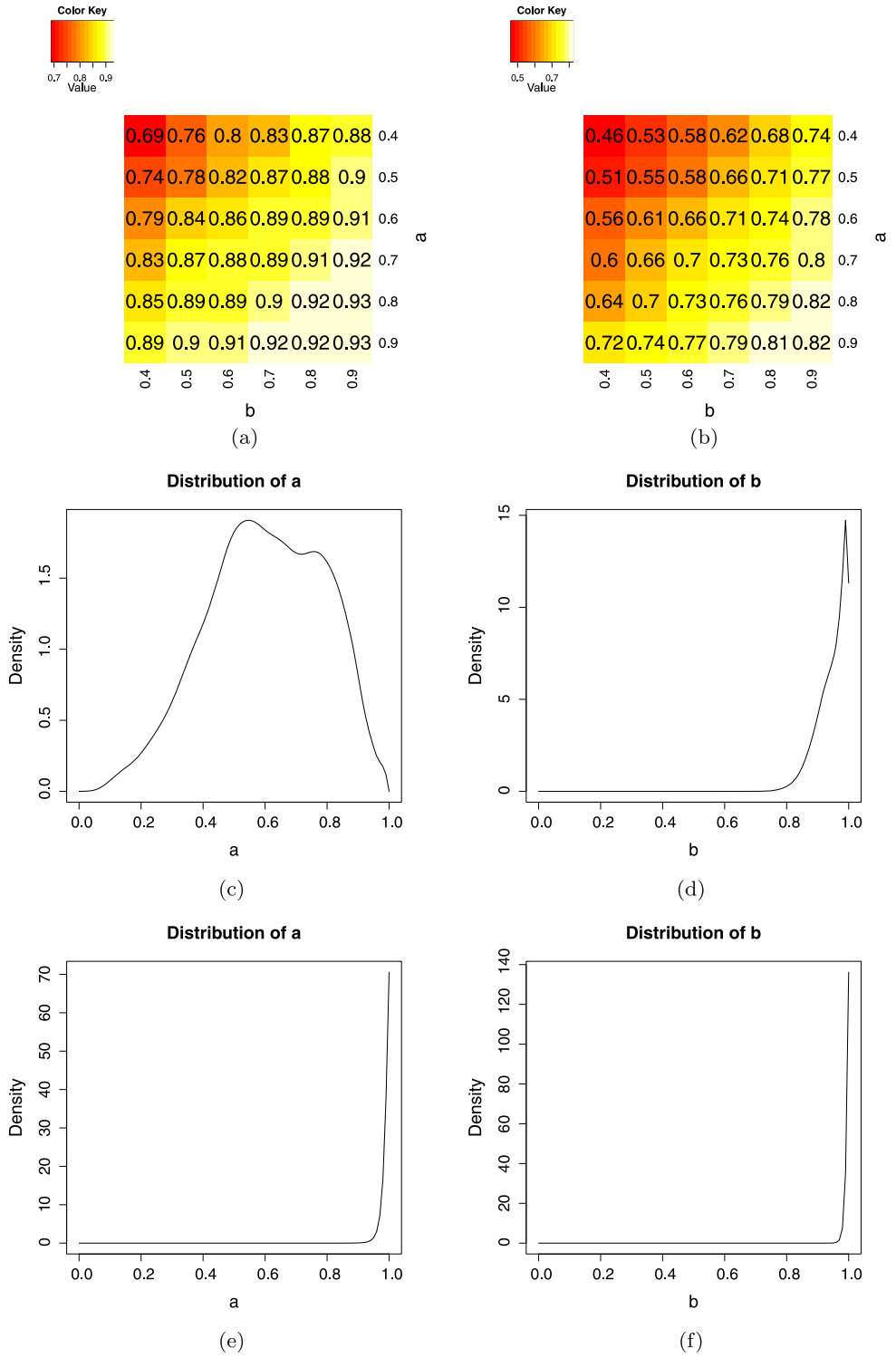


Fig. 5 Results for the binomial prior with hard coupling on the simulated data without mismatch among the structures. Panel (**a**) shows the AUROC scores for different values of the hyperparameters a and b. Panel (**b**) shows a corresponding plot for the AUPRC scores. Panels (**c**) and (**d**) show the marginal posterior distribution of the hyperparameters a and b, as obtained with the MCMC sampler adapted from Lèbre et al. (2010) and described in Sect. 2.6. Panels (**e**) and (**f**) show the marginal posterior distribution of the hyperparameters a and b, as obtained with the new MCMC sampler proposed in Sect. 3.6. The marginal distributions of a and b are obtained from the sampled values of the level-2 hyperparameters α, $\overline{\alpha}$, γ, $\overline{\gamma}$ and from the sampled networks using a kernel density estimator with the beta distribution from Eq. (45). The level-2 hyperparameters were given a uniform prior over the discrete set $\{1, 2, \ldots, 100\}$

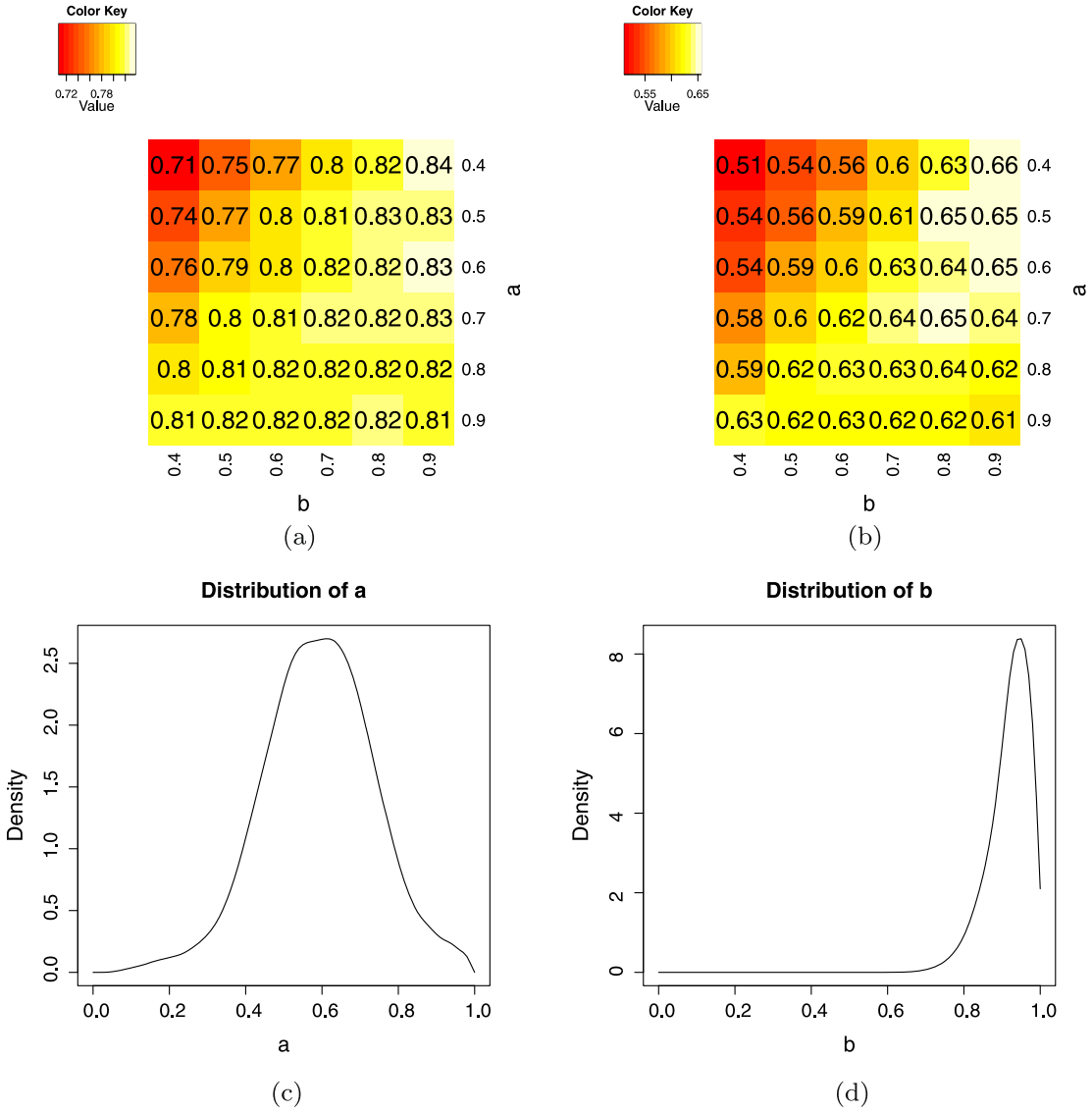


Fig. 6 Results for the binomial prior with hard coupling on the simulated data with mismatch among the structures. Panel (**a**) shows the AUROC scores for different values of the hyperparameters a and b. Panel (**b**) shows a corresponding plot for the AUPRC scores. Panels (**c**) and (**d**) show the marginal posterior distribution of the hyperparameters a and b, as obtained with the novel MCMC sampler proposed in Sect. 3.6. The marginal distributions of a and b were obtained from the sampled values of the level-2 hyperparameters α, $\overline{\alpha}$, γ, $\overline{\gamma}$ and from the sampled networks using a kernel density estimator with the beta distribution from Eq. (45). The level-2 hyperparameters were given a uniform prior over the discrete set $\{1, 2, \ldots, 100\}$

the set $\{1, 2, \ldots, 100\}$ as a prior on the match hyperparameters α and γ. The rationale behind this choice is that the regularization scheme is intended to encourage similarity rather than dissimilarity between adjacent network structures. However, repeating the MCMC simulations for different values of the level-2 hyperparameters revealed that the setting $\overline{\alpha} = \overline{\gamma} = 1$ is too restrictive and that the network reconstruction accuracy can be improved by relaxing this constraint (see Fig. 8).

