Article original : doi 10.3389/frym.2024.1266091

LA BIOINFORMATIQUE : L'ANALYSE DE DONNÉES PROVENANT D'ÊTRES VIVANTS

*Carmine Fruggiero†, Gaetano Aufiero† and Nunzio D'Agostino **

Département des sciences de l'agriculture, Université Frederico II de Naples, Portici, Italie

†Ces auteurs ont contribué également à ce travail

Sais-tu quelle quantité d'informations se cache à l'intérieur de tes cellules ? C'est une quantité énorme. Es-tu curieux de savoir comment les scientifiques déchiffrent et traitent ce grand volume de données ? L'informatique et les mathématiques ont aidé les biologistes à créer une nouvelle science pour analyser, organiser et comprendre les données biologiques : la bioinformatique. La bioinformatique nous permet de gérer d'énormes quantités de données biologiques et de leur donner un sens. En d'autres termes, la bioinformatique nous permet d'explorer les mystères de la vie et de trouver des réponses à des questions complexes sur le fonctionnement des êtres vivants. Dans cet article, nous sommes heureux de t'expliquer ce qu'est ce nouveau domaine fascinant et à quoi il sert. Nous pensons que cela te plaira !

LA BIOINFORMATIQUE : INTERPRÉTATION DES SYMBOLES BIOLOGIQUES

Tu seras peut-être surpris d'apprendre que, à l'exception des globules rouges, chaque cellule de ton corps contient un « manuel » contenant toutes les instructions nécessaires à la construction et à l'entretien de l'ensemble de ton corps, et cela dans un espace de 4 à 6 µm (un µm correspond à un millionième de mètre) ! Ces informations sont stockées dans l'ADN, que l'on retrouve dans tous les organismes (sauf quelques

virus) et qui constitue le « langage universel » de la vie. Dans les cellules humaines, ce langage est constitué de plus de trois milliards de « lettres ». Comment les scientifiques peuvent-ils étudier et interpréter l'énorme quantité d'informations inscrites dans l'ADN ? Cette tâche difficile a donné naissance à une nouvelle science, la bioinformatique [1].

BIOINFORMATIQUE. Science qui consiste à stocker de grandes quantités de données biologiques complexes et à les analyser pour en tirer de nouvelles conclusions.

GÉNOME. Ensemble des instructions de l'ADN que l'on trouve dans une cellule.

LE MANUEL DE LA VIE

Les scientifiques appellent génome le manuel de l'ADN, ou l'ensemble des instructions de l'ADN présentes dans chaque cellule (Figure 1). Ce manuel est divisé en « chapitres » appelés chromosomes. Chaque espèce possède un nombre spécifique de chapitres ou, en d'autres termes, chaque espèce possède un nombre déterminé de chromosomes. Certains organismes ont même plusieurs exemplaires de chaque chromosome : ainsi, les humains ont 23 paires de chromosomes et les tomates 12 paires. L'ADN est une longue molécule constituée d'une séquence de quatre « lettres » correspondant à quatre composés chimiques : l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). Voici un exemple de séquence d'ADN : ATGGTCCCATGCTAGCTAGCTATCGATGCTACGTACGTAGCATAAA TCGCGATAGCTA.

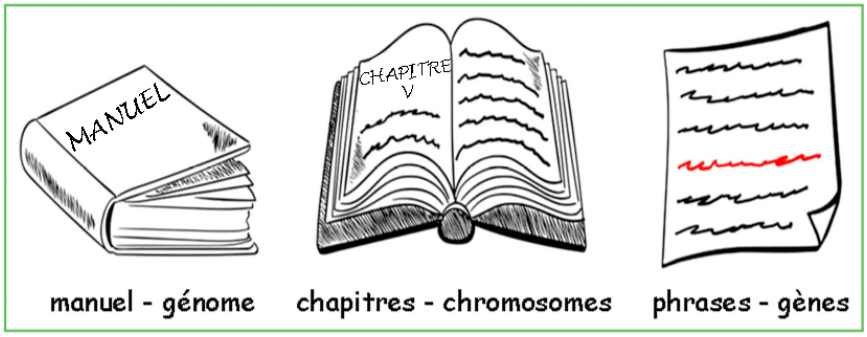

Figure 1. Tu peux considérer ton génome comme un manuel contenant toutes les instructions nécessaires à la fabrication de ton corps et à son fonctionnement. Chaque « chapitre » de ce manuel est un chromosome, et l'humain en possède 23 paires. À l'intérieur de chaque chapitre, les « phrases » contenant les instructions sont appelées gènes (en rouge).

