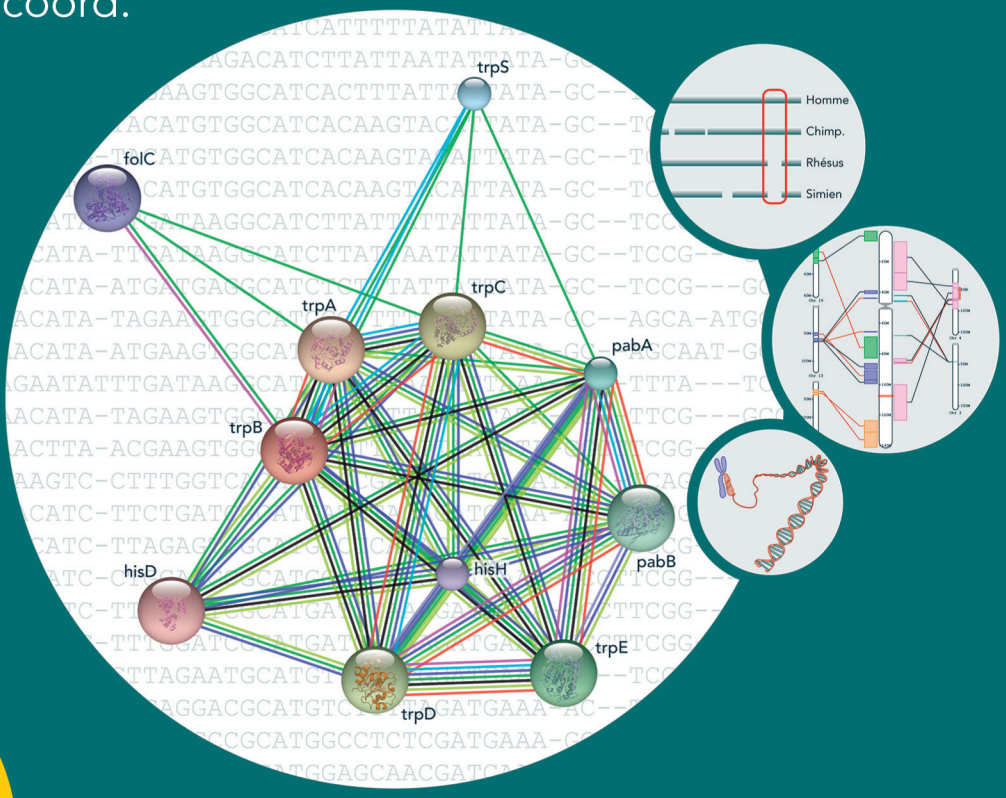
Savoir faire

Bio-informatique

Principes d'utilisation des outils

Denis Tagu, Jean-Loup Risler, coord.



éditions Quæ

Bio-informatique

Principes d'utilisation des outils

Denis Tagu, Jean-Loup Risler,
coordinateurs

Éditions Quæ

Collection *Savoir-faire*

Nutrition minérale des ruminants
François Meschy
2010, 212 p.

La gestion du trait de côte
Ministère de l'Écologie, de l'Énergie,
du Développement durable et de la Mer
2010, 304 p.

Évaluation économique de la biodiversité
Méthodes et exemples pour les forêts tempérées
Élodie Brahic, Jean-Philippe Terreaux
2009, 200 p.

Le campagnol terrestre
Prévention et contrôle des populations
Pierre Delattre, Patrick Giraudoux, coord.
2009, 304 p.

Retenues d'altitude
Laurent Peyras, Patrice Mériaux, coord.
2009, 352 p.

Référentiel pédologique 2008
Association française pour l'étude du sol
Denis Baize, Michel-Claude Girard, coord.
2009, 432 p.

