Avant-propos

Annie CHATEAU[1] et Mikaël SALSON[2]
[1] *Université de Montpellier, CNRS, LIRMM, Montpellier, France*
[2] *Université de Lille, CNRS, CRIStAL, Lille, France*

Pour étudier le vivant, depuis longtemps les scientifiques l'observent, tout autant à une échelle macroscopique en analysant l'aspect extérieur des organismes ou leur fonctionnement interne global, qu'à une échelle plus microscopique. L'observation poussée à son paroxysme consiste à étudier le cœur des cellules et les molécules du vivant qui en définissent leur fonctionnement : l'ADN (acide désoxyribonucléique) et l'ARN (acide ribonucléique). En effet, dans un organisme, l'ADN est le support de l'information génétique, qu'on appelle *génome*. Il tient donc une place centrale. Mais le génome n'est pas tout, celui-ci est composé de gènes qui permettent la production d'ARN ayant des rôles divers comme la synthèse de protéines ou la régulation de l'activité des cellules. Sous forme numérique, ces molécules d'ADN ou d'ARN sont le plus souvent représentées comme des textes sur des alphabets à quatre lettres (A, C, G et T pour l'ADN ; A, C, G et U pour l'ARN). À partir de ces séquences d'ADN et d'ARN, des méthodes informatiques permettent de répondre à un certain nombre de questions biologiques. C'est le cœur de cet ouvrage. Vous trouverez dans les différents chapitres des réponses, ainsi que leurs limites, à quelques questions fondamentales ayant parcouru, et parcourant encore, la bioinformatique. Comment rechercher rapidement une courte séquence, de quelques centaines de nucléotides, dans un génome qui peut en faire quelques milliards ? Comment comparer des séquences entre elles ? Comment arrive-t-on à reconstituer la séquence complète d'un génome ? Comment identifier les bactéries composant notre flore intestinale ? Comment, à partir de leurs séquences, prédire la structure que prendront certains ARN ?

Des séquences aux graphes,
coordonné par Annie CHATEAU et Mikaël SALSON.

Les méthodes qui sont décrites dans cet ouvrage tirent leur source de deux domaines ancrés sur des fondations anciennes, qui ont longtemps évolué l'un à côté de l'autre sans interaction profonde : l'informatique et la biologie. C'est au XX^{e} siècle que la symbiose entre les méthodes informatiques et les questions biologiques ont mené à un travail de modélisation conjointe et à la conception d'algorithmes, de méthodes et outils bioinformatiques. Les premiers séquençages d'ADN voient le jour à la fin des années 1970, avec des volumes faibles et des coûts gigantesques. Le besoin de stocker et de manipuler ces données de façon automatique se fait rapidement très pressant. On voit naître ainsi au milieu des années 1980 les premières bases de données de séquences. Ces bases de données se nourrissent des expériences de séquençage qui sont mises en commun par la communauté, grandissent toujours plus, et nécessitent des méthodes plus performantes. C'est ainsi que sont mises en place des méthodes d'alignement de séquences qui sont *dédiées* à ces séquences génomiques, et sont conçues dans un souci d'optimiser le temps passé ainsi que l'espace utilisé pour cette opération. Ces bases de données sont non seulement maintenues, mais également étendues et rendues publiques à l'échelle internationale, accélérant encore l'accès aux connaissances. L'accélération dans l'acquisition des données est également à l'œuvre, avec les premiers génomes complets de bactéries ou levures dès le milieu des années 1990, et le projet de génome humain qui a occupé de nombreuses équipes pendant plus d'une décennie. L'accès à la connaissance de ces génomes permet de questionner le vivant avec un point de vue tout à fait nouveau, et donne des perspectives à plusieurs champs d'application, notamment en santé, mais aussi en écologie et évolution, tout en augmentant la connaissance fondamentale des organismes et de leur fonctionnement.

Depuis le milieu des années 2000, les données génomiques sont acquises à un rythme bien plus soutenu suite à l'avènement des *séquenceurs à haut débit*, permettant, d'une certaine façon, de transformer à bas coût et à un rythme de plus en plus effréné des molécules d'ADN ou d'ARN en courtes séquences de lettres. On parle maintenant de projets concernant plusieurs milliers, voire dizaines de milliers, de génomes complets d'individus d'intérêt. Grâce à ces évolutions, il est possible de questionner le vivant d'une façon plus fine, à l'échelle des variétés et des individus d'une même espèce, mais aussi à l'échelle des différents tissus qui composent un organisme, ou encore à l'échelle d'un échantillon de milieu naturel contenant des milliers d'organismes différents. De nouvelles questions s'accompagnent du besoin de modéliser les données dans leur ensemble, de façon structurée, et de méthodes ayant pour vocation d'y répondre.