Les combinaisons de ces quatre lettres sont utilisées par les cellules pour savoir comment se comporter. Le génome de chacun est un peu différent, et la somme de ces différences conduit à ton apparence unique et même influencent ta façon de penser et d'agir.

Comment les scientifiques peuvent-ils lire et comprendre le génome ? « Lire » un génome, c'est connaître la séquence des lettres qui le composent. Pour ce faire, les scientifiques utilisent une technique appelée séquençage du génome (pour en savoir plus sur le séquençage du génome regarde cet article). Mais ne connaître que la séquence, c'est comme avoir un livre écrit dans une langue inconnue qu'il faut déchiffrer. Tout d'abord, les bioinformaticiens cherchent à

SÉQUENÇAGE DU GÉNOME. Procédé qui détermine l'ordre des éléments chimiques, appelés bases, qui composent la molécule d'ADN.

identifier des phrases spécifiques dans le génome, appelées gènes.

Chaque gène contient l'information nécessaire à la fabrication d'une protéine spécifique. En fait, pour que la cellule soit capable d'une activité particulière, des lignes de texte (c'est-à-dire des gènes) sont sélectionnées dans le manuel ce qui permet la fabrication des protéines nécessaires à cette activité. Les protéines sont les « ouvrières » de l'organisme, chacune ayant une tâche particulière à accomplir. Les protéines peuvent aider à déplacer des éléments à l'intérieur des cellules, à construire des cellules, etc. Les protéines peuvent interagir entre elles ou avec d'autres composés chimiques pour fabriquer les muscles, les cheveux et les ongles, par exemple.

GÈNE. Unité fonctionnelle de l'hérédité qui porte l'information spécifiant les caractéristiques transmises de génération en génération.

PROTÉINE. Grande molécule complexe. Les protéines jouent de nombreux rôles importants dans les organismes vivants.

DÉMOTIQUE. La plus simple des trois écritures utilisées par les égyptiens ; elle est d'usage courant à partir du VIIIème siècle avant JC pour les documents de la vie quotidienne.

CODE GÉNÉTIQUE. Code mettant en relation les séquences de nucléotides (A, T, G, C) de l'ADN et les acides aminés des protéines. Chaque acide aminé correspond à un ou plusieurs groupes de 3 nucléotides (un codon).

ACIDE AMINÉ. Molécule utilisée par tous les êtres vivants, constituant des protéines.