Éditions Quæ
RD 10
78026 Versailles Cedex, France

 ISBN : 978-2-7592-0871-5 ISSN : 1952-1251

Avant-propos

Les biologistes font évoluer les connaissances sur le vivant par l'observation et l'expérimentation, dont l'efficacité dépend souvent de la performance d'outils de mesure ou d'analyse. L'histoire des sciences de la vie est ainsi ponctuée d'avancées qui ont été permises par des progrès techniques. Souvent, ces explorations du vivant dépendent de la disponibilité de nouvelles technologies issues de domaines autres que la biologie, comme la physique ou la chimie, l'automatique ou les mathématiques. L'accès à la connaissance de la séquence de génomes — qui marque les années 2000 — a bénéficié grandement de ces évolutions ; et la description des génomes ne peut pas se passer de l'informatique appliquée à la biologie : la bio-informatique, qui se situe à l'interface entre la biologie — plus particulièrement, mais pas uniquement, la génomique — et l'informatique.

Aujourd'hui, dans les laboratoires s'intéressant de près ou de loin à la structure, au fonctionnement et à l'évolution des génomes, s'approprier les outils d'analyse, de stockage et de visualisation des séquences d'acides nucléiques (ADN, ARN) et d'acides aminés (peptides, protéines) est devenu une nécessité. Les informaticiens spécialisés dans l'analyse du vivant développent des algorithmes, des bases de données, des méthodes et des outils, après avoir écouté les besoins exprimés par les biologistes ; quant à ces derniers — qui endossent alors la blouse de bio-analyste —, ils les utilisent. La particularité des approches des sciences du vivant fait de la bio-informatique un véritable terrain de recherche en informatique.

L'objectif de cet ouvrage n'est pas d'apprendre aux biologistes à programmer, mais de les amener à comprendre les outils d'analyse bio-informatique des acides nucléiques et des protéines à disposition, ainsi que leurs principes de fonctionnement, afin qu'ils soient à même de choisir celui qui sera ponctuellement le plus approprié à leur besoin. Ce livre s'adresse donc à toute personne qui, quel que soit son niveau de connaissance en génomique, travaille dans le cadre de programmes ou sur des projets de biologie moléculaire, de génomique ou de génétique.

L'ouvrage est structuré en cinquante-huit fiches regroupées thématiquement. Étudiées pour que le lecteur accède très efficacement à l'information recherchée, les fiches trouvent matière à approfondissement, à la fin de chaque thématique, sous la forme d'une sélection de références à des articles scientifiques, à des ouvrages et à des sites Web. Cet ouvrage ne se veut pas exhaustif, et les retours des lecteurs auprès des éditions Quæ seront appréciés afin que ces derniers participent également à une éventuelle deuxième édition de *Bio-informatique. Principes d'utilisation des outils*.

Sommaire

Domaines protéiques

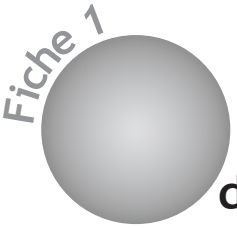
Reconstruction phylogénétique

Annotation des génomes

Comparaison des génomes

Analyse du transcriptome

Généralités



Bio-informatique et bio-analyse : définitions

Jean-Loup Risler

La « bio-informatique ». J'entends ce mot depuis bien longtemps, mais je ne sais toujours pas ce qu'il veut dire…

Je me souviens d'une rencontre organisée au CNRS entre informaticiens et biologistes, destinée à resserrer les liens entre les deux communautés. À l'époque (fin 1970-début 1980), les « bio-informaticiens » étaient essentiellement des structuralistes (rayons X et RMN). Les informaticiens étaient déjà des informaticiens. La réunion a essentiellement consisté en un long exposé théorique donné par un informaticien. Les biologistes ont tenté d'expliquer qu'ils avaient besoin des ordinateurs mais ne savaient pas forcément les programmer et/ou les utiliser, ce à quoi les informaticiens ont répondu qu'ils n'étaient pas des prestataires de services. Vous imaginez bien que le tout s'est terminé sur un retentissant constat d'échec.