The findings of our simulation study can be summarized as follows. A naive extension of the MCMC sampler of Lèbre et al. (2010), as described in Sect. 3.6, leads to a poor network reconstruction accuracy for high values of the hyperparameters; this problem can be resolved with the novel proposal scheme introduced in Sect. 3.6. With this new proposal

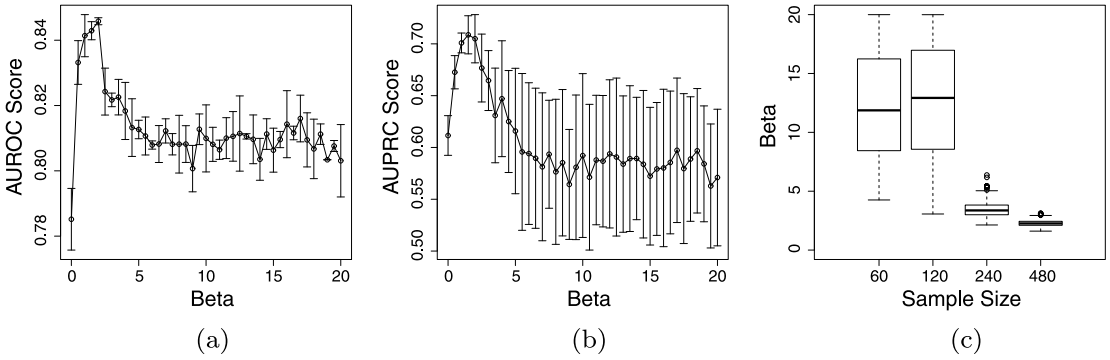


Fig. 7 Results for the exponential prior with hard coupling on the simulated data with mismatch among the structures. Panel (**a**) shows the AUROC scores and their standard deviations for different values of the hyperparameter β. Panel (**b**) shows a corresponding plot for the AUPRC scores. Panel (**c**) shows box plot representations of the inferred posterior distribution of β, for different sample sizes, using the MCMC scheme from Sect. 3.6. The *horizontal bar* shows the median, the *box margins* show the 25th and 75th percentiles, the *whiskers* indicate data within 2 times the interquartile range, and *circles* are outliers. The simulations were repeated on 10 independent data instantiations of time series length $n = 60$

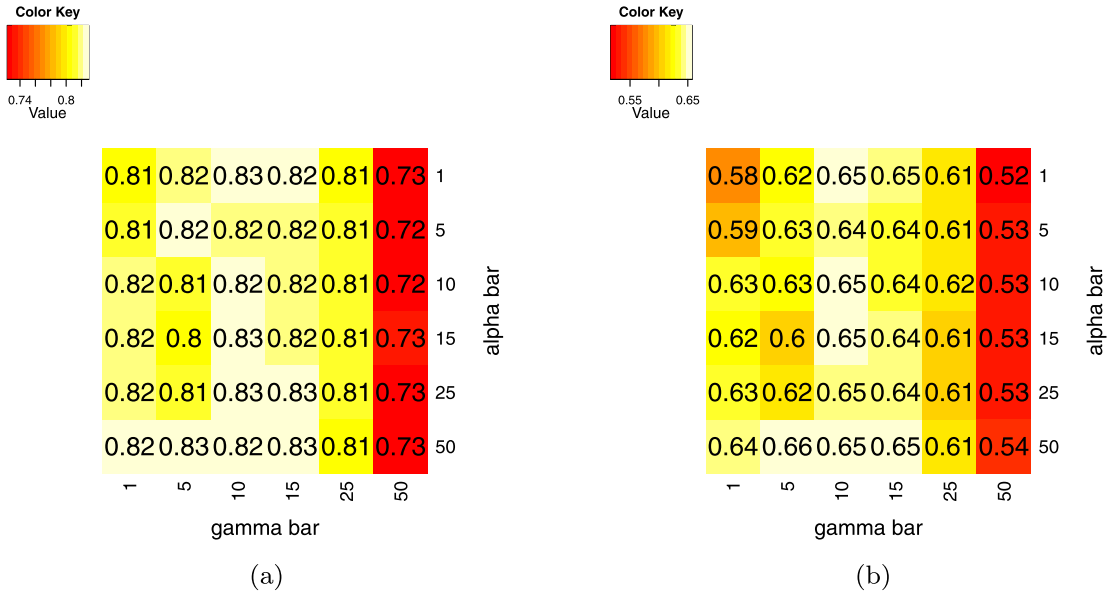


Fig. 8 Results for the binomial prior with hard coupling on the simulated data with mismatch among the structures: dependence of the reconstruction accuracy on the higher-level hyperparameters. Panel (**a**) shows the AUROC scores for different values of the level-2 hyperparameters $\overline{\alpha}$ and $\overline{\gamma}$. Panel (**b**) shows a corresponding plot for the AUPRC scores. The results indicate that setting $\overline{\alpha} = \overline{\gamma} = 1$ is over-restrictive and that the reconstruction accuracy improves as a consequence of employing a non-informative prior

scheme, information sharing with the binomial prior leads to a significant improvement in the network reconstruction accuracy in all cases, while information sharing with the exponential prior leads to a significant improvement when the true network structures are sufficiently similar. A detailed analysis of hyperparameter inference shows that the Bayesian inference scheme is consistent for the binomial prior in the sense that the high posterior probability region of the hyperparameters concurs with the one that optimizes the network reconstruction accuracy. For the exponential prior, this consistency is only given when the data set size is sufficiently large; otherwise a more restrictive hyperprior (i.e. prior on β) is needed. On the other hand, a restrictive setting for the level-2 hyperparameters of the bino-

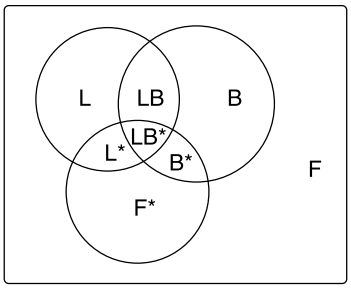


Fig. 9 Illustration of a hypothetical network scenario, where edges fall into several categories. Edges in sets L, LB, L^* and LB^* are true edges, which means they are included in the network corresponding to the current time series segment. Edges in sets L and LB are 'true positives' in that they contribute a score $A > 1$ to the likelihood. Edges in sets L^* and LB^* are 'false negatives', which contribute the neutral score of 1 to the likelihood. The edges in sets F^* and B^* are 'false positives', which contribute a score $A^* > A > 1$ to the likelihood. The edges in sets LB, LB^*, B^* and B are consistent with the prior network, all those in the complementary sets are not found in the prior network. Edges in set F are neither included in the network associated with the current segment, nor can they be found in the prior network. They also don't contribute any score to the log likelihood (i.e. they contribute a neutral score of 1 to the likelihood). An overview can be found in Table 2

mial prior is counter-productive, and better network reconstruction scores are obtained with a non-informative hyperprior.

