Classements consensuels de données biologiques massives

Pierre Andrieu, Sarah Cohen-Boulakia, Alain Denise

Classements consensuels de données biologiques massives

Pierre Andrieu[1], Sarah Cohen-Boulakia[1] et Alain Denise[1,2]

[1] Université Paris-Saclay, LISN (Laboratoire Interdisciplinaire des Sciences du Numérique)
[2] Université Paris-Saclay, I2BC (Institut de Biologie Intégrative de la Cellule)

Introduction

Nous sommes depuis plus de deux décennies dans l'ère du *big data* dont l'émergence a été facilitée par d'importantes évolutions technologiques. La capacité de stockage en particulier a été considérablement augmentée, passant du mégaoctet dans les années 1990 au gigaoctet dans les années 2000 pour atteindre aujourd'hui le téraoctet. Une conséquence directe est la quantité de plus en plus démesurée de données qu'il est possible de rendre accessible publiquement. Parallèlement, en biologie, l'avènement du séquençage haut débit a entraîné une baisse significative du coût de séquençage et ainsi une nette augmentation du nombre de séquences déterminées (d'ADN ou d'ARN). À titre d'exemple, alors qu'il fallut dix ans et plusieurs milliards de dollars pour le premier séquençage du génome humain, il suffit désormais d'une semaine et de quelques milliers de dollars pour aboutir au même résultat. L'essor d'approches pluridisciplinaires a considérablement accéléré le traitement et l'analyse de ces données en impliquant notamment les mathématiques et l'informatique (modélisation, algorithmique, ...). Ainsi, les vingt dernières années ont vu une augmentation très importante des données associées aux séquences biologiques. Sont notamment concernées la biologie moléculaire (compréhension des mécanismes de réplication de l'ADN, de transcription d'ADN en ARNm, de traduction d'ARNm en protéines), la biologie structurale (structures d'ARN, structures protéiques, ...), la génétique (mutations, phylogénie, ...), la médecine (associations gènes-maladies, cibles thérapeutiques, ...). Regrouper et organiser ces données sont donc devenus un enjeu très important pour la mise à disposition et l'approfondissement des connaissances.

1 Accès aux données biologiques

1.1 Des bases de données nombreuses et hétérogènes

De nombreuses bases de données biologiques sont créées pour tenter de répondre à ce besoin croissant de regrouper et d'organiser les données. Le journal *Nucleic Acids Research* (NAR), qui en répertorie chaque année une part importante, comptabilise plus d'un millier de références dont 89 nouvelles pour la seule année

2021 [39]. Mais ces nouvelles données biologiques ne sont pas seulement très nombreuses, elles sont aussi de natures diverses à plusieurs niveaux [15]. Elles sont à la fois hétérogènes (séquences nucléiques et protéiques, interactions entre entités biologiques, lien entre gènes et maladies, conformation 3D d'ARN et protéines, variants génétiques...) et dépendantes de contextes biologiques (pathologie, espèce, étape du développement...). Pour répondre à cette problématique, les bases de données biologiques s'organisent en deux types : les bases généralistes constituées selon la nature des données à stocker et les bases de données spécialisées autour de thématiques biologiques. La modélisation des objets biologiques varie parfois beaucoup d'une base à l'autre, et les données sont associées à des degrés de qualité variables. Les bases de données "thématiques" garantissent souvent une plus grande fiabilité des données par rapport aux bases généralistes, grâce notamment à une vérification manuelle du contenu. La contrepartie est que la quantité d'information peut être insuffisante et qu'il faut parfois les compléter en utilisant d'autres bases.

1.2 Les portails

La diversité et la complémentarité des bases de données biologiques a de ce fait rendu indispensable l'utilisation de portails web permettant un accès centralisé à ces nombreuses données hétérogènes, au point que des organisations gouvernementales sont dédiées à cette tâche depuis une vingtaine d'années : le *National Center for Biotechnology Information* (NCBI) aux États-Unis d'Amérique, le *European Bioinformatics Institute* (EBI) en Europe et la *DNA Data Bank of Japan* (DDBJ) au Japon.

Ces portails tirent profit de la complémentarité des bases de données et de l'abondance des données en exploitant des références établies entre elles à deux niveaux. Premièrement, des références croisées sont établies entre entités de bases de données différentes. Ainsi, tel qu'illustré par l'extrait d'une fiche d'une base de données génétique en figure 1, la fiche d'un gène codant une protéine contient un lien vers l'entrée correspondant à cette protéine au sein d'une base de données protéique. Deuxièmement, les éléments d'une même base de données peuvent être connectés entre eux grâce à des algorithmes. Par exemple, en consultant une séquence nucléotidique, on a la possibilité de retrouver celles qui lui "ressemblent". De la même manière, à partir d'un article, on peut trouver d'autres articles provenant de la même base de données portant sur le "même" sujet grâce à une étude sémantique des entrées (mots en commun, fréquence et proximité de ces mots, ...). Lorsqu'un utilisateur interroge une base de données avec un mot-clé, une analyse de texte est également à l'oeuvre : le mot-clé est recherché dans les différentes entrées de la base permettant un tri par pertinence. Par exemple, interroger la base de données *Gene* du NCBI [31] avec pour mot-clé un nom de maladie permet d'établir des associations gènes-maladies.

