Avant-propos

Christine FROIDEVAUX[1], Marie-Laure MARTIN-MAGNIETTE[2,3] et Guillem RIGAILL[2,4]

[1] *Université Paris-Saclay, CNRS, LISN, Orsay, France*

[2] *IPS2, Université Paris-Saclay, CNRS, INRAE, Université d'Évry, Université Paris-Cité, Gif-sur-Yvette, France*

[3] *MIA Paris-Saclay, Université Paris-Saclay, AgroParis Tech, INRAE, Paris, France*

[4] *LaMME, Université Paris-Saclay, CNRS, Université d'Évry, Évry-Courcouronnes, France*

A.1. Introduction

L'étude des données biologiques a connu des changements profonds ces dernières années. Tout d'abord, le volume de ces données a considérablement augmenté en raison des nouvelles techniques haut débit pour les expériences. Ensuite, les progrès remarquables tant des méthodes d'analyses informatiques et statistiques que des infrastructures ont rendu possible le traitement de ces données volumineuses. Il convient alors d'intégrer ces données, c'est-à-dire d'en exploiter la complémentarité dans l'espoir de faire avancer la connaissance biologique. L'intégration des données pour permettre l'analyse la plus exhaustive possible constitue ainsi un enjeu majeur de la biologie.

Cet ouvrage propose d'aborder de façon pédagogique des travaux de recherche dans la science des données biologiques, en s'intéressant d'abord aux approches informatiques pour l'intégration des données biologiques et ensuite aux approches statistiques pour l'intégration des données omiques.

Intégration de données biologiques,
coordonné par Christine FROIDEVAUX, Marie-Laure MARTIN-MAGNIETTE
et Guillem RIGAILL. © ISTE Editions 2022.

A.2. Approches informatiques pour l'intégration des données biologiques

A.2.1. *Défis de l'intégration des connaissances biologiques*

Les connaissances biologiques ont donné lieu à de nouveaux champs d'application : au-delà de la biologie intégrative et systémique, elles sont précieuses pour la santé et l'environnement. En particulier la mise en relation des données omiques avec les connaissances sur les pathologies et les données cliniques a permis l'apparition de la médecine de précision, une formidable promesse pour la santé des individus. Mais pour cela il convient de pouvoir analyser de façon intégrée toutes les connaissances dont on dispose.

L'intégration des données des sciences de la vie doit faire face à plusieurs difficultés : outre le fait qu'elles sont massives (Big Data), elles sont hétérogènes (formats très variés), dispersées (on les trouve dans de très nombreuses bases de données), de diverses granularités (données génomiques ou pathologies) et de qualité très variable (les bases de données n'offrent pas toutes la même garantie de vérification (*curation*)).

À la différence d'autres domaines d'application où le processus d'intégration repose sur l'identification de concepts structurés en ontologies et sur lesquels on apparie les données, l'intégration de données biologiques procède de la réconciliation des données, par des approches d'algorithmique, d'apprentissage et de statistique. Cette intégration tente de plus en plus souvent de replacer l'humain au centre du processus.

A.2.2. *Solutions informatiques*

On a assisté à l'émergence d'un nouveau paradigme. On ne procède plus à deux phases distinctes, où la première visait à rassembler des données distribuées dans différentes bases et à les intégrer tandis que la seconde effectuait des analyses sur les données intégrées, mais les deux phases sont entrelacées : l'intégration sert à l'analyse, qui à son tour sert à mieux intégrer.

Un certain nombre d'entrepôts de données ont été développés pour rassembler de façon intégrée, c'est-à-dire structurée, cohérente et complémentaire, les données disparates portant sur un même domaine de la biologie. La constitution de ces entrepôts s'accompagne de méthodes d'interrogation des données pour en permettre l'analyse. Ces données peuvent être annotées à l'aide de termes conceptuels issus des ontologies, qui permettent de garder trace de la connaissance profonde qui leur est associée. Les ontologies permettent non seulement d'enrichir les connaissances avec des annotations mais aussi de raisonner sur ces connaissances. Elles sont au cœur du Web sémantique qui vise une représentation fine des données pour faciliter l'intégration et l'interprétation automatiques des données (Chen *et al*. 2012).