5.2 Closed-form inference for the exponential prior

The results in Fig. 7 indicated that for the exponential prior, the Bayesian inference scheme might fail to find the hyperparameters that optimize the network reconstruction accuracy. Our conjecture is that this is not a consequence of poor mixing and convergence of the MCMC sampler, but intrinsic to the Bayesian inference scheme *per se*. As a demonstration, we reproduce the observation from Fig. 7 with a simpler model for which a closed-form expression of the posterior distribution of the hyperparameter can be derived. We consider the scenario depicted in Fig. 9, where edges of a hypothetical network can be divided into different categories, depending on whether or not they are true, supported by the data, or included in the prior network. An overview of the notation is presented in Table 2. With the simplifying assumption of posterior independence of the edges, the likelihood is given by

$$P(\boldsymbol{x}|\boldsymbol{G}) = A^{(n_L+n_{LB})} A^{*(n_{B^*}+n_{F^*})} \tag{57}$$

where n_S counts the number of elements in set S for network $\boldsymbol{G}$, and the symbols denoting the sets have been defined in Table 2. Assuming a uniform prior on β, the posterior distribution of the hyperparameter becomes:

$$\begin{aligned} P(\beta|\boldsymbol{x}) \propto P(\boldsymbol{x},\beta) &= \sum_{\boldsymbol{G}} P(\boldsymbol{x}|\boldsymbol{G}) P(\boldsymbol{G}|\beta) P(\beta) \\ &\propto \frac{1}{Z(\beta)} \sum_{\boldsymbol{G}} P(\boldsymbol{x}|\boldsymbol{G}) \exp\bigl(-\beta \bigl|\boldsymbol{G} - \boldsymbol{G}^0\bigr|\bigr) \end{aligned} \tag{58}$$

Table 2 Likelihood and prior scores for the edges contained in the sets defined in Fig. 9. The product of the prior and the likelihood defines the rank of the edge; the truth indicator is shown in the second column

Set	True edge	Supported by the data	Supported by the prior	Likelihood	Prior	Number of edges
L	yes	yes	no	A	$e^{-\beta}$	N_L
LB	yes	yes	yes	A	1	N_{LB}
LB^*	yes	no	yes	1	1	N_{LB^*}
L^*	yes	no	no	1	$e^{-\beta}$	N_{L^*}
B	no	no	yes	1	1	N_B
B^*	no	yes	yes	A^*	1	N_{B^*}
F^*	no	yes	no	A^*	$e^{-\beta}$	N_{F^*}
F	no	no	no	1	$e^{-\beta}$	N_F

where $\boldsymbol{G}^0$ represents our prior knowledge. Inserting (57) into (58) we get, with Eq. (31) for $Z(\beta)$ and under the assumption of a uniform prior on β:

$$
\begin{aligned}
P(\beta|\boldsymbol{x}) \propto \frac{1}{(1+e^{-\beta})^N} & \sum_{n_L=0}^{N_L} \sum_{n_{LB}=0}^{N_{LB}} \sum_{n_B=0}^{N_B} \sum_{n_F=0}^{N_F} \sum_{n_{L^*}=0}^{N_{L^*}} \sum_{n_{LB^*}=0}^{N_{LB^*}} \sum_{n_{B^*}=0}^{N_{B^*}} \sum_{n_{F^*}=0}^{N_{F^*}} \\
& \times \binom{N_L}{n_L}\binom{N_{LB}}{n_{LB}}\binom{N_B}{n_B}\binom{N_F}{n_F}\binom{N_{L^*}}{n_{L^*}}\binom{N_{LB^*}}{n_{LB^*}}\binom{N_{B^*}}{n_{B^*}}\binom{N_{F^*}}{n_{F^*}} \\
& \times A^{(n_L+n_{LB})} A^{*(n_{B^*}+n_{F^*})} \\
& \times \exp\big(-\beta[n_L + n_F + N_{LB} - n_{LB} + N_B - n_B \\
& + n_{L^*} + n_{F^*} + N_{LB^*} - n_{LB^*} + N_{B^*} - n_{B^*}]\big)
\end{aligned} \tag{59}
$$

A plot of (59) is shown in Fig. 10. The optimal network reconstruction in terms of AUROC and AUPRC scores is achieved for a finite value of $\beta \approx 1$. The effect of the data set size is emulated by varying the settings of the parameters entering the likelihood. For small values of A and A^*, corresponding to small data sets, the posterior probability increases monotonically in β, and the Bayesian inference scheme intrinsically fails to find the range of hyperparameters that optimizes the network reconstruction accuracy. When we increase the data set size, this mismatch disappears, and the two regions concur. These findings are consistent with those presented in Fig. 7 and suggest that the observed mismatch is a genuine inference feature rather than an MCMC artifact.

To further analyse this effect, we have investigated the values of A and A^* for which the posterior distribution shows a peak for a finite value of β. Analytically, this corresponds to finding values for A and A^* such that the equation $\frac{dP(\beta|\boldsymbol{x})}{d\beta} = 0$ has a solution. Unfortunately, it is non-trivial to determine the existence of a solution analytically; we have therefore resorted to numerically calculating $\frac{dP(\beta|\boldsymbol{x})}{d\beta}$ for $\beta = 20$. At $\beta = 0$, we have $\frac{dP(\beta|\boldsymbol{x})}{d\beta} > 0$; therefore, if $\frac{dP(\beta|\boldsymbol{x})}{d\beta} < 0$ at $\beta = 20$, this indicates that the distribution has a peak on the interval $[\beta, 20]$. On the other hand, under the assumption of unimodality, $\frac{dP(\beta|\boldsymbol{x})}{d\beta} > 0$ at $\beta = 20$ indicates that the marginal posterior probability of β increases monotonically with β. The results of this analysis are shown in Fig. 11, which shows a clear phase shift towards distributions with a peak as A and A^* increase.

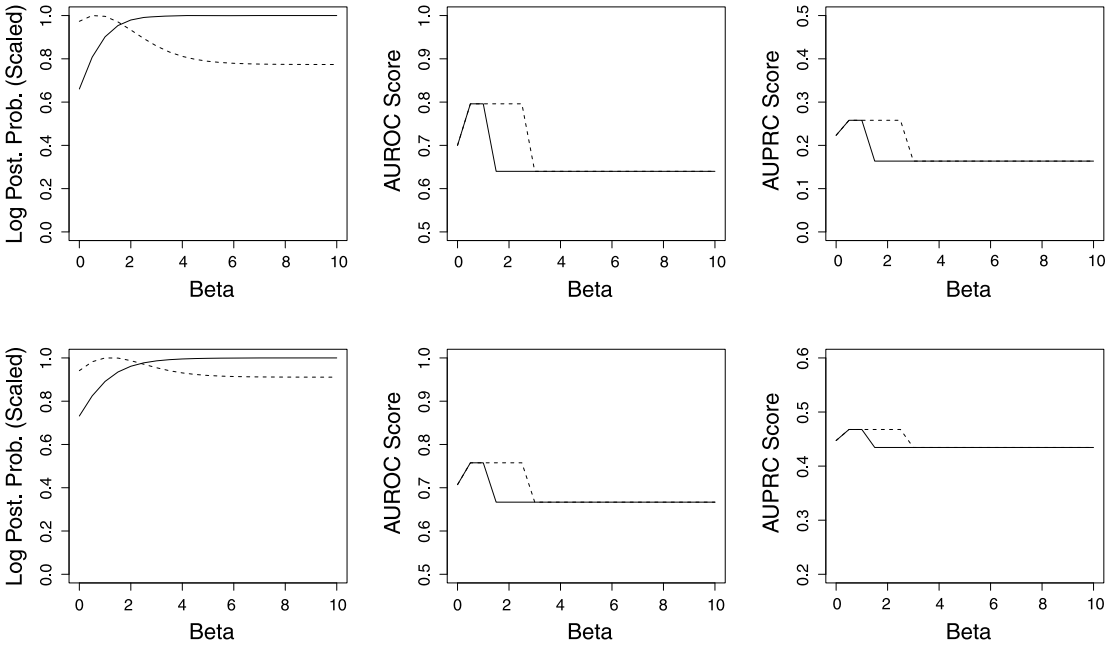


Fig. 10 Results for the simplified model with exponential prior. The *leftmost column* shows the marginal posterior distribution of β, computed from Eq. (59). The *middle column* shows the AUROC score as β varies. The *rightmost column* shows the AUPRC score as β varies. *Solid line*: $A = 2$, $A^* = 4$, *dashed line*: $A = 12$, $A^* = 14$. The *top* and *bottom rows* correspond to two different settings of the set sizes. *Top row*: $\{L : 15, LB : 0, B : 40, F : 60, L^* : 0, LB^* : 10, B^* : 25, F^* : 0\}$. *Bottom row*: $\{L : 15, LB : 20, B : 10, F : 25, L^* : 0, LB^* : 10, B^* : 20, F^* : 0\}$