2 Utilisation de mots-clés synonymes

Avec la recherche de mots clés au sein des fiches apparaît une problématique récurrente en biologie : comment faire lorsque des travaux de recherche utilisent

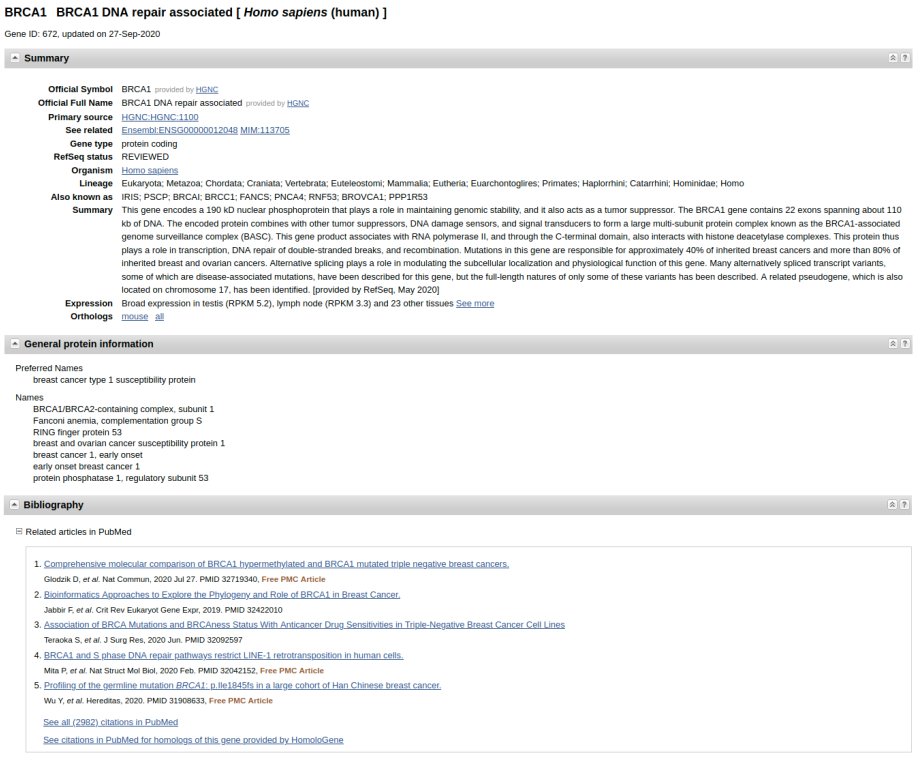

Fig. 1. Extrait de la fiche d'annotation du gène *BRCA1* au sein de la base de données *Gene* du NCBI.

des dénominations différentes pour désigner un même gène ou une même maladie? Une conséquence de ce problème est qu'une analyse sémantique au sein d'une base de données de gènes peut aboutir à des résultats très différents en fonction de la dénomination choisie pour lancer la recherche.

2.1 Illustration avec le cancer du sein

Prenons l'exemple de médecins qui s'interrogent sur les gènes les plus impliqués dans le cancer du sein. Pour cela, ils vont interroger la base de données *Gene* du *NCBI* avec le mot-clé *Breast cancer* (les premiers résultats de cette requête sont visibles en Figure 2). La base de données renvoie une liste de 19567 gènes dont 4433 pour l'espèce humaine classés par ordre de pertinence. Nous pouvons voir sur la Figure 2 que le gène *BRCA2* apparaît en première position, le gène *BRCA1* en deuxième position et ainsi de suite. La pertinence d'un gène donné vis-à-vis du mot-clé (*Breast Cancer* pour notre exemple) est déterminée par le *NCBI* en fonction du nombre de fois que le mot-clé est présent dans la fiche du gène. Il existe un cas de figure particulier : des gènes peuvent partager une

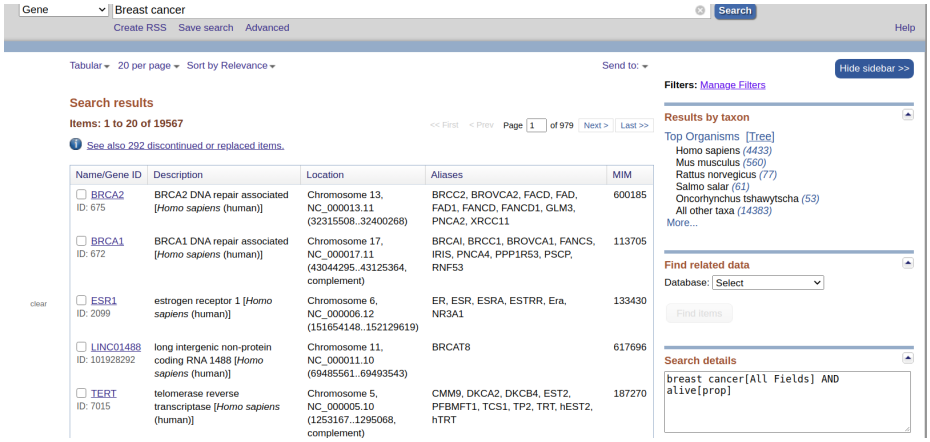

Fig. 2. Résultat de la requête "Breast Cancer" en interrogeant la base de données *Gene* du NCBI.

même position dans un classement, ils sont alors considérés ex-aequo en termes de pertinence.