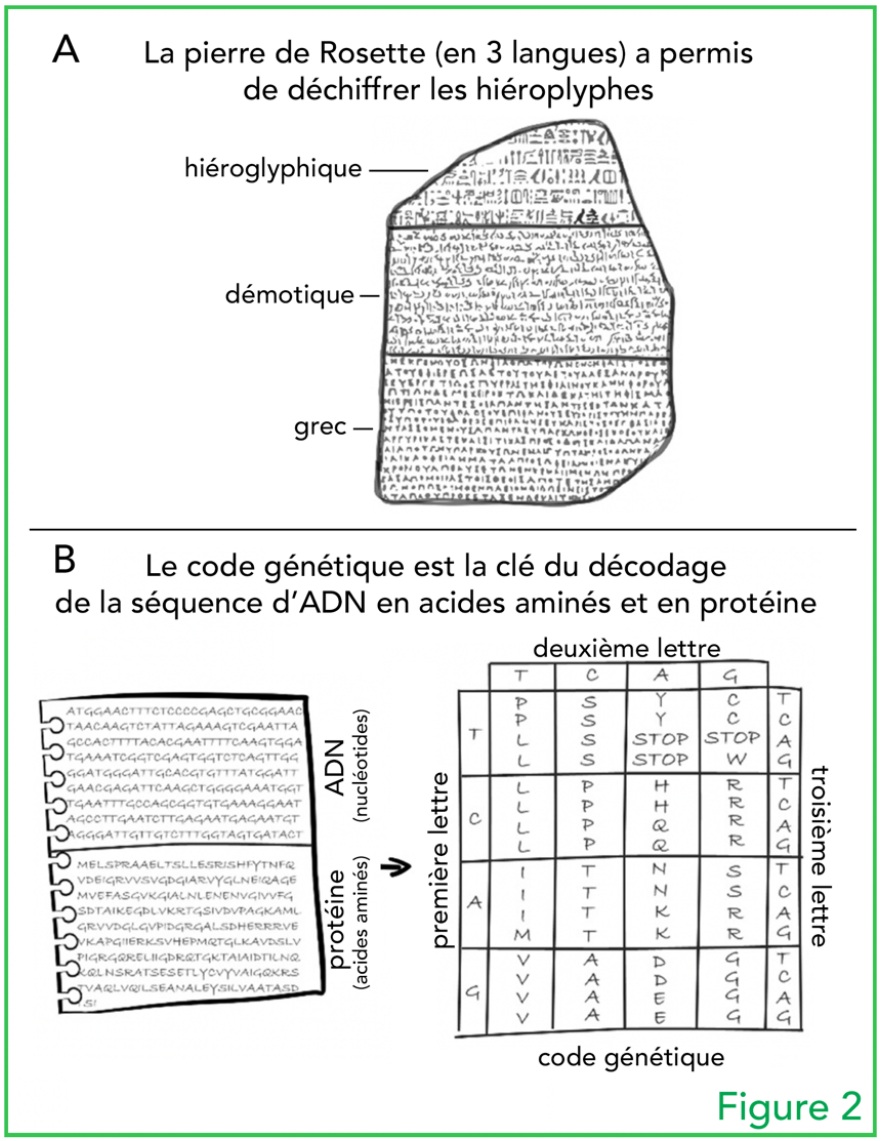

Figure 2. (**A**) La pierre de Rosette a permis de déchiffrer les hiéroglyphes (le même texte y est écrit en hiéroglyphes, en démotique, et en grec). (**B**) Le code génétique est la clé du décodage de la séquence d'ADN pour savoir quelle protéine est codée. Par exemple, le triplet de nucléotides ATG code pour l'acide aminé M (méthionine).

Les protéines sont des molécules composées de divers assortiments de 20 composés chimiques, les acides aminés. Un acide aminé peut être considéré comme un bloc de construction d'une protéine, comme lorsque tu assembles les briques de différentes couleurs de ton LEGO.

Selon la manière dont tu assembles les blocs, tu peux obtenir des créations particulières. De la même manière, une cellule assemble des acides aminés pour obtenir une protéine. Chaque fois que la cellule a besoin d'une protéine, elle traduit les instructions codées dans le gène correspondant en une séquence particulière d'acides aminés.

Pour effectuer cette traduction, la cellule applique un ensemble de règles connu sous le nom de code génétique. Le code génétique permet aux scientifiques de décoder la séquence d'ADN, tout comme les trois traductions de la pierre de Rosette ont aidé les gens à comprendre les hiéroglyphes égyptiens (Figure 2) [2]. Ton génome contient environ 20 000 gènes codant pour au moins 80 000 protéines qui, ensemble, assurent les fonctions de ton corps.

En résumé, les gènes comme les protéines peuvent être représentés par une chaîne de caractères, et le rôle passionnant des bioinformaticiens est d'étudier ces séquences pour en déchiffrer les secrets.

Si cela t'intéresse et que tu es impatient de devenir bioinformaticien, ne t'inquiéte pas ! Tout ce dont tu as besoin, c'est d'un ordinateur portable pour commencer ! La première chose que fait le bioinformaticien est de se familiariser avec les fichiers textes des ordinateurs où sont stockées les séquences d'ADN ou de protéines. Ensuite, à l'aide de quelques logiciels, il peut naviguer, déchiffrer et interpréter ce monde fait de lettres.