En parallèle, les capacités de stockage et de traitement de l'information, ainsi que les performances de calcul permises par des processeurs de plus en plus puissants et exploitant du parallélisme de plus en plus complexe, ont accompagné des progrès fulgurants dans le domaine de l'algorithmique et de la modélisation des problèmes par

des structures discrètes élaborées. Certaines opérations qui semblaient inaccessibles sont devenues courantes à moindre coût dans les programmes modernes et il n'est pas rare à l'heure actuelle de lancer ses calculs sur une grille dont les capacités excèdent de loin ce que l'on pouvait imaginer il y a une vingtaine d'années. Pour autant, cela n'est pas suffisant pour rendre réalisables toutes les études que l'on souhaite mener sur les données de séquençage et leurs dérivés.

Les données produites par les séquenceurs, par leur quantité (jusqu'à 10 millions de nucléotides par seconde) et par leurs particularités (dont les longueurs et types d'erreurs varient selon les technologies de séquençage), requièrent des méthodes appropriées afin d'en tirer des informations pertinentes dans un temps raisonnable sans recourir à des infrastructures de calcul gigantesques.

Les méthodes développées, même si elles sont généralement indépendantes de la technologie, doivent prendre en compte les contraintes de celle-ci afin de disposer de solutions applicables en pratique. En particulier, l'augmentation du volume de données à traiter rend certaines solutions impraticables et nécessite le recours à des heuristiques bien plus rapides. De ce fait, les méthodes utilisées en bioinformatique sont le plus souvent à la croisée entre méthodes exactes et approchées.

Afin de saisir au mieux les termes et les outils spécifiques de la bioinformatique, nous avons choisi d'en introduire la plupart dans le chapitre 1. Il détaille également les données (ADN, ARN et protéines) sur lesquelles nous travaillons, la manière dont elles sont obtenues. Ce chapitre présente aussi des notions d'algorithmiques utiles pour la compréhension de l'ouvrage, mais aborde plus largement des concepts utilisés en bioinformatique. Les autres chapitres exposent les problèmes les plus couramment étudiés en bioinformatique à partir des données génomiques. Certains chapitres sont plus axés sur les outils, d'autres sur les méthodes, et enfin d'autres encore s'attachent à une description plus détaillée des données. Nous présentons succinctement les questions auxquelles répondent les chapitres de l'ouvrage.

– Indexation de séquences. Face à l'afflux de données, comment faire pour les stocker, les interroger et les manipuler facilement ? C'est l'objet du chapitre 2 qui explique comment répondre à ces différents aspects. Les enjeux cruciaux ici sont la conservation de l'information, la flexibilité de la structure et sa capacité à répondre en un temps raisonnable aux questions les plus courantes, comme « est-ce que cette séquence est bien dans mon génome ? ».

– Alignement de séquences. Lorsque l'on étudie une ou plusieurs séquences, une question arrive très rapidement : comment savoir si des séquences sont similaires, si une séquence se retrouve approximativement dans une autre, et également déterminer un score qui permet de classer ces comparaisons entre elles. Répondre à la question « quelles sont les occurrences les plus significatives de mon motif dans ma

séquence ? » est crucial en bioinformatique. C'est ce que l'on appelle l'alignement de séquences, objet du chapitre 3. Il aborde également les aspects de comparaison sans alignement où, pour faire face au volume de données et à des taux d'erreurs parfois importants, recourir à un alignement n'est pas faisable et des heuristiques sont développées dans ce but.

– Assemblage des génomes. Dans le chapitre 4, la question abordée est : « Comment obtient-on la séquence complète d'un organisme à partir des lectures produites par le séquençage ? » Ce problème fondamental est donc né d'une difficulté technique qui rend impossible la lecture du génome d'un organisme en un seul morceau à partir de ses cellules. Cette difficulté technique a pour vocation à disparaître si les progrès en matière de séquençage rendent cette lecture en une seule passe possible, mais l'assemblage est pour le moment indispensable à la connaissance du génome et soulève de nombreuses problématiques, comme « comment choisir entre deux possibilités pour assembler les lectures ? » ou « comment évaluer la qualité de la reconstruction ? ». Les graphes sont une modélisation très intéressante dans ce contexte de reconstruction.