Enfin, les analyses effectuées sur les données utilisent une multitude d'outils très variés. Le processus de traitement des données qui enchaîne plusieurs outils, appelé *workflow*, devient un élément fondamental de l'analyse des données et est au cœur du changement de paradigme mentionné dans l'introduction. Concevoir et exécuter ces chaînes de traitements de données bioinformatiques est un enjeu important.

A.2.3. *Présentation de la première partie*

Le chapitre 1 présente les entrepôts de données pour les sciences de la vie, en se concentrant sur les données cliniques. Le chapitre 2 introduit les concepts et techniques du Web sémantique pour l'intégration de données omiques. Enfin, le chapitre 3 expose les problèmes et solutions bioinformatiques pour la conception et l'exécution des workflows scientifiques.

Ces chapitres mettent en lumière les liens étroits entre une bonne intégration et les principes de données FAIR (*Findable, Accessible, Interoperable, Reproducible*), et insistent sur l'importance de la provenance des données (Zheng *et al*. 2015). Ils pointent les enjeux éthiques de la protection des données personnelles stockées, notamment dans le domaine de la santé, en lien avec la sécurité des systèmes informatiques.

À travers ces chapitres, le lecteur verra comment, en termes d'intégration des données, les avancées de la recherche informatique bénéficient aux sciences de la vie, et comment une adoption plus large des méthodes informatiques pourrait encore davantage leur profiter. Inversement, les sciences de la vie offrent un formidable champ d'investigation pour le développement de méthodes informatiques innovantes.

A.3. Approches statistiques pour l'intégration de données omiques

A.3.1. *Défis statistiques de l'intégration*

L'intégration de données omiques est une thématique très vaste. Il est bien difficile d'en définir précisément les contours. Notre vision de l'intégration de données omiques est assez proche de celle présentée par Ritchie *et al*. (2015) :

> « [...] l'intégration d'informations (multi)-omiques d'une manière sensée pour fournir une analyse plus complète d'un point d'intérêt biologique[1]. »

Cette définition met l'accent sur les objectifs de l'intégration. L'analyse doit être sensée, cela va de soi, mais surtout apporter un éclairage nouveau sur une question

1. Traduction des auteurs.

biologique d'intérêt : autrement dit elle doit faire « mieux » qu'une analyse non intégrative.

Sur le plan biologique, une vision systémique du fonctionnement de la cellule motive parfaitement le développement de méthodologies pour intégrer les informations omiques. En effet, comment pourrions-nous comprendre les régulations de la cellule sans étudier ou comprendre les très nombreuses interactions moléculaires qui ont lieu en son sein : ADN-ADN, ADN-ARN, ARN-protéines, etc. Toutefois, l'intégration de données omiques n'est pas aisée. Ce n'est pas une solution miraculeuse et la démonstration qu'une analyse intégrative offre une vision biologique plus complète qu'une analyse non intégrative n'est pas toujours simple. Nous évoquons très brièvement ici quelques-unes des difficultés statistiques associées à l'intégration des données (Ritchie *et al*. 2015).

A.3.1.1. *Données hétérogènes et complexes*

Une des premières difficultés que l'on rencontre est certainement la diversité des données. Par exemple :

1) il faut intégrer des données de formats très différents : graphes, matrices, signaux, etc. ;

2) il faut intégrer des données correspondant à des échelles moléculaires très variées, par exemple, des données transcriptomiques et protéomiques ;

3) il faut intégrer des jeux de données déséquilibrés où certains échantillons ne sont pas présents dans tous les jeux de données.

A.3.1.2. *Données de qualité*

Comme bien rappelé par Ritchie *et al*. (2015), avant d'intégrer des données il faut analyser séparément chaque jeu de données et valider leur qualité. Pour avoir des résultats d'une analyse intégrative de grande qualité il faut des données de grande qualité.

A.3.1.3. *Données de grande dimension*

En génomique nous sommes souvent confrontés au problème de la grande dimension (Giraud 2014) : le nombre de variables p (gènes, protéines, transcrits) est souvent beaucoup plus grand que le nombre d'observations n (individus, échantillons). L'intégration a tendance à aggraver le problème. Supposons par simplicité que sur chaque jeu de données à intégrer d on observe les mêmes n échantillons et qu'on mesure p_d variables. Si sur chaque jeu de données on a déjà $n \ll p_d$, *a fortiori* $n \ll \sum_d p_d$.