Malheureusement ce système est biaisé en raison du problème de nomenclature soulevé plus haut. Ainsi, en pratique, il est fréquent que certains gènes très pertinents et donc théoriquement attendus en tête de liste se retrouvent mal classés voire absents du classement retourné. Afin de pallier ce problème, les médecins interrogent la base de données à plusieurs reprises en utilisant différentes dénominations synonymes de la maladie qu'on appelle *reformulations*. Ainsi, le gène *MUC1* qui apparaît en positon 170 en utilisant le mot-clé *Breast cancer* se trouve dans les dix premiers lorsqu'on utilise la reformulation *Breast carcinoma* ou encore *Malignant neoplasm of breast*. Il est important de noter que les différents classements de gènes obtenus ne contiennent pas le même nombre de gènes (6129 pour *Breast carcinoma* dont 3321 pour l'espèce humaine par exemple) et que les gènes contenus dans les "plus petits" classements ne sont pas nécessairement inclus dans l'ensemble des gènes contenus dans les "plus grands classements". En revenant à l'exemple du gène *MUC1*, il est absent des résultats lorsqu'on utilise la reformulation *Cancer of the breast*.

2.2 Plusieurs classements à combiner

Le problème est que les maladies peuvent avoir de nombreuses reformulations, aboutissant donc à de nombreux classements de plusieurs centaines voire milliers de gènes différents. Il est alors complexe pour les médecins de savoir quels gènes étudier en priorité. Imaginons un cas de figure caricatural mais tout à fait possible pour illustrer la difficulté de considérer plusieurs classements : gène A est devant gène B dans une majorité de classements, B est avant C dans une majorité de classements, et C est avant A dans une majorité de classements. Choisir un gène parmi ces trois pour une étude expérimentale s'avère être un

problème complexe. L'agrégation de classements apporte des réponses à de tels problèmes.

3 Méthodes d'agrégation de classements

L'agrégation de classements consiste à partir de plusieurs classements en entrée et à calculer un classement dit consensuel censé représenter au mieux les points communs entre les classements d'entrée. Cette problématique est étudiée depuis des siècles. Un manuscrit écrit par le majorquin Ramon Llull daté de la fin du XIIIe siècle et intitulé *'De arte eleccionis'* a été récemment découvert [16]. Ce manuscrit présente certains concepts et méthodes d'agrégation de classements considérés comme majeurs aujourd'hui. Nous présentons maintenant la méthode Kemeny-Young, particulièrement adaptée à notre contexte biologique pour agréger les classements.

3.1 La méthode Kemeny-Young pour agréger les classements

Une méthode très utilisée dans le cadre de l'agrégation de classements est la *méthode Kemeny-Young* qui consiste à déterminer un classement aussi proche que possible des classements d'entrée, en minimisant une fonction de distance appelée *distance de Kendall-τ*. Un tel classement est appelé *classement consensuel optimal* ou *médiane*. Ce problème est connu pour être difficile si le nombre de classements est pair et supérieur ou égal à 4 [9, 22] ou impair et supérieur ou égal à 7 [7], ce qui signifie qu'un algorithme capable de trouver un classement consensuel optimal à coup sûr ne peut être que de complexité exponentielle en fonction de la taille des données.

En pratique, il peut falloir plusieurs heures à un tel calcul si le nombre d'éléments à classer est de l'ordre de la centaine. La complexité du problème pour les cas où le nombre de classements est égal à 3 ou 5 est encore inconnue et fait l'objet de travaux théoriques (par exemple, [34]).

Si l'agrégation de classements intéresse aujourd'hui plusieurs domaines dont l'algorithmique ([1,2,7–9]), les bases de données ([23–25]), la physique ([35,36]), la biologie et la bioinformatique ([14, 28, 29, 42]), elle a été initialement étudiée dans un contexte d'élections où les électeurs étaient invités à classer l'ensemble des candidats par ordre de préférence ([17,21]). Or, les classements obtenus dans ce contexte particulier d'élections sont très différents des classements obtenus en biologie. D'une part, dans le contexte d'élections en théorie du choix social, on considère que les classements sont complets (chaque votant doit trier l'ensemble des candidats) et sans égalités (un votant ne peut pas mettre deux candidats *ex-aequo*) [6,13,17,18,21]. En conséquence, la majorité des travaux théoriques ne s'appliquent qu'à ces classements particuliers [1, 2, 8, 9, 17–21, 27, 32–34] et sont donc inutilisables avec des classements incomplets et/ou avec égalités comme c'est le plus souvent le cas en biologie. D'autre part, dans le contexte d'élections, on a généralement beaucoup de classements et peu d'éléments dans chaque

classement (beaucoup de votants et peu de candidats). En biologie le problème est généralement orthogonal : on se retrouve avec quelques dizaines de classements dont chacun peut contenir plusieurs milliers d'entités biologiques (gènes, protéines, ...). Or, le paramètre à l'origine de la difficulté du problème d'agrégation de classements est le nombre d'éléments à trier. Ainsi, l'utilisation d'algorithmes exacts sur des données biologiques est compromise, ces derniers ne permettant pas de calculer un classement consensuel optimal en un temps raisonnable s'il y a plus de quelques dizaines d'éléments à trier.

Premier enjeu. Pour répondre à ce problème, il est donc fondamental de concevoir des algorithmes que l'on appelle des heuristiques, c'est-à-dire des algorithmes qui ne renvoient pas toujours la solution exacte mais qui sont capables de calculer un classement consensuel de bonne qualité en un temps raisonnable.

3.2 Plusieurs classements consensuels optimaux possibles

Pour rappel, l'agrégation de classements dans le contexte de la méthode Kemeny-Young consiste à calculer un classement aussi proche que possible des classements de départ en minimisant une fonction de distance. Le problème est que dans certains jeux de données, un grand nombre de classements parfois très différents peuvent minimiser cette distance. Dans un tel cas, la position des éléments dans le classement consensuel n'est pas robuste, même si ce dernier est optimal.