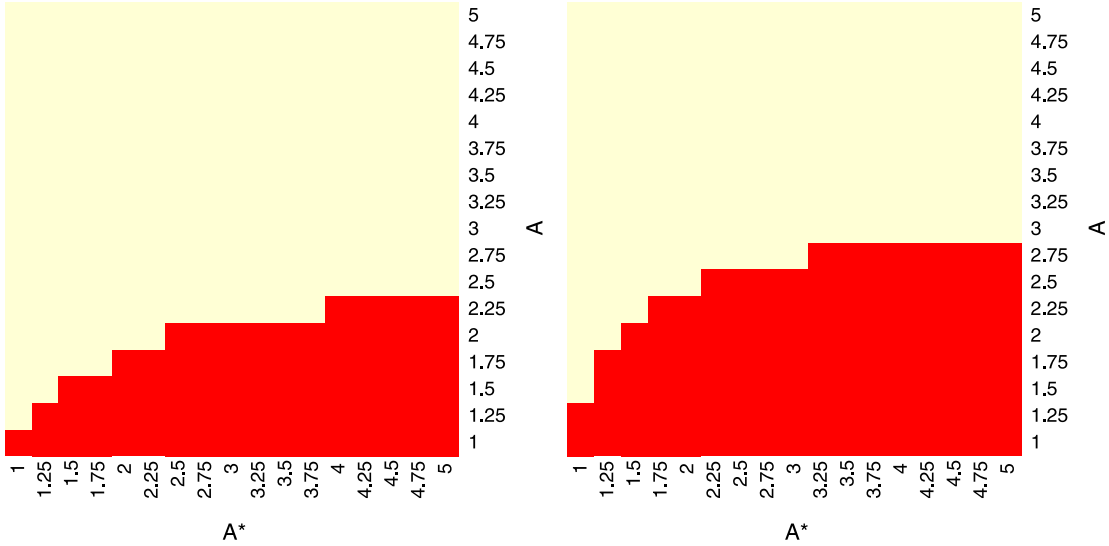


Fig. 11 Existence of a peak in the posterior distribution of β for the simplified model with exponential prior. The *two plots* show values of A and A^* for which the marginal posterior probability of β monotonically increases as β increases (*red tiles*), and those where the posterior probability decreases for high β (*white tiles*), indicating the existence of a peak in the distribution. We used the same settings of the set sizes as in Fig. 10. *Left*: $\{L : 15, LB : 0, B : 40, F : 60, L^* : 0, LB^* : 10, B^* : 25, F^* : 0\}$. *Right*: $\{L : 15, LB : 20, B : 10, F : 25, L^* : 0, LB^* : 10, B^* : 20, F^* : 0\}$

What does this analysis entail for the general applicability of the exponential prior? It is clear that when the data set size is too small, then the marginal posterior distribution of β will be biased towards high values. The exact definition of "too small" will crucially depend on the nature of the dataset. Given that we have shown in Sect. 5.1 that the binomial prior avoids this weakness and outperforms the exponential prior in terms of network reconstruction accuracy, we would recommend that this form of information sharing prior be used in preference of the exponential prior.

6 Real-world applications

6.1 Morphogenesis in *Drosophila melanogaster*

During its life-cycle, *Drosophila melanogaster* undergoes four major stages of morphogenesis: embryo, larva, pupa and adult. Arbeitman et al. (2002) obtained a gene expression time series covering all four stages. We have applied our methods to a subset of this gene expression time series consisting of eleven genes involved in wing muscle development. First, we investigated whether the changepoints inferred by our methods correspond to the known transitions between stages. Figure 12(a) shows the posterior probabilities of inferred changepoints for any gene using TVDBN-0 (unregularized by information sharing, see Table 1), while Figs. 12(c)–12(d) show the posterior probabilities for the information sharing methods. We compared this performance to the method proposed in Ahmed and Xing (2009), using the authors' software package TESLA (Fig. 12(b)). In addition, Robinson and Hartemink (2009) used a discrete non-homogeneous DBN to analyse the same data set, and a plot corresponding to Fig. 12(b) can be found in their paper.

An analysis of the results suggests that our non-homogeneous DBN methods are generally more successful than TESLA. We recover changepoints for all three transitions (embryo $\rightarrow$ larva, larva $\rightarrow$ pupa, and pupa $\rightarrow$ adult). As shown in Fig. 12(b), the last transition, pupa $\rightarrow$ adult, is less clearly detected with TESLA, and it is completely absent in Robinson and Hartemink (2009). Furthermore, TESLA and our method both detect additional changepoints during the embryo stage, which are missing in Robinson and Hartemink (2009). It is not implausible that additional transitions at the gene regulatory network level should occur within one morphogenic phase. One would expect that a complex gene regulatory network is unlikely to transition into a new phase all at once, and some pathways might have to undergo activational changes earlier in preparation for the morphogenic transition. However, a failure to detect a known transition represents a shortcoming of a method, and so we can say that in this aspect, our model appears to outperform the two alternative approaches.

In addition to the changepoints, we have inferred network structures for the morphogenic stages of embryo, larva, pupa and adult (see Fig. 13). An objective assessment of the reconstruction accuracy is not feasible due to the limited existing biological knowledge and the absence of a gold standard. However, our reconstructed networks show many similarities with the networks discovered by Robinson and Hartemink (2009), Guo et al. (2007) and Zhao et al. (2006). For instance, we recover the interaction between two genes, *eve* and *twi*. This interaction is also reported in Guo et al. (2007) and Zhao et al. (2006), while Robinson and Hartemink (2009) seem to have missed it. We also recover a cluster of interactions among the genes *myo61f*, *msp300*, *mhc*, *prm*, *mlc1* and *up* during all morphogenic phases. This result is not implausible, as all genes (except *up*) belong to the myosin family. However, unlike Robinson and Hartemink (2009), we find that *actn* also participates as a regulator in this cluster. There is some indication of this in Zhao et al. (2006), where *actn* is found to regulate *prm*. We have further validated our reconstructed networks using genetic and protein

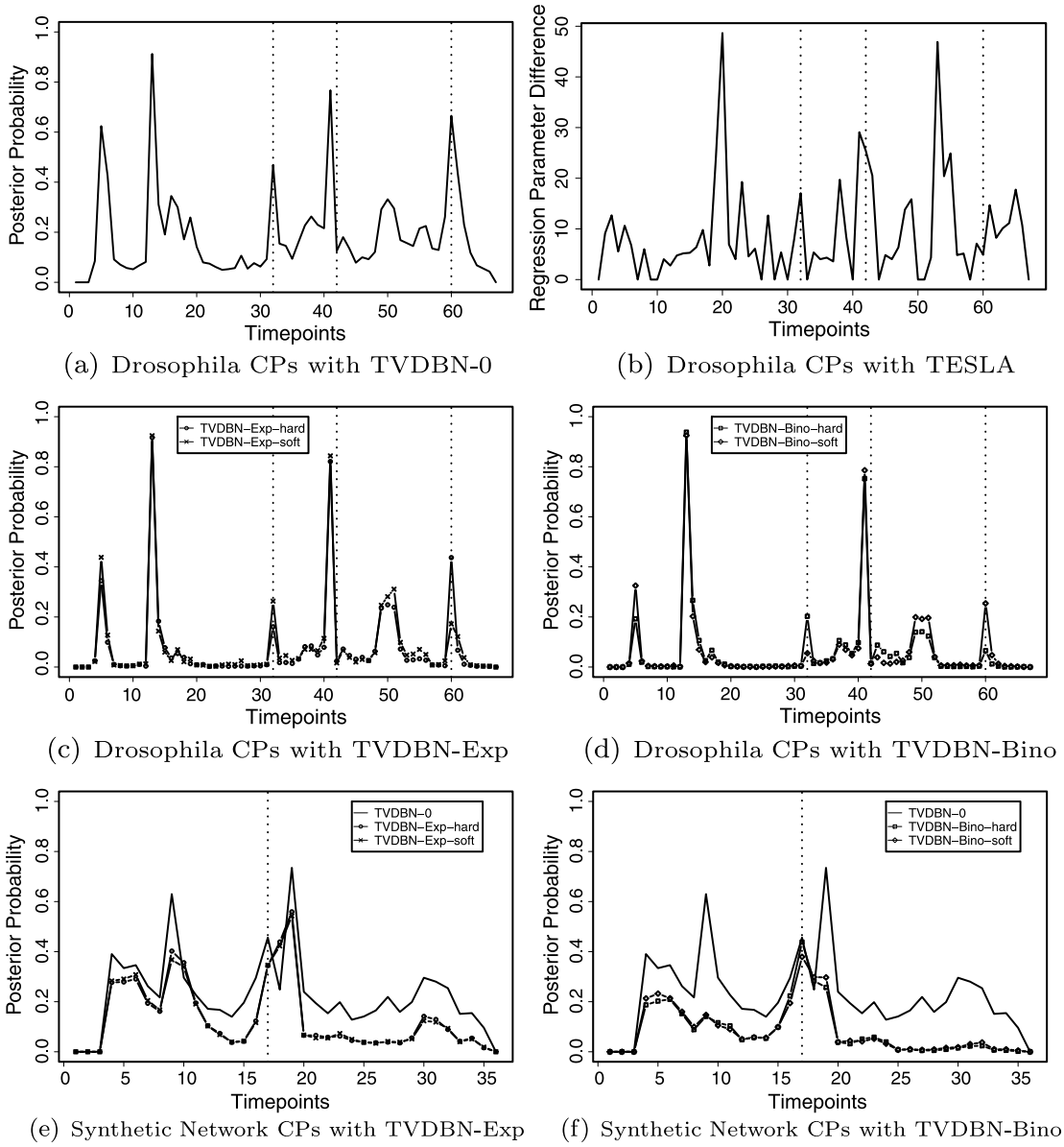


Fig. 12 Changepoints inferred from gene expression time series related to morphogenesis in *Drosophila melanogaster*, and synthetic biology in *Saccharomyces cerevisiae* (yeast). (**a**): TVDBN-0 changepoints for Drosophila (no information sharing). (**b**): TESLA, L1-norm of the difference of the regression parameter vectors associated with two adjacent time points plotted against time. (**c**) and (**d**): TVDBN changepoints for Drosophila with information sharing; the method is indicated by the legend. (**e**) and (**f**): TVDBN changepoints for the synthetic gene regulatory network in yeast. All figures using TVDBN plot the posterior probability of a changepoint occurring for any node at a given time (ordinate) against time (abscissa). In (**a**)–(**d**), the *vertical dotted lines* indicate the three morphogenic transitions, while in (**e**) and (**f**) the *line* indicates the boundary between the "switch on" (galactose) and "switch off" (glucose) phases

interactions recorded in the FLIGHT database (Sims et al. 2006). We found that a number of the inferred interactions over all segments correspond to interactions that have been reported in the literature. Some of these result from indirect interactions, where the intermediate gene is missing in the data. Table 3 gives an overview of the identified interactions with references to the biological literature.