Pour réduire l'importance de ce problème une solution consiste à réduire la dimension de chaque jeu de données. Il existe de très nombreuses techniques pour le faire, par exemple des techniques de *data mining* ou encore l'utilisation de bases de connaissances.

A.3.2. *Intégration et acquisition des connaissances intégration omique ou multiomique*

On se focalise souvent sur le besoin d'intégration de données multiomiques. Ce besoin est indéniable. Toutefois, au niveau statistique, il ne faudrait pas oublier le besoin d'intégration mono-omique. De nombreux outils d'analyse classique modélisent les entités biologiques de manière indépendante (ou quasi indépendante). Par exemple, pour l'étude de données RNA-seq, on a recours le plus souvent à une analyse différentielle où les gènes sont analysés de manière presque indépendante (Robinson *et al*. 2010 ; Love *et al*. 2014). Il y a une forme d'intégration au niveau de l'estimation du paramètre de surdispersion ou encore d'analyses de *pathways*. Cette intégration soulève déjà des difficultés statistiques très importantes. Mais il faudrait aller plus loin dans la modélisation des dépendances au sein d'un type de données omiques (voir par exemple les chapitres 4 et 5).

Clairement, intégrer des données doit permettre de profiter de très nombreux jeux de données déjà disponibles et réaliser des méta-analyses puissantes. On prédit souvent une science plus centrée sur les données, le calcul et la simulation. Néanmoins, il nous semble important de penser l'intégration dans le processus d'acquisition de connaissances (voir figure 1 de (Camacho *et al*. 2018)). Dans ce cadre une question importante est de savoir comment générer des données « faciles à intégrer ». En statistique on parle de « planification expérimentale ». La réponse dépendra évidemment de ce que l'on veut prédire ou comprendre biologiquement et des techniques de validation dont on dispose.

A.3.3. *Présentation de la seconde partie*

Pour résumer, l'intégration de données omiques semble un objectif clé pour une biologie plus intégrative et systémique. Sur le plan statistique, il reste quelques verrous méthodologiques, notamment la grande dimension, la gestion des données manquantes, la prédiction en contexte incertain et la validation.

Par ailleurs il ne faut pas oublier l'objectif : répondre à une question biologique. Définir cette question n'est pas toujours simple. S'agit-il de prédire ou comprendre un processus biologique d'intérêt ? L'analyse sera-t-elle supervisée, non supervisée ou semi-supervisée ? Quelles sont les hypothèses implicites ou explicites de l'analyse réalisée, sont-elles en accord avec la question biologique ?

Les chapitres statistiques de cet ouvrage illustrent, nous l'espérons, comment l'intégration permet d'avancer sur un point biologique d'intérêt, la diversité des approches méthodologiques et certaines des difficultés rencontrées. Les chapitres 4 et 5 abordent

l'intégration de données mono-omiques pour prédire un phénotype. Les chapitres 6 et 7 présentent des techniques exploratoires pour l'analyse multiomique.

Nous avons demandé aux auteurs de présenter leurs travaux de manière pédagogique et de fournir les codes de leurs analyses et simulations. Un comité de lecture composé de statisticiens, mathématiciens, bioinformaticiens et biologistes a pu apprécier et valider leurs efforts. Nous espérons que cela rendra ces chapitres accessibles au plus grand nombre. Nous remercions très chaleureusement les auteurs des chapitres et les relecteurs pour leur travail.

A.4. Bibliographie

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., Collins, J.J. (2018). Next-generation machine learning for biological networks. *Cell*, 173(7), 1581–1592.

Chen, H., Yu, T., Chen, J.Y. (2012). Semantic web meets integrative biology: a survey. *Briefings in Bioinformatics*, 14(1), 109–125.

Giraud, C. (2014). *Introduction to high-dimensional statistics*. Chapman & Hall/CRC Press, Londres.

Love, M.I., Huber, W., Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12), 550.

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2), 85–97.

Robinson, M.D., McCarthy, D.J., Smyth, G.K. (2010). Edger: a bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, 26(1), 139–140.

Zheng, C.L., Ratnakar, V., Gil, Y., McWeeney, S.K. (2015). Use of semantic workflows to enhance transparency and reproducibility in clinical omics. *Genome Medicine*, 7(73).