Les choses se sont améliorées depuis, mais il subsiste un problème qui, à mon avis, tient essentiellement à la définition même du mot « informaticien ». Dans la communauté académique *française* — je mets en italique l'adjectif « française », car ce qui va suivre est une spécialité hexagonale —, un informaticien est un chercheur qui se livre à des recherches en informatique — cette phrase tient parfaitement si elle est mise au féminin. Le travail réalisé par un(e) informaticien(ne) est donc essentiellement théorique. Un chercheur en informatique (mathématique/statistique) n'est pas censé écrire des applications : c'est le rôle des ingénieurs. Et comme il y a un manque cruel d'ingénieurs…

Il y a donc une première vision de ce qu'est la bio-informatique : c'est une recherche originale en informatique, voire en mathématiques/statistiques, suscitée par un problème biologique, qui peut éventuellement conduire à l'acquisition de connaissances en biologie. C'est le cas, par exemple, des recherches menées sur les répétitions (exactes, inexactes, palindromiques) dans les séquences d'ADN et leur compression ; ou de la démonstration théorique par Karlin que les scores d'alignement (sans *gaps*) des séquences biologiques suivent une distribution dite des valeurs extrêmes — ce qui donnera lieu à l'écriture du programme BLAST ; ou encore de la mise en évidence de mots sur- ou sous-représentés dans les séquences

nucléotidiques, qui pose des problèmes statistiques épineux. Les recherches de ce type, qui sont publiées dans des journaux spécialisés, sont le plus souvent inconnues des biologistes.

La bio-informatique, ce peut être aussi la mise en œuvre — pas forcément triviale — de méthodes, de concepts ou d'algorithmes éprouvés pour résoudre un problème posé par les biologistes : par exemple, la comparaison de séquences génomiques complètes, ou l'utilisation de la transformée de Fourier pour créer des alignements multiples, ou encore la mise en musique des chaînes de Markov pour repérer les gènes codant les protéines dans les séquences génomiques. Il y a là production, par des informaticiens/mathématiciens/statisticiens, de programmes que les biologistes utiliseront.

Ce qui nous amène à une troisième définition possible de la bio-informatique, à savoir l'utilisation, généralement par un biologiste, d'un programme, le plus souvent écrit par un informaticien, pour produire de la connaissance en biologie. La plupart du temps, c'est à cette définition que pense un biologiste quand il se réfère à la « bio-informatique ». Le bio-informaticien est alors quelqu'un qui sait utiliser de façon raisonnée les nombreux programmes disponibles, sans pour autant être théoricien. L'informatique est ici un outil qui sert, par exemple, à analyser des séquences. Pour cette raison, le mot (ou le terme) bio-analyse est parfois employé.

Les frontières entre ces trois définitions ne sont évidemment pas étanches. Les Anglo-Saxons, pour ne citer qu'eux, ne rechignent pas à mettre les mains dans le « cambouis », tandis que les théoriciens produisent souvent des programmes utilisables par tout un chacun.

Fiche 2

Quelques généralités sur les gènes et les génomes

Denis Tagu

Les informations de cette fiche sont extraites de : Tagu D., Moussard C. (éds), 2006. *Principes des techniques de biologie moléculaire*. Versailles, éditions Quae, coll. « Mieux comprendre », 186 p., 2e éd.

Les matériaux de base des analyses bio-informatiques sont les polymères constituant les « molécules du vivant » que sont principalement les acides nucléiques et les protéines. La définition d'un gène est très variable selon l'angle de vue des chercheurs et leur spécialité : un biochimiste le définira comme un enchaînement de nucléotides assemblés par des réactions enzymatiques, alors qu'un biologiste de l'évolution le définira comme une entité sujette à sélection (par exemple). La question ici n'est pas de revenir sur cette large question, mais simplement de préciser quelques éléments de nomenclatures. La grande majorité des analyses des génomes portent actuellement sur les gènes codant des protéines, car celles-ci sont considérées comme les actrices majeures de la vie cellulaire. Cependant, la proportion dans un génome de gènes codant des protéines peut être faible (quelques % pour un génome eucaryote) ; coexistent également avec ces gènes de protéines, des séquences répétées, des transposons « actifs » ou « inactifs », des gènes correspondant à des ARN non codants, etc. L'identification de ces séquences d'ADN ne codant pas des protéines reste délicate encore actuellement, comme nous le verrons dans les fiches qui suivent.