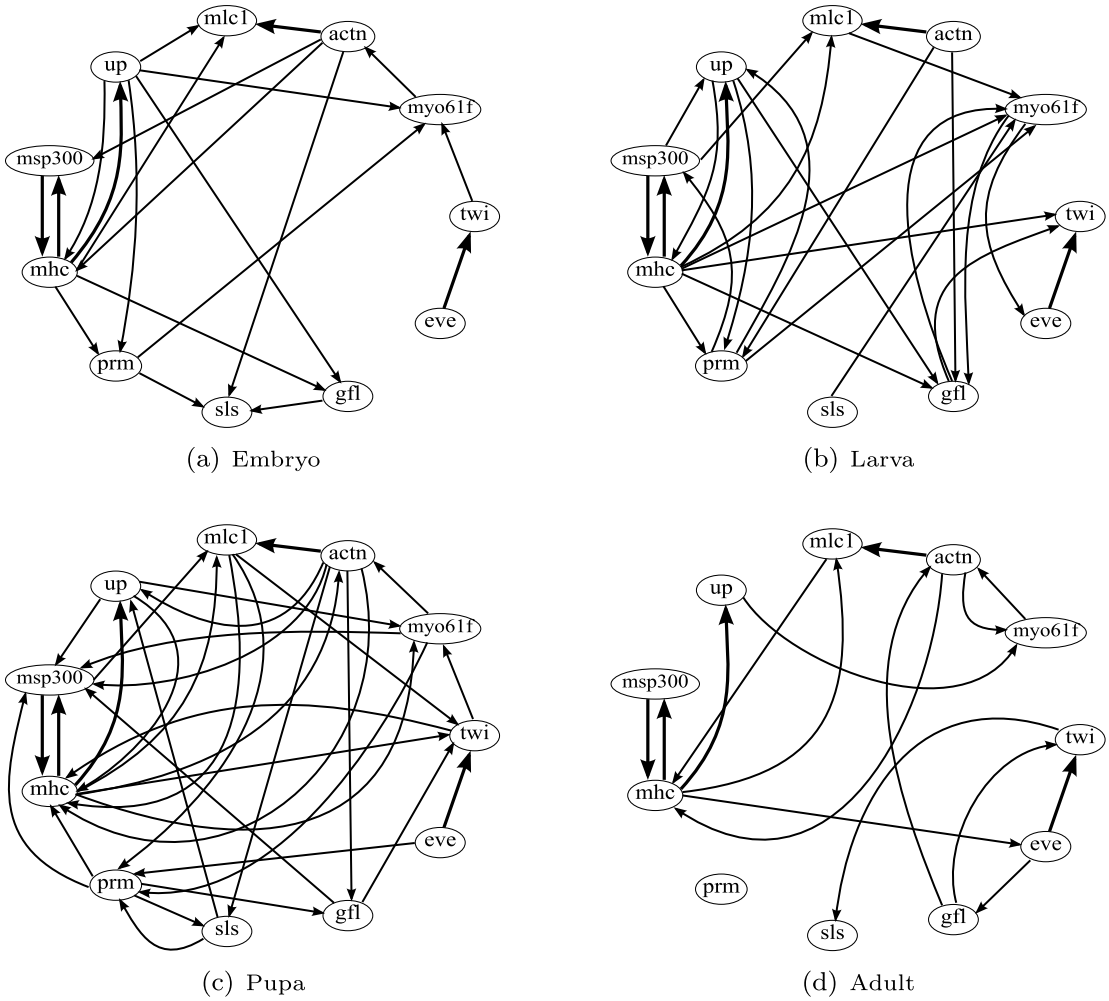


Fig. 13 Gene regulatory networks inferred from gene expression time series related to morphogenesis in *Drosophila melanogaster*, using TVDBN-Bino-hard. The networks were obtained by applying a threshold of 0.25 to the marginal posterior probabilities of the gene interactions. We have reconstructed a network for each morphological phase; interactions that were consistent across all four phases are marked in *bold*

6.2 Synthetic biology in *Saccharomyces cerevisiae*

Synthetic biology is a rapidly developing and highly topical discipline that aims to combine the biological sciences and engineering (Andrianantoandro et al. 2006). One of its aims is to design new gene regulatory networks in living cells. We make use of these endeavours by using gene expression time series obtained *in vivo* from cells with a known gene regulatory network structure to objectively assess the network reconstruction accuracy. Our work is based on Cantone et al. (2009), where the authors constructed a synthetic regulatory network with 5 genes in *Saccharomyces cerevisiae* (yeast). Then they measured gene expression time series with RT-PCR for 16 and 21 time points under two experimental conditions, related to the carbon source: galactose ("switch on"), and glucose ("switch off"). The authors applied two established state-of-the-art methods from computational systems biology to reconstruct the known underlying network from these time series. One is based on ODEs: ordinary differential equations (TSNI), the other is based on conventional DBNs (Banjo); see Cantone et al. (2009) for details. Both methods are optimization-based and only output a single network. By comparison with the known network, the authors calculated the

Table 3 Reconstructed interactions in the *Drosophila melanogaster* wing muscle development network that have been validated using the FLIGHT database (Sims et al. 2006)

Interaction	References	Interaction	Notes
actn ↔ *mhc*	Homyk and Emerson (1988); Nongthomba et al. (2003),; Montana and Littleton (2004)	Protein	Via missing gene *wupA*
actn → *up*	Homyk and Emerson (1988); Nongthomba et al. (2003)	Protein	Via missing gene *wupA*
eve → *twi*	Parkhurst and Ish-Horowicz (1991)	Protein	Via missing gene *RpIII40*
up ↔ *mhc*	Homyk and Emerson (1988); Nongthomba et al. (2003); Montana and Littleton (2004)	Protein	Direct interaction
actn → *msp300*	Formstecher et al. (2005)	Gene	Via missing gene *TSG101* or missing gene *Hrs*
actn → *sls*	Sanchez et al. (1999)	Gene	Direct Interaction
actn → *prm*	Formstecher et al. (2005)	Gene	Via missing gene *exo70*
prm ↔ *sls*	Sanchez et al. (1999); Formstecher et al. (2005)	Gene	Via missing gene *exo70* and present gene *actn*
sls → *up*	Sanchez et al. (1999); Formstecher et al. (2005)	Protein and Gene	Via missing gene *Act88F*

precision (proportion of predicted regulatory interactions in the network that are correct) and recall (proportion of predicted true interactions) scores. Figure 14 shows the true networks, the reconstructed networks for TSNI and Banjo, as well as the reconstructed networks using TVDBN-Bino-hard, where we have applied a threshold of 0.75 to the inferred marginal posterior probabilities of the gene interactions to obtain absence/presence values for the edges.[8]

In our study, we merged the time series from the two experimental conditions under exclusion of the boundary point,[9] and applied the non-homogeneous DBNs from Table 1. Figures 12(e) and 12(f) show the inferred marginal posterior probabilities of potential changepoints. The salient changepoint is at the boundary between the "switch on" (galactose) and "switch off" (glucose) phases, confirming that the true changepoint is consistently identified. However, in the absence of information sharing, we observe additional spurious changepoints. These changepoints are successfully suppressed with the proposed Bayesian information-coupling schemes, with the binomial prior having a slightly stronger regularizing effect than the exponential one.

As described in Sect. 4, the Bayesian inference scheme provides a ranking of the potential gene interactions in terms of their marginal posterior probabilities. From this ranking we computed the precision-recall curves (Davis and Goadrich 2006) shown in Fig. 15. By using information sharing, our non-homogeneous DBN outperforms Banjo and TSNI both in the "switch on" and the "switch off" phase. The information sharing methods also perform better than TVDBN-0 on the "switch off" data, but are slightly worse on

[8]Note that while our TVDBN methods are in principle capable of inferring the type of interaction (activation or inhibition) by sampling regression weights, we have not investigated this for the purpose of this paper. Therefore in Fig. 14, the arrows in the networks reconstructed using TVDBN-Bino-hard only record the presence or absence of an interaction, and not its type.

[9]When merging two time series $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$, only the pairs $x_i \to x_j$ and $y_i \to y_j$ are presented to the DBN, while the pair $x_m \to y_1$ is excluded due to the obvious discontinuity.