– Métagénomique et métatranscriptomique. Lorsque plusieurs organismes sont mélangés dans un échantillon, par exemple de sol, de milieu marin, de milieu interne à un organisme (le fameux microbiote), de nouvelles questions s'ajoutent à celles déjà présentes lors de l'assemblage. Ainsi, « comment déterminer quelles espèces sont présentes ? », ou encore « comment assembler les génomes lorsqu'ils sont mélangés ? ». C'est l'objet du chapitre 5.

– Repliement de l'ARN. La donnée ARN est une donnée à part, car sa structure secondaire joue un rôle fondamental dans le fonctionnement des organismes. Le chapitre 6 propose un aperçu des méthodes destinées à modéliser et à inférer cette structure secondaire à partir de la donnée de la séquence. La question fondamentale ici est donc : « Comment trouver et évaluer le repliement d'un mot sur lui-même, en tenant compte des affinités entre les caractères de ce mot ? »

Au-delà des solutions apportées pour répondre à ces nombreuses questions, il est plus que jamais nécessaire de prendre un recul salutaire par rapport aux traitements que l'on peut faire sur ces données. Que signifie « retrouver un morceau de séquence » de tel organisme dans un autre, quelle est la significativité de cette affirmation au regard des paramètres choisis, des méthodes employées et de leurs limitations en matière de modélisation ? Quelle est la place de cette méthode dans le paysage des méthodes déjà développées et des problématiques soulevées par l'actualité des données et des connaissances ?

C'est cet esprit critique, qui ne peut se développer qu'en connaissance de cause, que nous souhaitons aussi cultiver chez les lecteurs et lectrices. Comprendre que les

structures discrètes employées sont des modèles que l'on construit, des objets mathématiques qui ne sont pas des vérités absolues, que l'on conçoit en fonction d'un objectif, garder en tête que ces modèles peuvent avoir leurs limites, en termes de représentativité des données, en termes de pouvoir d'expression, en termes de méthodes classiques que l'on peut appliquer, et enfin faire des compromis entre ces limitations et le besoin d'avoir des réponses rapides à des questions cruciales qui ont pour vocation d'accroître la connaissance du vivant, tel est le travail illustré ici. Il est certes illusoire de comprendre chacun des mécanismes qui sous-tendent toutes ces méthodes, mais le schéma général qui est derrière peut être un guide utile pour notre pratique scientifique.

Cet ouvrage fait partie d'une série dédiée aux méthodes en bioinformatique, en particulier liées à l'analyse des séquences caractérisant l'information génétique dans les organismes. Cet ouvrage s'adresse à tous et à toutes, étudiants et étudiantes à partir du master, doctorants et doctorantes, jeunes et moins jeunes chercheurs et chercheuses. L'idée est de concentrer dans cet ouvrage des exemples de modélisations et de résolutions de problèmes utilisant les structures discrètes. Loin d'être exhaustif, cet ouvrage se veut un rappel et une ouverture pour celles et ceux qui se consacrent de manière pointue à l'un des domaines traités, et un ouvrage d'introduction pour la future génération de bioinformaticiens et bioinformaticiennes. Rédigé par des chercheurs et chercheuses du domaine, issues de formations diverses, la plupart avec une implication non négligeable au sein des formations en bioinformatique, cet ouvrage a été pensé avec un effort pédagogique qui le rend accessible à un public un peu moins averti. Nous espérons sincèrement que vous prendrez plaisir à le parcourir et à le conseiller autour de vous.

Afin de réaliser cet ouvrage, il a fallu concentrer l'énergie et la volonté de plusieurs personnes que nous tenons particulièrement à remercier. Tout d'abord les chercheurs et chercheuses qui ont accepté de rédiger les différents chapitres, exercice qui s'ajoute à des emplois du temps déjà souvent chargés, et ce, avec une qualité que nous saluons. Ensuite, nous tenons à remercier nos coordinatrices, Hélène Touzet et Anne Siegel, pour leur gentillesse, leurs conseils et leur accompagnement tout au long de cette aventure. Enfin, nous remercions ISTE Editions de nous avoir fait confiance et laissé une liberté complète sur l'organisation et la coordination de cet ouvrage.