Nous reprenons dans la figure 2.1 la représentation classique d'un gène codant une protéine chez un eucaryote. Un gène eucaryote comporte une séquence codante bordée de séquences régulatrices. Ces dernières servent de signaux de début et de fin de transcription du gène par l'ARN polymérase II. Certaines de ces séquences (p. ex. la boîte TATA du promoteur) sont reconnues par des protéines appelées « facteurs de transcription généraux », car elles assistent cette enzyme dans les étapes d'initiation de la transcription. D'autres séquences d'ADN, en aval ou en amont de la séquence codante, sont reconnues par des « facteurs de transcription spécifiques », qui modulent l'expression des gènes dans l'espace (p. ex. selon le type de cellule), dans le temps (p. ex. au cours du développement) et/ou sous l'effet de facteurs biotiques ou abiotiques (p. ex. le stress).

La séquence codante, chez un eucaryote, est constituée d'exons et d'introns. Ces deux types de séquences sont transcrits (transcrit primaire), mais les introns sont éliminés lors de l'épissage de l'ARN prémessager. L'ARN est d'abord modifié en 5′-P (addition d'une coiffe) et en 3′-OH (addition de nucléotides à adénine, ou « queue poly-A ») avant d'être transporté dans le cytoplasme. Là, les ribosomes se fixent sur l'ARNm (ARN messager) et, par l'intermédiaire des ARNt (ARN de transfert), l'ARNm est traduit en polypeptide.

L'ADN est constitué de deux brins antiparallèles et de séquences complémentaires. Lors de sa transcription en ARN, seul un des deux brins est lu et copié par l'ARN polymérase. Le brin d'ARN obtenu est donc complémentaire du brin matrice qui a servi de copie. Par définition, ce brin d'ADN matrice est le brin antisens ; l'autre brin, qui a la même séquence que l'ARNm, est le brin sens. Par convention, la séquence d'ADN identique à l'ARNm (donc le brin sens) est appelée « brin codant » ; l'écriture d'un gène sur le papier (ou un écran d'ordinateur…) correspond au brin codant dans son orientation 5′-P vers l'extrémité 3′-OH (figure 2.2).

Données actuelles (partielles) sur les génomes

De nombreux génomes procaryotes et eucaryotes ont été séquencés ou sont en cours de séquençage. Le lecteur peut se référer au site du *National Center for Biotechnology Information* (NCBI)[1], qui héberge l'un des serveurs Web compilant les avancées sur les génomes. La notion de « séquençage complet » est ambiguë : il y a très peu de génomes eucaryotes pour lesquels la séquence complète est réellement connue, car, le plus souvent, la qualité de séquençage et d'assemblage des séquences ne permet de couvrir qu'environ 80 % de la totalité d'un génome. La présence de nombreuses séquences répétées dans les génomes eucaryotes est un frein majeur à la complétion d'un séquençage et d'un assemblage.

▸▸ Procaryotes

Le premier génome séquencé a été celui de *Haemophilus influenzae* en 1995. Actuellement, près de 1 000 génomes de bactéries (eubactéries et archébactéries) sont séquencés.

▸▸ Eucaryotes

Plusieurs centaines de génomes eucaryotes sont répertoriés sur le serveur du NCBI, dont 24 complets (séquencés et assemblés, juillet 2009). Les autres sont en cours de séquençages ou d'assemblage. Parmi ces 602 génomes, on en trouve 237 d'animaux, 194 de champignons, 78 de plantes et 93 de protistes. Au sein des animaux, les mammifères (105) et les insectes (51) sont les classes les plus représentées, probablement pour l'intérêt de la communauté scientifique à nos origines proches (les mammifères) et la relative petite taille des génomes d'insectes actuellement séquencés. La taille des génomes est très variable d'une espèce à l'autre (environ 200 Mb pour la drosophile et 3 000 Mb pour l'homme). Par ailleurs, il n'y a pas de corrélation entre la taille physique de l'organisme et la taille des génomes : l'homme et la souris ont des tailles de génomes du même ordre de grandeur. Enfin, un organisme à grand génome n'est pas forcément plus riche en gènes : le génome de l'homme est plus de dix fois plus grand que celui de la drosophile, mais présente tout au plus deux fois plus de gènes.

[1] http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj (consulté le 27.09.2010).