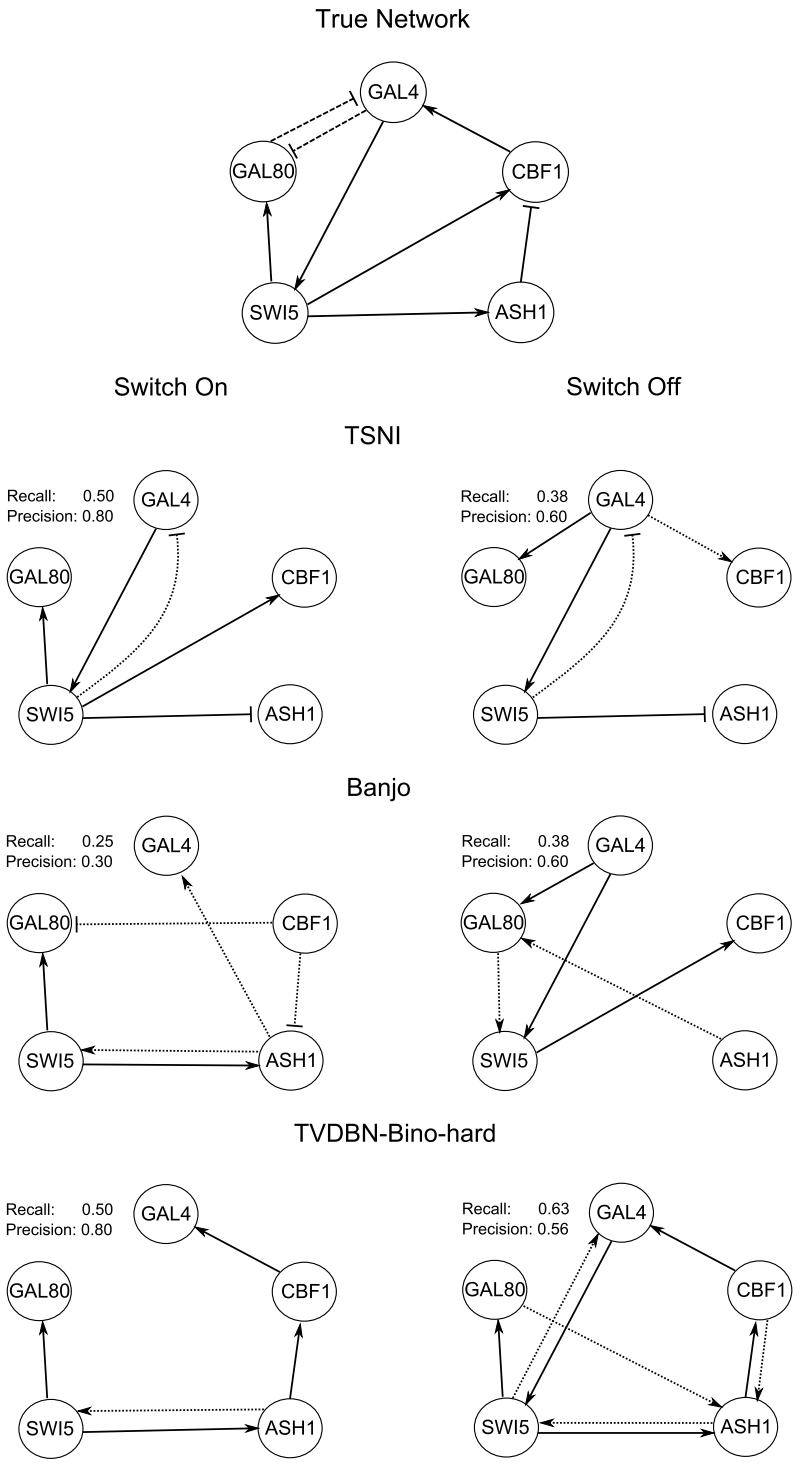


Fig. 14 True and reconstructed networks for a synthetic biology gene regulatory network in *Saccharomyces cerevisiae* (yeast). *Top row*: True network as described in Cantone et al. (2009). *2nd row*: Networks reconstructed using TSNI, a method based on ordinary differential equations (ODEs). *3rd row*: Networks reconstructed using Banjo, a conventional DBN. *Bottom row*: Networks reconstructed using TVDBN-Bino-hard, applying a threshold of 0.75 on the marginal posterior probabilities of gene interactions to obtain an absence/presence value for each edge. All reconstructed networks were reconstructed from two gene expression time series obtained with RT-PCR in two experimental conditions, reflecting the switch in the carbon source from galactose ("switch on") to glucose ("switch off"). The *dashed lines* in the true network indicate protein-protein regulation. The *dotted lines* in the reconstructed networks indicate false positive gene interactions. The networks found by Banjo and TSNI are reproduced from Cantone et al. (2009)

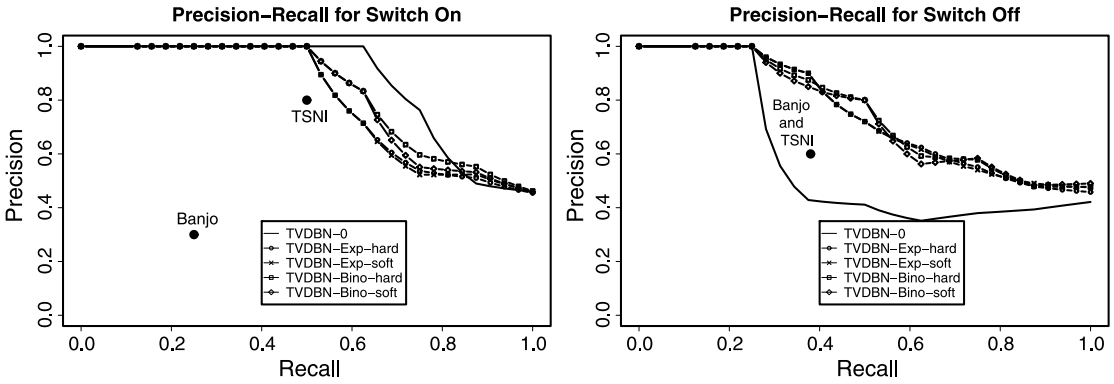


Fig. 15 Reconstruction of a gene regulatory network designed with synthetic biology in *Saccharomyces cerevisiae*. The network was reconstructed from two gene expression time series obtained with RT-PCR in two experimental conditions, reflecting the switch in the carbon source from galactose ("switch on") to glucose ("switch off"). The reconstruction accuracy of the methods proposed in Sect. 3 and Table 1, where the legend is explained, is shown in terms of precision (*vertical axis*)–recall (*horizontal axis*) curves. Results were averaged over 10 independent MCMC simulations. For comparison, fixed precision/recall scores are shown for two state-of-the-art methods, as reported in Cantone et al. (2009): Banjo, a conventional DBN, and TSNI, a method based on ordinary differential equations (ODEs)

the "switch on" data. Cantone et al. (2009) showed that in general, the reconstruction accuracy on the "switch off" data is poorer than on the "switch on" data. This lends credence to our results, suggesting that the proposed Bayesian regularization and information sharing schemes substantially improve the gene network reconstruction accuracy on the poorer time series segment, at the cost of a slightly degraded performance on the stronger one. Overall, the effect of information sharing is a performance improvement, as shown by the average areas under the PR curves, averaged over both phases ("switch on and off"): TVDBN-0 = 0.68, TVDBN-Exp-hard = 0.74, TVDBN-Exp-soft = 0.74, TVDBN-Bino-hard = 0.76, TVDBN-Bino-soft = 0.75.

We complete our investigation of the yeast network by providing an analysis of the network reconstruction performance (in terms of average area under the PR curve) as the hyperparameters vary. This is analogous to the evaluation we performed in Sect. 5.1 on simulated data. The results are shown in Fig. 16. As expected, higher values of the hyperparameter β, which correspond to stronger coupling, result in a better performance (Fig. 16(a)). Figure 16(b) shows the effect of different values for κ in Eq. (37). There is no discernible trend, which suggests that the strength of the coupling scheme does not matter much for this application, and that when moving closer to the hard coupling scheme (higher κ while keeping the mean μ of the gamma distribution fixed), the network reconstruction performance does not change significantly. The results obtained with the binomial prior demonstrate that, for this application, encouraging agreement related to the presence of interactions is more important than agreement related to the absence of interactions (Fig. 16(c)). Figure 16(d) confirms that our sampled hyperparameters a and b are in the correct range for optimal network reconstruction.