LA TÂCHE FONDAMENTALE DE LA BIOINFORMATIQUE : COMPARER DES SÉQUENCES

BIOSÉQUENCE. Ordre des éléments constituant les molécules biologiques, telles que l'ADN et les protéines.

Les scientifiques appellent les séquences d'ADN et de protéines des bioséquences. Tes propres bioséquences peuvent être comparées à toutes les autres bioséquences connues pour en déduire une descendance évolutive commune ou une structure/fonction similaire. Mais que signifie comparer des bioséquences ? Il s'agit d'utiliser des programmes informatiques pour les aligner par paires, afin de faire correspondre autant de lettres que possible (Figure 3). Le résultat de cet alignement révèle les concordances (symboles identiques), les non-concordances (symboles différents) et les lacunes (ajouts ou suppressions dans une séquence par rapport à l'autre). En d'autres termes, l'alignement de deux séquences d'ADN permet d'identifier les régions d'identité [3]. Les séquences de protéines peuvent également être comparées. Il existe des outils bioinformatiques qui permettent aux scientifiques de traduire une séquence d'ADN en une séquence de protéines, puis de procéder à l'alignement et à l'analyse.

Au lieu de procéder à une comparaison par paire, les bioinformaticiens peuvent également utiliser des outils qui leur permettent de rechercher une séquence spécifique dans une base de données. Par exemple, s'ils disposent d'une séquence et veulent savoir pour quel type de protéine

elle code ou quelle pourrait être sa fonction, un logiciel spécifique peut comparer cette nouvelle bioséquence avec toutes les bioséquences connues stockées dans des bases de données dédiées. Le logiciel renvoie des alignements par paire de toutes les séquences connues qui sont similaires à la séquence inconnue. Plus la similarité est grande, plus il est probable que les gènes ou les protéines remplissent les mêmes fonctions à l'intérieur de la cellule.

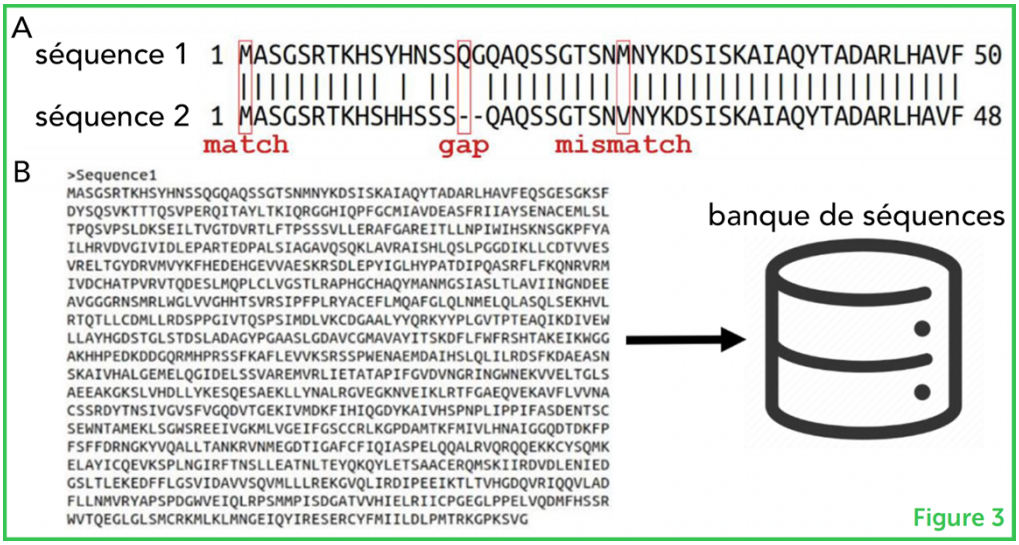

Figure 3. (**A**) Alignement des séquences de deux protéines. Des acides aminés concordants (*match*) et non-concordants (*mismatch*) ainsi qu'une lacune (*gap*) sont entourés en rouge dans les deux séquences. (**B**) Une nouvelle séquence, définie comme une « requête », peut être comparée à une base de données de séquences connues. Cette opération, connue sous le nom de recherche dans une base de données, renvoie une liste d'alignements par paire de séquences et peut indiquer aux scientifiques si la séquence est déjà connue ou à quel point elle est similaire à des séquences connues.