Deuxième enjeu. Un enjeu très important dans le cadre des données réelles est ainsi de fournir à l'utilisateur des indicateurs sur la robustesse du classement consensuel renvoyé afin qu'il soit averti du degré de confiance qu'il peut avoir vis-à-vis de ce classement.

4 Données réelles et classements incomplets

Se confronter à des données réelles implique également de se poser des questions qualitatives sur ces données. En particulier, dans le contexte d'agrégation de classements, il est légitime de se demander comment les éléments manquants dans les classements doivent être interprétés vis-à-vis de ceux présents.

Pour illustrer cette interrogation, reprenons le cas de l'agrégation de classements de gènes issus des reformulations d'une maladie puis considérons un contexte biologique différent.

4.1 Premier cas d'utilisation : reformulations synonymes de maladies

Pour rappel, les classements issus des reformulations peuvent être incomplets sans pour autant que l'ensemble des gènes contenus dans les "plus petits classements" soient inclus dans l'ensemble des gènes contenus dans les "plus grands classements". En utilisant la reformulation *Cancer of the breast*, le gène *MUC1*

est absent de la liste retournée alors que le gène *MDM2* est présent. On peut en déduire que le mot-clé *Cancer of the breast* n'existe pas dans la fiche de *MUC1* alors qu'il est présent dans la fiche de *MDM2*. Il paraît alors raisonnable de considérer que, vis-à-vis du mot-clé *Cancer of the breast*, *MDM2* doive être considéré comme plus pertinent que *MUC1*. On préférera donc pénaliser un gène absent d'un classement, afin de ne pas mettre en avant des gènes sans aucune pertinence réelle. On retrouve cette interprétation des données manquantes dans d'autres contextes, notamment dans certains systèmes électoraux modernes où les électeurs peuvent classer jusqu'à trois candidats par ordre de préférence (c'est le cas des élections présidentielles au Sri Lanka par exemple). Dans ce contexte, il est raisonnable de considérer qu'un électeur préfère les candidats classés que les candidats non classés. Les candidats absents d'un bulletin doivent donc être pénalisés vis-à-vis des candidats présents sur le bulletin, au risque sinon de faire élire un candidat ne reflétant pas du tout les préférences des électeurs.

4.2 Deuxième cas d'utilisation : les classements issus d'expériences multi-omiques

Intéressons nous maintenant à un second cas d'utilisation possible en biologie : l'agrégation de données issues d'expériences multi-omiques. Ces expériences permettent d'observer des entités biologiques de natures différentes (gènes, ARN et protéines) mais complémentaires. Les expériences impliquant les protéines aboutissent à des classements potentiellement incomplets puisque des protocoles expérimentaux excluent certaines protéines de la mesure. Ce cas d'utilisation est ainsi très différent du précédent. En effet, dans le cas des expériences multi-omiques, les absences sont dues à un biais de protocole et ne doivent surtout pas être interprétées comme un signe de non pertinence : pénaliser l'absence peut avoir pour conséquence de passer à côté de protéines particulièrement importantes pour le phénomène biologique étudié. Les éléments présents et les éléments absents sont incomparables. Cette interprétation des données manquantes est retrouvée dans d'autres contextes, par exemple dans celui des plateformes de films que les utilisateurs peuvent noter. Même le plus cinéphile des utilisateurs n'aura vu qu'une petite proportion des films à disposition dans la base de données. Si un film n'a pas été noté par un utilisateur, on ne peut pas en conclure que cet utilisateur l'a trouvé pire que le film qu'il a le plus mal noté. Pénaliser les éléments absents reviendrait à favoriser les films les plus regardés au lieu de favoriser les films les plus appréciés. De la même manière, dans un contexte universitaire où des étudiants choisissent des matières optionnelles, on ne peut pas considérer que les étudiants qui n'ont pas choisi une option ont moins bien réussi l'examen que les étudiants inscrits à l'option et donc présents à l'examen.

Troisième enjeu. Le troisième enjeu est de permettre une prise en compte du contexte dans lesquels les classements sont obtenus afin d'agréger les classements incomplets de façon pertinente.

5 Résumé des contributions

Le projet hyQualiBio regroupe des experts de différentes communautés afin de répondre aux enjeux évoqués dans les sous-sections 3.1, 3.2 et la section 4. Ainsi, les travaux présentés ci-dessous sont issus de collaborations entre des biologistes (experts entre autres des données multi-omiques), des médecins (experts dans l'étude des associations gènes-maladies), ainsi que des membres de plusieurs communautés informatiques : classement et gestion des données (bases de données), techniques et approches pour des calculs efficaces d'agrégation (algorithmique combinatoire), identification de "bonnes propriétés" des approches d'agrégation de classements (représentation des connaissances) pour aider au choix d'un algorithme adapté à la gestion pertinente des données manquantes.

Cette collaboration a permis de développer des méthodes permettant de répondre aux enjeux cités précédemment. Ces méthodes ont été implémentées dans deux outils en ligne mis à la disposition de la communauté scientifique. Le premier, ConQuR-Bio, apporte une réponse aux deux premiers enjeux présentés. Le deuxième, CoRankCo, apporte une réponse au troisième enjeu présenté.

