Modélisation de l'expression des gènes à partir de données de séquence ADN

May Taha[1,2,3*], Chloé Bessière[1,2*], Florent Petitprez[1,2], Jimmy vandel[1,4], Jean-Michel Marin[1,3], Laurent Bréhélin[1,4#], Sophie Lèbre[1,3#] Charles-Henri Lecellier[1,2#]

[1]Institut de Biologie Computationnelle
[2]Institut de Génétique Moléculaire de Montpellier
[3]Institut Montpelliérain Alexander Grothendieck
[4]Laboratoire d'informatique, de robotique et de microélectronique de Montpellier

[*,#] These authors contributed equally to this work

Résumé

L'expression des gènes est étroitement contrôlée pour assurer une grande variété de fonctions et de types cellulaires. Le développement des maladies, en particulier les cancers, est invariablement lié à la dérégulation de ces contrôles. Notre objectif est de modéliser le lien entre l'expression des gènes et la composition nucléotidique des différentes régions régulatrices du génome. Nous proposons d'aborder ce problème dans un cadre de régression avec une approche Lasso couplée à un arbre de régression. Nous utilisons exclusivement des données de séquences et nous apprenons un modèle différent pour chaque type cellulaire. Nous montrons (i) que les différentes régions régulatrices apportent des informations différentes et complémentaires et (ii) que la seule information de leur composition nucléotidique permet de prédire l'expression des gènes avec une erreur comparable à celle obtenue en utilisant des données expérimentales. En outre, le modèle linéaire appris n'est pas aussi performant pour tous les gènes, mais modélise mieux certaines classes de gènes avec des compositions nucléotidiques particulières.

Mots-clés Régression, Lasso, Arbre de régression, Régulation de l'expression des gènes, Cancer

Abstract

Gene expression is tightly controlled to ensure a wide variety of cell types and functions. The development of diseases, particularly cancers, is invariably related to deregulations of these controls. Our objectif is to model the link between gene expression and nucleotide composition of different regulatory regions in the genome. We propose to address this problem in a regression framework using a Lasso approach coupled to a regression tree. We use exclusively sequence data and we fit a different model for each cell type. We show that (i) different regulatory regions provide particular and complementary informations and that (ii) the only information contained in the nucleotide compositions allows to predict gene expression with an error comparable to that obtained using experimental data. Moreover, the fitted linear model is not as powerful for all genes, but better fits certain groups of genes with particular nucleotides compositions.

Keywords Regression, Lasso, Regression trees, Regulation of gene expression, Cancer

1 Introduction

L'expression des gènes est étroitement contrôlée pour assurer une grande variété de fonctions cellulaires. Ces contrôles ont lieu à plusieurs niveaux (transcriptomique, post-transcriptomique, ...) et sont associés à différentes régions génomiques : promoteurs proches et distants, amplificateurs, exons, introns, ... Alors que ces régions génomiques sont les mêmes dans les différentes cellules et dans les différents types cellulaires, la diversité de leurs réponses est assurée par la variété des facteurs qui viennent se lier à ces différentes régions. Les facteurs de transcription par exemple, viennent se fixer sur les régions régulatrices de l'ADN (promoteurs, UTR, introns, ...) et jouent le rôle d'activateurs ou d'inhibiteurs de la transcription des gènes qu'ils contrôlent. La régulation de l'expression des gènes est ainsi le résultat d'une combinaison complexe et précise de régulateurs, qui reconnaissent des séquences ADN particulières dans des régions de régulation spécifiques. Un défi actuel est de comprendre comment la machinerie de régulation des gènes est orchestrée dans chaque type cellulaire, et d'identifier les facteurs et les régions régulatrices les plus importants.