7 Discussion

In the present paper we have addressed some of the challenges encountered in systems biology when attempting to reconstruct gene regulatory networks from gene expression time series. We have looked at the case where the network structure may change

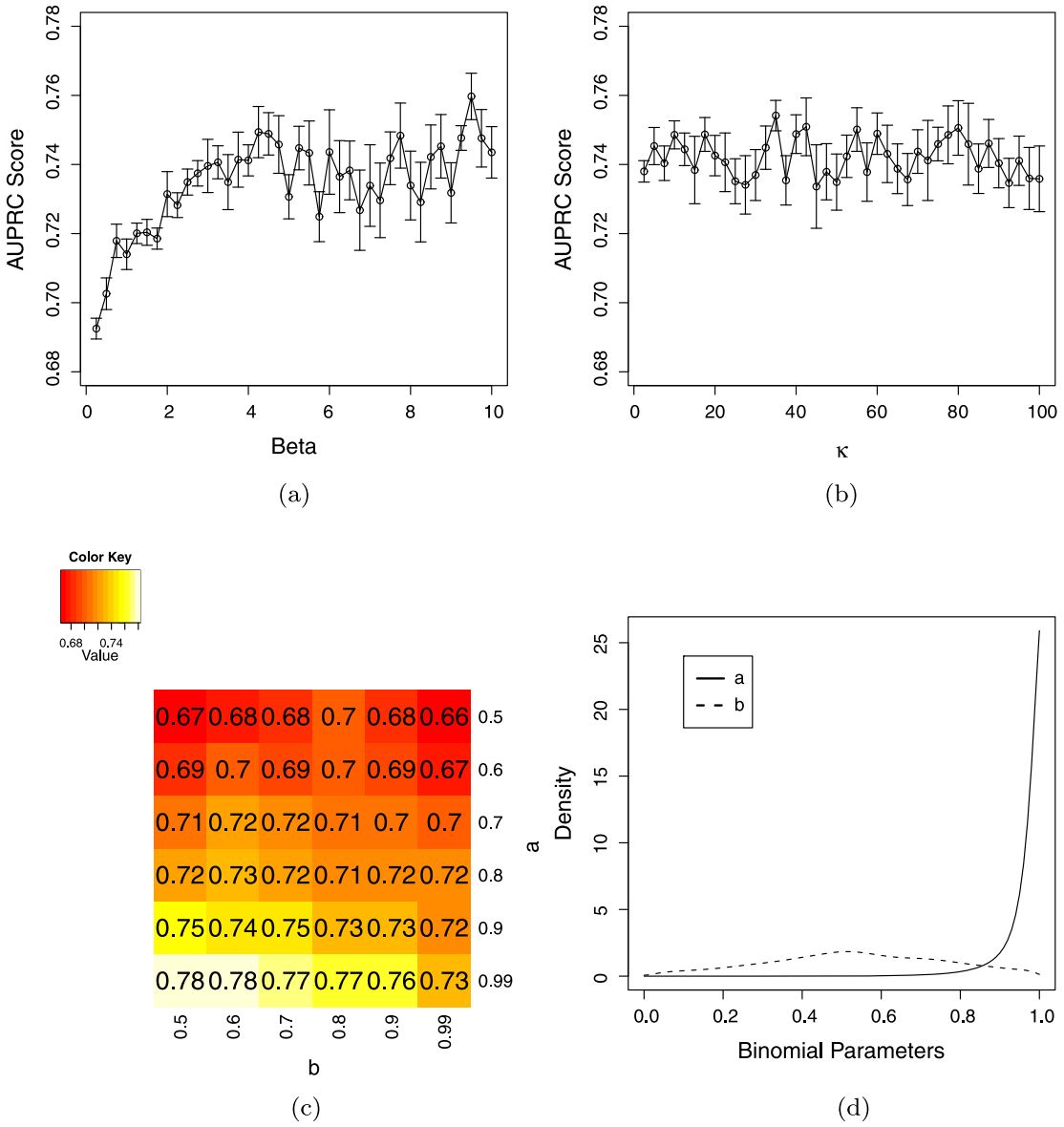


Fig. 16 Effect of the hyperparameters on the reconstruction of a known gene regulatory network from synthetic biology in yeast. The reconstruction accuracy is measured in terms of the average area under the precision–recall curve (AUPRC). Results were averaged over 10 independent MCMC simulations. (**a**): Variation of the hyperparameter β for the exponential information sharing prior with hard coupling. (**b**): Variation of the level-2 hyperparameter κ for the exponential prior with soft coupling, where the mean of the gamma distribution is kept fixed at $\mu = 5$. (**c**): Variation of hyperparameters a and b for the binomial prior. (**d**): Sampled distributions of hyperparameters a and b for the binomial prior with hard coupling. These distributions were obtained from the sampled values of the level-2 hyperparameters α, $\overline{\alpha}$, γ, $\overline{\gamma}$ using a kernel density estimator with the beta distribution from Eq. (45)

over time due to developmental or environmental causes. To deal with this situation, we have developed a non-homogeneous DBN, which has various advantages over existing schemes: it does not require the data to be discretized (as opposed to Robinson and Hartemink 2009, 2010); it allows the network structure to change with time (as opposed to Grzegorczyk and Husmeier 2009, 2011); it includes four different regularization schemes based on inter-time segment information sharing (as opposed to Lèbre 2007;

Lèbre et al. 2010); and it allows all hyperparameters to be inferred from the data via a consistent Bayesian inference scheme (as opposed to Ahmed and Xing 2009).

We note that the model of Robinson and Hartemink (2009, 2010) is conceptually similar to our exponential information sharing prior with hard coupling described in Sect. 3.2. By including three alternative information sharing schemes, we have extended the model of Robinson and Hartemink (2009, 2010) in two further respects:

(1) We allow for different penalties between edges and non-edges. The method in Robinson and Hartemink (2009, 2010) simply penalizes the number of different edges, i.e. the Hamming distance, between two adjacent structures. This corresponds to the approach taken for the exponential prior in Sects. 3.2 and 3.3. The inclusion of an extra edge leads to the same penalty as the deletion of an existing edge. This might not always be appropriate. Removing a rate-limiting reaction step of a critical signalling pathway is a more substantial change than including some redundant bypass pathway. Our two models based on the binomial prior (Sects. 3.4 and 3.5) allow for that by introducing different prior penalties for the deviation between edges and for the deviation between non-edges. In Sect. 5.1 we have experimentally shown that an information sharing approach based on different penalties for edges and non-edges can outperform the simpler approach when the number of changes among segments is small, but non-zero.
(2) We allow for different nodes of the network to have different penalty terms. The model in Robinson and Hartemink (2009, 2010) has a single hyperparameter for penalizing differences between structures: λ_s. This might not be appropriate if different subnetworks are conserved to a different degree. For instance, we would assume that molecular network substructures related to generic functionality, e.g. to maintain an essential baseline metabolism, are conserved to a greater extent than more peripheral pathways. By introducing node-dependent hyperparameters, the priors described in Sects. 3.3 and 3.5 generalize the approach in Robinson and Hartemink (2009, 2010) by allowing different parts of the network to be conserved during the temporal process to a different extent.

A further difference to Robinson and Hartemink (2009, 2010) merits some additional discussion. In our model, the changepoints are node-dependent. This gives us extra model flexibility, which is biologically motivated: on infection of an organism by a pathogen, genes involved in defence pathways are likely to be up-regulated, while others are not. Hence, it is plausible that different genes respond to changes in the environment differently, and this is directly incorporated in our model. In Robinson and Hartemink (2010), node-specific changepoints can be obtained indirectly: the calculation of the sufficient statistics for computing the marginal likelihood depends on the intervals during which each parent set is active. The marginal likelihood is recomputed for epochs, where an epoch is the union of consecutive time intervals during which a node-dependent substructure does not change. Since these unions of sets can be different for different nodes, the model does allow different changepoint sets to be associated with different nodes. However, there is a considerable price to pay for that: a changepoint in Robinson and Hartemink (2010) is intrinsically associated with a structure change, whereas in our model, a changepoint can be related to either a structure or a parameter change, or both. This gives us extra model flexibility, which is important for systems biology: when adapting to environmental change, several molecular interactions in signalling pathways may be up- or down-regulated, rather than switched on or off altogether.

An evaluation on simulated data has demonstrated that the proposed Bayesian regularization and information sharing schemes lead to an improved performance over Lèbre (2007) and Lèbre et al. (2010). We have carried out a comparative evaluation of four different information coupling schemes: a binomial versus an exponential prior, and hard versus soft

information coupling. This comparison has revealed that the binomial prior allows for more consistent inference of the right level of information sharing, while the exponential prior tends to enforce overly-strong information sharing. The difference between hard and soft information coupling seems negligible in the scenarios we investigated. A detailed investigation of the hyperparameter inference has allowed us to improve the MCMC sampler for better convergence, and to explore the limitations of the exponential information sharing prior.

The application of our method to gene expression time series taken during the life cycle of *Drosophila melanogaster* has revealed better agreement with known morphogenic transitions than the methods of Robinson and Hartemink (2009, 2010) and Ahmed and Xing (2009), and we have been able to identify several gene and protein interactions that are known from the literature. In an application to data from a topical study in synthetic biology (Cantone et al. 2009), our methods have outperformed two established network reconstruction methods from computational systems biology, and information sharing has allowed us to reconstruct the true underlying gene network with higher overall precision and recall than would have been possible without it.