5.1 ConQuR-Bio : outil en ligne à destination des biologistes

Le but de ConQuR-Bio est de fournir une liste de gènes associés à une maladie (exprimée sous la forme de mot-clé par les utilisateurs). ConQuR-Bio exploite le fait que les noms de maladies peuvent avoir différents synonymes. Chaque synonyme est associé à une liste différente de gènes et est récupérée par ConQuR-Bio afin de fournir aux utilisateurs un classement consensuel des gènes d'intérêt pour la maladie considérée.

5.1.1 Architecture du logiciel. L'architecture de ConQuR-Bio, décrite dans la figure 3, est composée de trois modules principaux.

Le premier module - **le module de reformulation** - prend comme entrée w, le mot-clé utilisateur (par exemple "breast cancer") et génère un ensemble de synonymes de w ($\{s_1, ..., s_m\}$). Les synonymes sont générés graĉe à une utilisation automatique de l'UMLS Metathesaurus [10] qui interroge les bases de données suivantes : MeSH (Medical Subject Headings) [30], SNOMED CT (SNOMED Clinical Terms) [43] CIM-9-CM et CIM-10-CM [37,38], OMIM (*Online Mendelian Inheritance in Man*) [26].

Le deuxième module - **le module d'interrogation** - prend le mot-clé fourni par l'utilisateur et la liste des synonymes fournis par le module de reformulation puis envoie une requête à la base de données *Gene* du NCBI par mot-clé et synonyme. Pour chaque mot-clé (et synonyme), *Gene* du NCBI fournit la liste de gènes associée. Dans la figure 3, $g1^w, ..., gn^w$ est la liste des gènes associée au mot-clé w. Chaque liste de gènes est classée par le NCBI par ordre décroissant du nombre d'occurrences du mot-clé (ou du synonyme) dans la fiche du gène. En conséquence, le module d'interrogation produit plusieurs listes triées de gènes (une par mot-clé et synonyme).

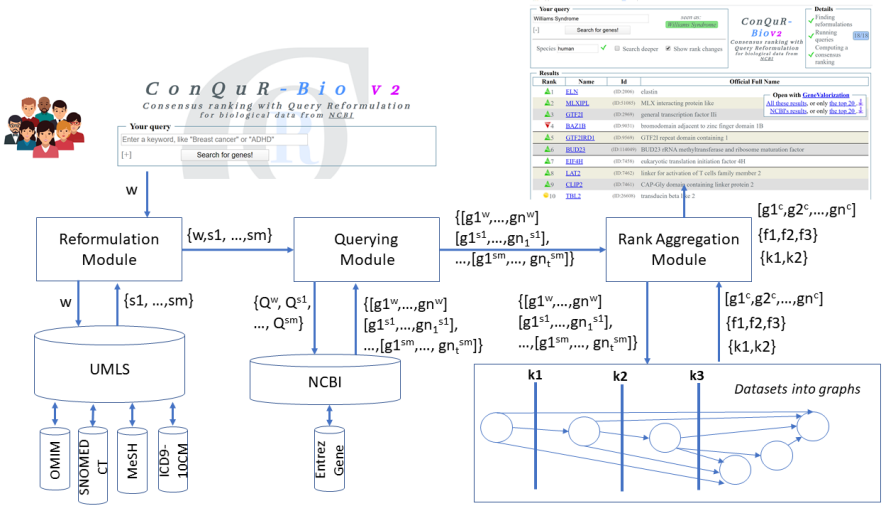

Fig. 3. Architecture de l'outil ConQuR-Bio

Le troisième module - **le module d'agrégation de classements** - est décrit dans la sous-section suivante.

5.1.2 Module d'agrégation de classements. Le module d'agrégation de classements prend en entrée les différents classements de gènes obtenues grâce aux deux premiers modules, et calcule un classement consensuel qui est renvoyé à l'utilisateur.

Une de nos contributions principales répond au premier enjeu présenté dans la sous-section 3.1 : nous avons conçu et implémenté dans ce troisième module une heuristique capable de gérer des données biologiques réelles (possiblement incomplètes et avec égalités). Une des étapes de cette heuristique consiste à diviser le problème initial en sous-problèmes indépendants, rendant possible l'utilisation d'algorithmes exacts [4, 5] (impossibles à utiliser initialement en raison d'un nombre souvent trop importants de gènes). Cette heuristique a fait l'objet d'une évaluation quantitative sur un grand nombre de données biologiques massives afin d'évaluer, entre autres, sa capacité à partitionner le problème initial en de nombreux sous-problèmes [4, 5] et d'une évaluation qualitative s'intéressant à la pertinence biologique des classements consensuels calculés [5].

5.1.3 Présentation du résultat à l'utilisateur. Le classement consensuel calculé par le module d'agrégation de classements est affiché à l'utilisateur. Pour rappel, pour un même jeu de données, il peut y avoir un grand nombre de classements consensuels optimaux, parfois très différents les uns des autres. L'utilisateur est donc intéressé à l'idée de savoir quelle est la robustesse du classement qui lui est présenté.

Notre deuxième contribution est de s'attaquer au problème de la non-unicité d'un classement consensuel optimal. Plus précisément, nous présentons un algorithme capable de calculer des *frontières* entre les gènes du classement consensuel : nous affichons une frontière en position k du classement consensuel si nous avons pu établir que l'ensemble des k premiers gènes est identique dans tous les classements consensuels optimaux.

La figure 4 présente la sortie utilisateur pour la maladie *Williams syndrome*. Les frontières correspondent aux lignes horizontales en gras.