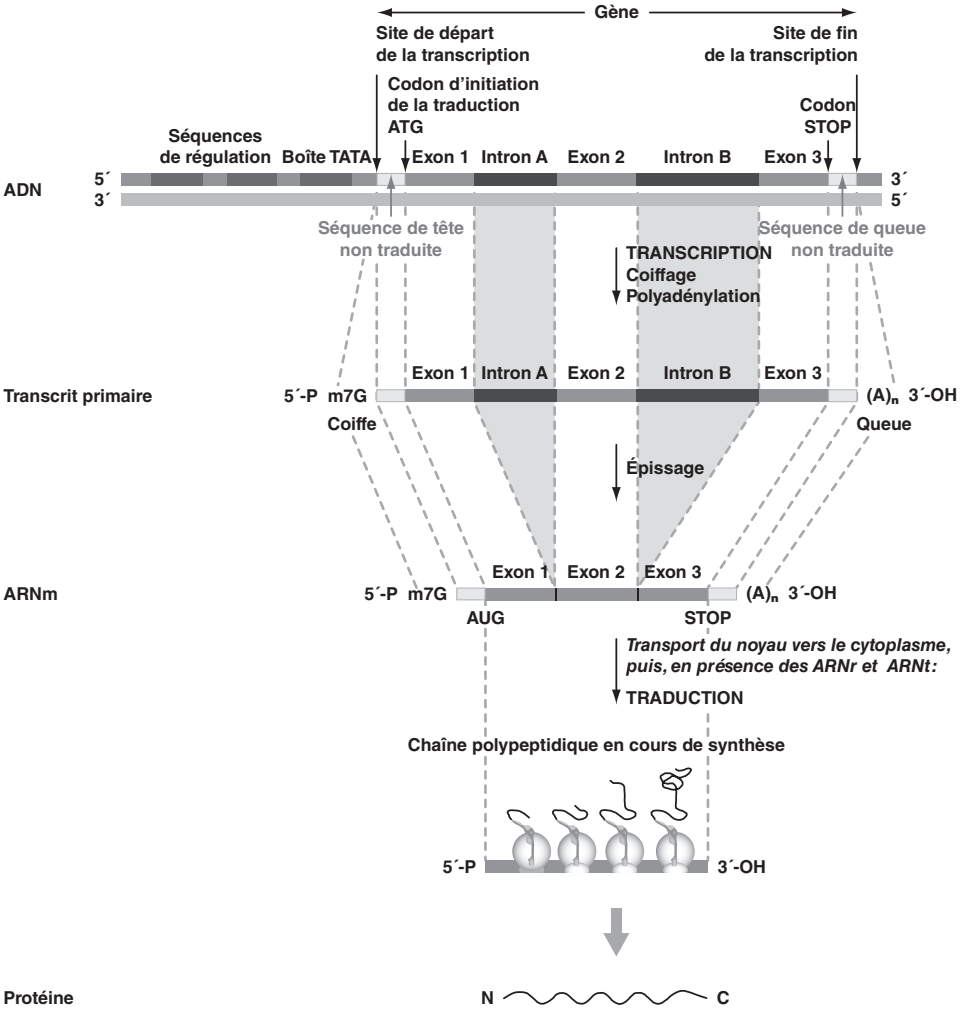


Figure 2.1. **Schéma de régulation de l'expression d'un gène eucaryote.**

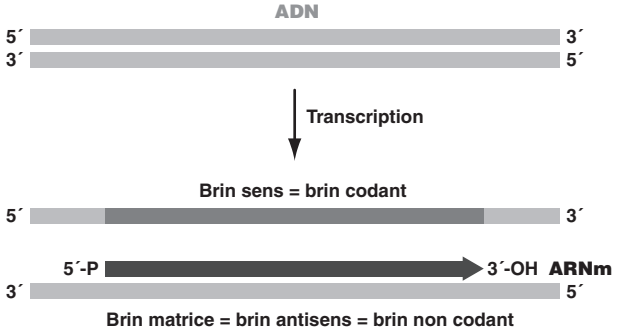


Figure 2.2. **Convention de nomenclature.**

Banques et bases de données en biologie