We have investigated the performance of our methods on datasets which arise from gene regulatory networks with temporal changes in the structure of the network. There are several special cases of this situation which merit further discussion. The simplest case occurs when the changes of the underlying process are limited to parameter changes, and the true structure of the network remains constant. We have shown in Sect. 5.1 that our methods can deal with this situation effectively thanks to information sharing among segments. A more complicated case could involve a reoccurring event that causes certain gene interactions to switch on or off, leading to repeated network structures. For example, in a circadian clock system such as Locke et al. (2006), Pokhilko et al. (2010), the absence of sunlight might deactivate the interaction between two genes in the network, causing its structure to change from A to B.[10] If gene expression levels are measured both during the day and at night for three days, then we will observe a sequence like ABABAB. While our methods can in principle represent repeated segments, the multiple changepoint process was not designed with this in mind. A better model for repeated segments might be a Hidden Markov Model (HMM), where each hidden state corresponds to a network structure, and transitions between states correspond to changes in the structure, in the same vein as applied to changing tree structures in phylogeny (Husmeier and McGuire 2003). The disadvantage of using HMMs is that they impose a geometric distribution on the segment lengths, and in that respect our changepoint process is more flexible. To have the same flexibility with HMMs, model extensions along the lines of hierarchical HMMs or HMMs with weighting times could be pursued, as known from speech processing, but this would come at significantly increased computational costs. Hence, this approach only appears to make sense if there is strong prior indication that repetitions occur.

An interesting topic for future work is to investigate other functional forms of the information sharing mechanism. In our work, we have investigated four different models, based on an exponential versus binomial distribution, with or without gene-specific hyperparameters. It has recently come to our attention that Wang et al. (2011) have experimented with a different approach, which effectively combines our exponential prior with an additional factor that encourages network sparsity. Sparsity in our model is encouraged by the truncated Poisson prior of Eq. (4), as explained in the paragraph under Eq. (30). It would be

[10] Note that our definition of a deactivated gene interaction includes interactions that no longer occur because one of the interacting genes is no longer expressed.

interesting to explore the effect of the additional factor used in Eq. (7) of Wang et al. (2011) in the context of gene network reconstruction.

Reconstructing gene regulatory networks from transcriptional profiles remains a challenging problem, which a flurry of ongoing methodological developments in the computational systems biology community are trying to address. We believe that our paper adds a valuable contribution to this field, by presenting a consistent and flexible Bayesian model for the case where the network structures change over time.

Acknowledgements Most of the work was carried out while Dirk Husmeier was employed at Biomathematics and Statistics Scotland, and the work was supported by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS). This work was partly funded by EU FP7 grant "Timet". Frank Dondelinger's PhD research is partly funded by the Engineering and Physical Sciences Research Council (EPSRC).

References

Ahmed, A., & Xing, E. P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, *106*, 11878–11883.

Andrianantoandro, E., Basu, S., Karig, D., & Weiss, R. (2006). Synthetic biology: new engineering rules for an emerging discipline. *Molecular Systems Biology*, *2*(1), E1–E14.

Andrieu, C., & Doucet, A. (1999). Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Transactions on Signal Processing*, *47*(10), 2667–2676.

Arbeitman, M., Furlong, E., Imam, F., Johnson, E., Null, B., Baker, B., Krasnow, M., Scott, M., Davis, R., & White, K. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, *297*(5590), 2270–2275.

Cantone, I., Marucci, L., Iorio, F., Ricci, M.A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., & Cosma, M. P. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, *137*(1), 172–181.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on machine learning* (p. 240). New York: ACM.

Dondelinger, F. (2012). *A machine learning approach to reconstructing signalling pathways and interaction networks in biology*. PhD thesis, University of Edinburgh (in preparation).

Dondelinger, F., Lebre, S., & Husmeier, D. (2010). Heterogeneous continuous dynamic Bayesian networks with flexible structure and inter-time segment information sharing. In *Proceedings of the 27th international conference on machine learning (ICML)*.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., et al. (2005). Protein interaction mapping: a Drosophila case study. *Genome Research*, *15*(3), 376.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Grzegorczyk, M., & Husmeier, D. (2009). Non-stationary continuous dynamic Bayesian networks. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems (NIPS)* (Vol. 22, pp. 682–690).

Grzegorczyk, M., & Husmeier, D. (2011). Non-homogeneous dynamic Bayesian networks for continuous data. *Machine Learning*, *83*, 355–419.

Guo, F., Hanneke, S., Fu, W., & Xing, E. (2007). Recovering temporally rewiring networks: a model-based approach. In *Proceedings of the 24th international conference on machine learning* (p. 328). New York: ACM.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Homyk, T. Jr, & Emerson, C. Jr (1988). Functional interactions between unlinked muscle genes within haploinsufficient regions of the Drosophila genome. *Genetics*, *119*(1), 105.

Husmeier, D., & McGuire, G. (2003). Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution*, *20*(3), 315–337.

Husmeier, D., Dondelinger, F., & Lèbre, S. (2010). Inter-time segment information sharing for non-homogeneous dynamic Bayesian networks. In J. Lafferty (Ed.), *Proceedings of the twenty-fourth annual conference on neural information processing systems (NIPS)* (Vol. 23, pp. 901–909). New York: Curran Associates.

Kolar, M., Song, L., & Xing, E. (2009). Sparsistent learning of varying-coefficient models with structural changes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems (NIPS)* (Vol. 22, pp. 1006–1014).

Larget, B., & Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, *16*(6), 750–759.

Lèbre, S. (2007). *Stochastic process analysis for genomics and dynamic Bayesian networks inference*. PhD thesis, Université d'Evry-Val-d'Essonne, France.

Lèbre, S., Becq, J., Devaux, F., Lelandais, G., & Stumpf, M. (2010). Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, *4*, 130.

Locke, J., Kozma-Bognár, L., Gould, P., Fehér, B., Kevei, E., Nagy, F., Turner, M., Hall, A., & Millar, A. (2006). Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Molecular Systems Biology*, *2*(1), 59.

Montana, E., & Littleton, J. (2004). Characterization of a hypercontraction-induced myopathy in Drosophila caused by mutations in *mhc*. *The Journal of Cell Biology*, *164*(7), 1045.

Nongthomba, U., Cummins, M., Clark, S., Vigoreaux, J., & Sparrow, J. (2003). Suppression of muscle hypercontraction by mutations in the myosin heavy chain gene of *Drosophila melanogaster*. *Genetics*, *164*(1), 209.

Parkhurst, S., & Ish-Horowicz, D. (1991). *WIMP*, a dominant maternal-effect mutation, reduces transcription of a specific subset of segmentation genes in Drosophila. *Genes & Development*, *5*(3), 341.

Pokhilko, A., Hodge, S., Stratford, K., Knox, K., Edwards, K., Thomson, A., Mizuno, T., & Millar, A. (2010). Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular Systems Biology*, *6*(1), 416.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., & Stolovitzky, G. (2010). Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE*, *5*(2), e9202.

Punskaya, E., Andrieu, C., Doucet, A., & Fitzgerald, W. (2002). Bayesian curve fitting using MCMC with applications to signal segmentation. *IEEE Transactions on Signal Processing*, *50*(3), 747–758.

Robinson, J. W., & Hartemink, A. J. (2009). Non-stationary dynamic Bayesian networks. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems (NIPS)* (Vol. 21, pp. 1369–1376). San Mateo: Morgan Kaufmann.

Robinson, J., & Hartemink, A. (2010). Learning non-stationary dynamic Bayesian networks. *Journal of Machine Learning Research*, *11*, 3647–3680.

Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Roeder, L., Euzenat, J., Rechenmann, F., & Jacq, B. (1999). Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an internet database. *Nucleic Acids Research*, *27*(1), 89.

Sims, D., Bursteinas, B., Gao, Q., Zvelebil, M., & Baum, B. (2006). FLIGHT: database and tools for the integration and cross-correlation of large-scale RNAi phenotypic datasets. *Nucleic Acids Research*, *34*(suppl 1), D479.

Talih, M., & Hengartner, N. (2005). Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society B*, *67*(3), 321–341.

Wang, Z., Kuruoglu, E., Yang, X., Xu, Y., & Huang, T. (2011). Time varying dynamic Bayesian network for non-stationary events modeling and online inference. *IEEE Transactions on Signal Processing*, *4*(59), 1553.

Werhli, A. V., & Husmeier, D. (2008). Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions. *Journal of Bioinformatics and Computational Biology*, *6*(3), 543–572.

Xuan, X., & Murphy, K. (2007). Modeling changing dependency structure in multivariate time series. In Z. Ghahramani (Ed.), *Proceedings of the 24th annual international conference on machine learning (ICML 2007)* (pp. 1055–1062). New York: Omnipress.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques* (pp. 233–243). Amsterdam: Elsevier.

Zhao, W., Serpedin, E., & Dougherty, E. (2006). Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics*, *22*(17), 2129.