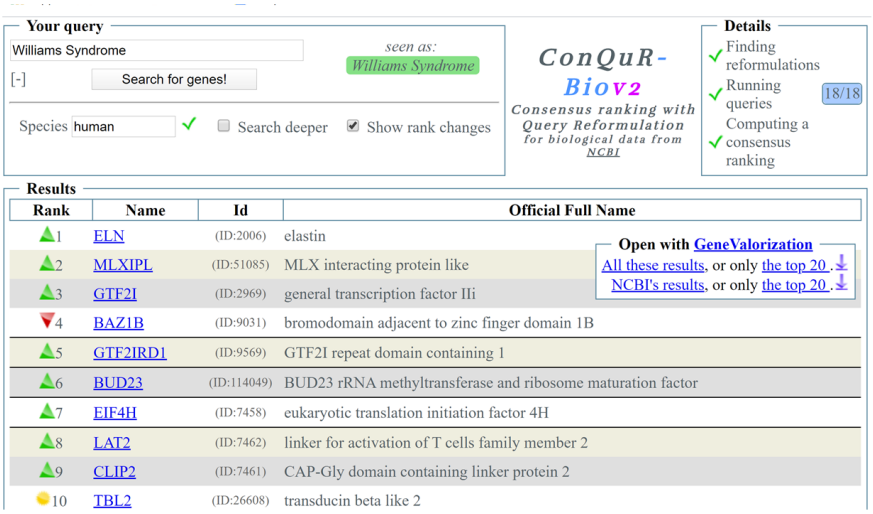

Fig. 4. Interface de ConQuR-BioV2 et classement consensuel obtenu pour le mot-clé (maladie) "Williams Syndrome".

Sur cette maladie, ConQuR-Bio fournit aux utilisateurs quatre frontières en position 4, 5, 6 et 7. Ces frontières indiquent que les gènes ELN, MLXIPL, GTF2I, BAZ1B doivent être considérés en priorité, puis GTF2IRD1, puis BUD23 et enfin EIF4H.

5.2 CoRankCo : outil en ligne pour l'agrégation de classements incomplets

Le deuxième outil mis à la disposition de la communauté est CoRankCo. Cet outil permet d'agréger des classements y compris dans le cas où les classements d'entrée sont incomplets (certains éléments sont absents de certains classements) et comportent des égalités (plusieurs éléments sont positionnés au même rang dans un même classement).

Ce logiciel est basé sur modèle présenté dans [3] dont les paramètres permettent d'intégrer un regard qualitatif sur les données, en permettant notamment

de choisir la façon dont les éléments manquants dans un classement doivent être pénalisés vis-à-vis des éléments présents. Ce modèle englobe les méthodes de l'état de l'art capable de gérer les classements incomplets. Nous avons adapté les algorithmes utilisés dans le cadre des classements complets pour intégrer les paramètres du modèle, ainsi que les méthodes correspondant aux deux premières contributions. Le modèle est évalué dans [3] sur des jeux de données réels de différentes natures ainsi que sur des jeux de données synthétiques.

La Figure 5 présente l'interface utilisateur de CoRankCo. Celle-ci peut être présentée en trois parties distinctes dont les deux principales sont détaillées ci-dessous.

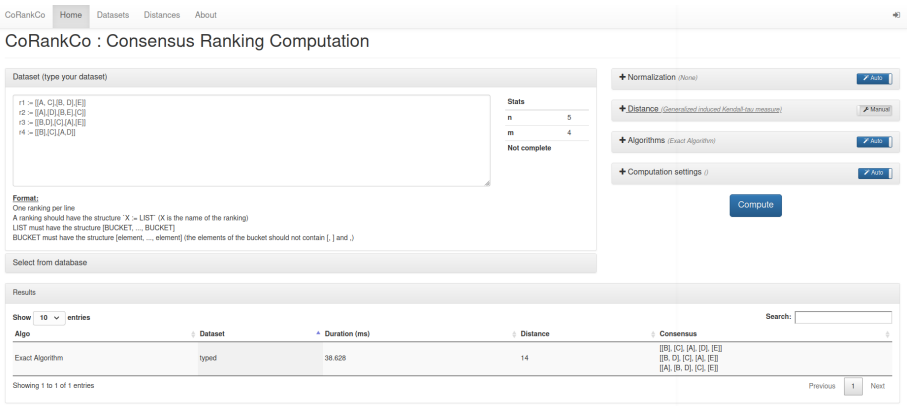

Fig. 5. Interface utilisateur de l'outil CoRankCo

5.2.1 Les jeux de données. Les utilisateurs peuvent utiliser les jeux de données présents dans la base (incluant de nombreux jeux de données biologiques ainsi que des jeux de données de votes de la base de données publique *preflib* https://www.preflib.org/) ou bien rentrer manuellement les classements à agréger dans un champ texte.

5.2.2 Les paramètres de calcul de classements consensuels. Le logiciel CoRankCo est décomposé en différents *panels* permettant de paramétrer le calcul de classements consensuels. Nous présentons ici le *panel normalisation* et le *panel distance* qui servent à la gestion des des données manquantes ainsi que le *panel algorithme* qui sert au choix de l'algorithme pour le calcul.

Gestion des données manquantes. Le *panel normalisation* et le *panel distance* visibles sur la Figure 5 permettent de gérer les données manquantes dans les classements.