Nous proposons d'aborder ce problème dans un cadre de régression avec une approche Lasso. Nous disposons de mesures d'expression des gènes dans un type cellulaire, ainsi que d'un ensemble de variables descriptives de la composition nucléotidique dans les régions régulatrices de ces gènes. Nous voulons alors apprendre un modèle qui puisse prédire l'expression d'un gène à partir des variables descriptives de ce gène. Le but est d'identifier par ce biais les variables importantes pour la régulation de l'expression dans le type cellulaire étudié. Dernièrement, plusieurs approches ont été proposées dans un cadre similaire [2]. Bien que ces approches soient relativement efficaces pour prédire l'expression des gènes, elles sont basées sur des données expérimentales (ChIP-seq, methylation, DNase hypersensitivity) qui sont limitées à des échantillons spécifiques et qui restent souvent difficiles à obtenir. L'originalité de notre approche est que nous utilisons exclusivement des données de séquences (a priori invariables pour les différents types cellulaires) et que nous apprenons un modèle différent pour chaque type cellulaire.

2 Méthodologie

2.1 Modèle

Des données d'expression de gènes ont été collectées sur sur la base de données The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/) pour 12 types de cancers différents sous la forme de quantités de transcrits mesurées par RNA-seq (20 tumeurs par type de cancer, soit 240 tumeurs au total). Afin d'intégrer différents modes de régulation des gènes, notamment les régulations transcriptionnelles et post-transcriptionnelles, nous avons défini pour chaque gène, 8 régions régulatrices, réparties sur le promoteur du gène (initiateur de la transcription) et le corps du gène (exons et introns). Chaque région est une sous-séquence d'ADN spécifique pour chaque gène dont nous mesurons les taux de nucléotides et de di-nucléotides. Certaines régions régulatrices sont définies par des annotations spécifique à chaque gène et leurs longueurs varient d'un gène à un autre. Le fait d'utiliser des taux (nombre de nucléotide/longueur) permet de limiter l'influence de ces variations. Dans cette étude, nous avons considéré 16294 gènes pour lesquels nous disposons de l'ensemble des données définissant ces 8 régions régulatrices. L'expression de l'ensemble

des gènes d'une tumeur est expliquée par le modèle linéaire suivant :

$$Y = X\beta + \epsilon \tag{1}$$

où $Y = (y_1, ..., y_n)$ est le vecteur d'expression de l'ensemble des n gènes, $X = (X_1, ...X_p)$ est la matrice des compositions nucléotidiques, telle que chaque colonne X_i contient le taux d'un nucléotide ou di-nucléotide au sein de l'une des 8 régions régulatrices pour l'ensemble des n gènes, $\beta = (\beta_1, ...\beta_p)$ est le vecteur des coefficients à estimer et $\epsilon = (\epsilon_1, ..., \epsilon_n)$ le vecteur des erreurs supposées indépendantes et identiquement distribuées. Un modèle est donc estimé spécifiquement pour chaque condition.

Notre étude repose sur un grand nombre de variables explicatives (160 variables) et une méthode de sélection de variables pertinente est nécessaire. Nous avons utilisé pour cela la régression linéaire pénalisée par la norme ℓ_1 ou Lasso (Least Absolute Shrinkage and Selection Operator par Tibshirani) [5]. L'avantage de cette méthode réside dans l'utilisation de la norme ℓ_1 qui, par sa géométrie, conduit naturellement à favoriser les coefficients nuls. Les estimations de β sont obtenues en minimisant l'erreur pénalisée :

$$\widehat{\beta}_{LASSO} = \underset{\beta}{\text{argmin}} \left(||Y - X\beta||^2 + \lambda \sum_{j=0}^{p} |\beta_j| \right) \tag{2}$$

Le choix du poids λ de la pénalité est crucial. Une grande valeur de λ impose une pénalisation forte, et ainsi la sélection d'un petit nombre de variables, alors qu'une valeur de λ égale à 0 se ramène à une régression linéaire ordinaire. La valeur de λ est estimée par validation croisée, de façon à minimiser l'erreur de prédiction sur les données de test.