Panel normalisation.
Deux normalisation sont proposées : (i) la projection [8] consiste à retirer des classements les éléments qui ne sont pas présents dans chaque classement et (ii) l'unification [40] consiste à ajouter à la fin de chaque classement incomplet l'ensemble des éléments manquants en dernière position. Ces deux normalisations pouvant induire certains biais, il est souvent préférable d'utiliser le *panel distance* décrit ci-dessous au lieu du *panel normalisation*.

Panel distance.
Nous rappelons que dans le cas où les classements sont complets et sans égalités, l'objectif est de calculer un classement qui minimise une fonction de distance (distance de Kendall-τ). Le *panel distance* permet de choisir l'adaptation de la distance Kendall-τ aux classements incomplets que l'on souhaite utiliser. Le choix de l'adaptation de la distance de Kendall-τ est étroitement lié au regard qualitatif que l'on porte sur les données manquantes. Par exemple, certaines distances permettent de considérer que les éléments manquants d'un classement sont moins pertinents vis-à-vis de ce classement que les éléments présents (cas d'utilisation 1 décrit en sous-section 4.1) et d'autres permettent de considérer que les éléments absents d'un classement sont incomparables avec les éléments présents (cas d'utilisation 2 décrit en sous-section 4.1).

Choix de l'algorithme. Le *panel algorithme* permet de sélectionner le ou les algorithmes qui vont être utilisés pour le calcul de classement(s) consensuel(s). Sont aujourd'hui disponibles les douze algorithmes suivants : BioCo [11], BioConsert [14], BordaCount [21], CopelandMethod [18], ExactAlgorithm [3], ExactAlgorithmPreprocessing [3], ParCons [5], KwikSort [2], MEDRank [25], Pick-a-Perm [2], Repeat Choice [1], SchulzeMethod [41]. Les algorithmes BioCo, BioConsert, BordaCount, CopelandMethod, KwikSort, MedRank, Pick-A-Perm et Repeat Choice sont décrits dans [12] et y ont été adaptés pour gérer les classements avec égalités. L'adaptation des 13 algorithmes aux données manquantes est explicitée dans [3].

5.2.3 Résultats du calcul de classement(s) consensuel(s). Pour chaque jeu de données sélectionné et chaque algorithme choisi, les classements consensuels sont calculés et affichés. Certains algorithmes ont la possibilité de renvoyer plusieurs classements consensuels, tous équivalents en terme de score de Kemeny. Pour chaque algorithme et chaque jeu de données sont affichés le score de Kemeny associé et le temps d'exécution de l'algorithme.

6 Conclusion

La recherche de classements consensuels est un problème très ancien qui apparaît dans de nombreux contextes. L'arrivée des données massives en biologie et en médecine est un contexte nouveau qui amène de nouveaux défis, non seulement

en termes de taille des données, mais aussi de propriétés de ces données, comme la gestion des éléments manquants.

Nous avons proposé de nouvelles solutions algorithmiques dans ce nouveau contexte, en proposant des heuristiques efficaces mais aussi un critère de robustesse des classements résultats. Dans le cas des données manquantes, nous avons aussi proposé un cadre général qui permet de paramétrer la recherche des classements consensus en fonction de la façon dont ces données manquantes doivent être interprétées. Ces résultats ont été implémentées dans deux logiciels librement diffusés à l'usage de la communauté scientifique.

Remerciements

References

1. Ailon, N. Aggregation of partial rankings, p-ratings and top-m lists. *Algorithmica 57*, 2 (Feb. 2010), 284–300.
2. Ailon, N., Charikar, M., and Newman, A. Aggregating inconsistent information: Ranking and clustering. *J. ACM 55*, 5 (Nov. 2008), 23:1–23:27.
3. Andrieu, P. *Passage à l'échelle, propriétés et qualité des algorithmes de classements consensuels pour les données biologiques massives.* Theses, Université Paris-Saclay, June 2021.
4. Andrieu, P., Brancotte, B., Bulteau, L., Cohen-Boulakia, S., Denise, A., Pierrot, A., and Vialette, S. Reliability-Aware and Graph-Based Approach for Rank Aggregation of Biological Data. In *2019 15th International Conference on eScience (eScience)* (San Diego, France, Sept. 2019), IEEE, pp. 136–145.
5. Andrieu, P., Brancotte, B., Bulteau, L., Cohen-Boulakia, S., Denise, A., Pierrot, A., and Vialette, S. Efficient, robust and effective rank aggregation for massive biological datasets. *Future Generation Computer Systems* (2021). Soumis après demande de révisions mineures par les reviewers.
6. Arrow, K., Sen, A., and Suzumura, K. *Handbook of Social Choice and Welfare*, vol. 1. Elsevier, 2002.
7. Bachmeier, G., Brandt, F., Geist, C., Harrenstein, P., Kardel, K., Peters, D., and Seedig, H. G. k-majority digraphs and the hardness of voting with a constant number of voters. *Journal of Computer and System Sciences 105* (2019), 130 – 157.
8. Betzler, N., Bredereck, R., and Niedermeier, R. Theoretical and empirical evaluation of data reduction for exact Kemeny rank aggregation. *Autonomous Agents and Multi-Agent Systems* (2013), 1–28.
9. Biedl, T., Brandenburg, F. J., and Deng, X. On the complexity of crossings in permutations. *Discrete Math. 309*, 7 (Apr. 2009), 1813–1823.
10. Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research 32* (01 2004), D267–D270.
11. Brancotte, B. *Rank aggregation with ties : algorithms, user guidance et applications to biologicals data.* Theses, Université Paris Sud - Paris XI, Sept. 2015.
12. Brancotte, B., Yang, B., Blin, G., Cohen Boulakia, S., Denise, A., and Hamel, S. Rank aggregation with ties: Experiments and analysis. *Proc. of the VLDB Endowment (PVLDB) 8*, 11 (Aug 2015), 2051.

13. Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. *Handbook of Computational Social Choice*, 1st ed. Cambridge University Press, USA, 2016.
14. Cohen-Boulakia, S., Denise, A., and Hamel, S. Using medians to generate consensus rankings for biological data. In *Scientific and Statistical Database Management* (07 2011), vol. 6809, pp. 73–90.
15. Cohen-Boulakia, S., and Valduriez, P. Interrogation et gestion de données bio-informatiques pour la biologie moléculaire. *Techniques de l'Ingenieur TIP140WEB* (Nov. 2015), BIO7055.
16. Colomer, J. Ramon llull: From ars electionis to social choice theory. *Social Choice and Welfare* (01 2012).
17. Condorcet, N. d. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Cambridge Library Collection - Mathematics. Cambridge University Press, 2014.
18. Copeland, A. H. A reasonable social welfare function, 1951. Seminar on Appl. of Mathematics to the social sciences, University of Michigan.
19. D'Ambrosio, A., Amodio, S., and Iorio, C. Two algorithms for finding optimal solutions of the kemeny rank aggregation problem for full rankings. *Electronic Journal of Applied Statistical Analysis 8* (2015), 198–213.
20. Davenport, A., and Kalagnanam, J. A computational study of the Kemeny rule for preference aggregation. In *Proceedings of the 19th National Conference on Artifical Intelligence* (2004), AAAI'04, AAAI Press, p. 697–702.
21. de Borda, J. C. *Mémoire sur les élections au scrutin*. Histoire de l'académie royale des sciences, 1781, pp. 657–664.
22. Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. Rank aggregation methods for the web. In *Proc. of the WWW Conference* (New York, NY, USA, 2001), WWW '01, ACM, pp. 613–622.
23. Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. Comparing and aggregating rankings with ties. In *Proc. of PODS* (2004), ACM, pp. 47–58.
24. Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. Comparing partial rankings. *SIAM J. Discret. Math. 20*, 3 (Mar. 2006), 628–648.
25. Fagin, R., Kumar, R., and Sivakumar, D. Efficient similarity search and classification via rank aggregation. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data* (2003), ACM, pp. 301–312.
26. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research 33* (01 2005), D514–D517.
27. Karpinski, M., and Schudy, W. Faster algorithms for feedback arc set tournament, kemeny rank aggregation and betweenness tournament. In *Algorithms and Computation* (Berlin, Heidelberg, 2010), O. Cheong, K.-Y. Chwa, and K. Park, Eds., Springer Berlin Heidelberg, pp. 3–14.
28. Kolde, R., Laur, S., Adler, P., and Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics (Oxford, England) 28* (02 2012), 573–80.
29. Li, X., Wang, X., and Xiao, G. A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications. *Briefings in Bioinformatics 20*, 1 (08 2017), 178–189.
30. Lipscomb, C. E. Medical subject headings (mesh). *Bulletin of the Medical Library Association 88* (July 2000), 265–271.

31. Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. Entrez gene: gene-centered information at NCBI. *Nucleic acids research 33* (2005), D54–D58.
32. Meilă, M., Phadnis, K., Patterson, A., and Bilmes, J. Consensus ranking under the exponential model. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence* (Arlington, Virginia, USA, 2007), UAI'07, AUAI Press, p. 285–294.
33. Milosz, R., and Hamel, S. Heuristic, branch-and-bound solver and improved space reduction for the median of permutations problem. In *Combinatorial Algorithms* (Cham, 2018), L. Brankovic, J. Ryan, and W. F. Smyth, Eds., Springer International Publishing, pp. 299–311.
34. Milosz, R., Hamel, S., and Pierrot, A. Median of 3 permutations, 3-cycles and 3-hitting set problem. In *Combinatorial Algorithms* (Cham, 2018), C. Iliopoulos, H. W. Leong, and W.-K. Sung, Eds., Springer International Publishing, pp. 224–236.
35. Muravyov, S., Baranov, P., and Emelyanova, E. How to transform all multiple solutions of the kemeny ranking problem into a single solution. *Journal of Physics: Conference Series 1379* (11 2019), 012053.
36. Muravyov, S., and Marinushkina, I. Intransitivity in multiple solutions of Kemeny ranking problem. *Journal of Physics Conference Series 459* (09 2013), 012006.
37. Organization, W. H. International classification of diseases : [9th] ninth revision, basic tabulation list with alphabetic index, 1978.
38. Organization, W. H. Icd-10 : international statistical classification of diseases and related health problems : tenth revision, 2004.
39. Rigden, D. J., and Fernández, X. M. The 2021 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Research 49*, D1 (12 2020), D1–D9.
40. Schalekamp, F., and van Zuylen, A. Rank aggregation: Together we're strong. In *Proc of ALENEX* (2009), pp. 38–51.
41. Schulze, M. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social Choice and Welfare 36*, 2 (2011), 267–303.
42. Song, Z.-y., Chao, F., Zhuo, Z., Ma, Z., Li, W., and Chen, G. Identification of hub genes in prostate cancer using robust rank aggregation and weighted gene co-expression network analysis. *Aging 11* (07 2019).
43. Stearns, M., Price, C., Spackman, K., and Wang, A. Snomed clinical terms: Overview of the development process and project status. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium* (02 2001), 662–6.