2.2 Évaluation des prédictions

Deux critères ont été utilisés afin d'évaluer les performances de prédiction : l'erreur quadratique moyenne et la corrélation de Spearman entre les valeurs observées et les valeurs prédites. Un premier résultat notable est que les performances de notre modèle utilisant uniquement la composition de la séquence ADN (corrélation médiane = 0.58) sont comparables à celles obtenues à partir de données expérimentales mesurant la quantité de facteurs de transcription fixés dans la région promotrice de chaque gène dans une approche récente [2] (corrélation médiane= 0.57).

2.3 Caractérisation des gènes bien prédits

Suivant les tumeurs, on observe que le modèle linéaire appris n'est pas aussi performant pour tous les gènes, mais modélise mieux certaines classes de gènes avec des compositions nucléotidiques particulières. Pour chaque tumeur, nous avons donc construit un arbre de régression (CART pour Classificatin and Regression Tree [1]) qui prédit l'erreur du modèle linéaire à partir des compositions nucléotidiques. Cela nous permet de caractériser les groupes de gènes bien ou mal prédits par le modèle linéaire en fonction des compositions nucléotidiques de leurs régions régulatrices.

2.4 Stabilité de sélection des variables

Le Lasso n'est pas stable et les variables sélectionnées changent en fonction de l'échantillon d'apprentissage. Afin d'identifier les sous-groupes de variables sélectionnées de façon stable pour chaque tumeur, nous avons utilisé l'approche proposée et étendue par Meinshausen et al. [4] [3]. Cette stratégie consiste à répéter l'estimation un grand nombre de fois, en utilisant à chaque répétition, seulement une partie des individus et en pondérant les variables explicatives par des poids uniformément distribués entre 0.5 et 1. Après avoir répété l'estimation 500 fois, nous avons conservé les variables sélectionnées sur plus de 70% des répétitions. Nous avons ainsi mis en évidence des compositions nucléotidiques sélectionnées de façon stable quel que soit le type de cancer, alors que d'autres sont spécifiques à certains types de cancer.

3 Conclusion et perspectives

Nous proposons une méthodologie pour identifier les régions importantes pour la régulation de l'expression dans un type cellulaire donné. Nous montrons que ces différentes régions apportent des informations différentes et complémentaires, et que la seule information de leur composition nucléotidique permet de prédire l'expression des gènes avec une erreur comparable à celle obtenue en utilisant des données expérimentales.

En analysant plus finement le modèle linéaire à l'aide d'un arbre de classification, nous avons pu identifier des classes de gènes qui sont bien ou mal modélisées. On identifie ainsi deux types de gènes. Les gènes pour lesquels nous avons vraisemblablement identifiés les principaux déterminants de la régulation, et les autres gènes, pour lesquels les variables génomiques impliquées dans leur régulation ne sont pas connues ou sont mal prises en compte par le modèle linéaire. Nous montrons en outre que ces groupes diffèrent sensiblement d'un type cellulaire à l'autre, et qu'ils sont associés à la structure 3D de la chromatine (TADs, pour Topologically Associating Domains).

Il y a plusieurs perspectives à ce travail. Une direction est de s'affranchir de l'hypothèse de linéarité et de proposer un modèle plus complexe (par exemple non-paramétrique) à même de mieux modéliser les combinaisons liant les variables génomiques à l'expression. Une autre approche serait d'étendre le modèle afin d'intégrer certaines interactions de variables, au sein d'une même région ou dans des régions différentes.

4 Remerciements

Ces travaux ont été financés par l'école doctorale Sciences Chimiques et Biologiques pour la Santé (CBS2 ED168) en interface avec l'école doctorale Information, Structures, Systèmes (I2S ED166).

Références

[1] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[2] Yue Li, Minggao Liang, and Zhaolei Zhang. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput Biol*, 10(10) :e1003908, 2014.

[3] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 2010.

[4] Martin Sill, Thomas Hielscher, Natalia Becker, and Manuela Zucknick. c060 : Extended inference with lasso and elastic-net regularized cox and generalized linear models. *Journal of Statistical Software*, 2014.

[5] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[6] YX Rachel Wang and Haiyan Huang. Review on statistical methods for gene network reconstruction using expression data. *Journal of theoretical biology*, 362 :53–61, 2014.