Imagine que tu viens d'obtenir une nouvelle séquence lors d'une expérience. La première chose à faire est de vérifier qu'elle code pour une protéine (séquence 1, Figure 3B). Tu ne sais rien des propriétés de cette protéine, ni de son rôle dans la cellule vivante. Tu vas alors comparer ta séquence (séquence 1, Figure 3A) avec une grande base de données de séquences dont les propriétés sont connues. Cela identifiera peut-être une séquence (séquence 2, Figure 3A) codant pour une protéine dont on connaît la fonction, par exemple une protéine qui fait entrer ou sortir des composés chimiques de la cellule. Tu pourras alors supposer que la nouvelle protéine que tu as découverte, pourrait également être impliquée dans le transport de molécules à travers la membrane d'une cellule. Dans l'ensemble, la comparaison des séquences est la première étape, et la plus instructive, de l'analyse de nouvelles séquences.

LA BIOINFORMATIQUE : UNE SCIENCE QUI ÉVOLUE CONTINUELLEMENT

Depuis 1978, date à laquelle le terme « bioinformatique » a été introduit, le domaine a connu une croissance explosive, en particulier au cours des

20 dernières années. La technologie a évolué pour rendre la bioinformatique encore plus puissante. Les données sont stockées en toute sécurité dans des superordinateurs et organisées en bases de données. Les données sont également analysées par des programmes informatiques de plus en plus efficaces, développés pour faire face à un volume croissant d'informations, appelées mégadonnées (big data).

L'utilisation de superordinateurs et de certaines formes d'intelligence artificielle peut contribuer à extraire des informations encore plus significatives des données biologiques. Au fur et à mesure que la bioinformatique se développe, cette science nous aidera à mieux comprendre le fonctionnement des organismes vivants et à prévoir toutes sortes de moyens passionnants pour aider les gens, qu'il s'agisse de défendre les humains contre les maladies ou d'aider les plantes à s'adapter au changement climatique.

REMERCIEMENTS

Les auteurs remercient tous les membres du laboratoire GenoPom pour les discussions stimulantes qu'ils ont eues pendant la préparation de cet article. Nous remercions tout particulièrement notre jeune amie (13 ans) Elizabeth Villmer, qui nous a fait part de ses commentaires du point de vue d'un jeune public.

RÉFÉRENCES

[1] Luscombe, N. M., Greenbaum, D., and Gerstein, M. 2001. What is bioinformatics ? An introduction and overview. *Yearbook Med. Informat.* 10:83–100. doi: 10.1055/s-0038-1638103

[2] Caskey, C. T., and Leder, P. 2014. The RNA code: Nature's Rosetta Stone. *Proc. Natl. Acad. Sci. U. S. A.* 111:5758–9. doi: 10.1073/pnas.1404819111

[3] Rosenberg, M. S. 2009. *Sequence Alignment: Methods, Models, Concepts, and Strategies*. Berkeley, CA: University of California Press.

VERSION FRANÇAISE

Cet article d'accès libre est une traduction avec modifications d'un article publié par Frontiers for Young Minds (doi : 10.3389/frym.2024.1266091 ; Fruggiero C, Aufiero G and D'Agostino N (2024) Bioinformatics: Analyzing Data From Living Things. *Front. Young Minds*. 12:1266091)

TRADUCTION : Nicole Pasteur, Association Jeunes Francophones et la Science

ÉDITION : Catherine Braun-Breton, Association Jeunes Francophones et la Science

MENTORS SCIENTIFIQUES : Océane Paris, Catherine Braun-Breton, Ula Hibner

REMERCIEMENTS : Merci à Benjamin Vierne, Josselin Gély et Romain Blanc pour leur accueil et leur implication dans l'édition de cet article par leurs élèves.

JEUNES ÉDITEURS :

CYRIL, NINO, ARWEN, CYRIAN, LOLA, CORENTIN, TARIK, YLIES, JANAËLLE, EVA,14-15 ANS

Nous sommes en classe de 3ème à Salon de Provence, dans le sud de la France. Nous aimons la science et nous nous intéressons au quotidien de la vie. Dans cet article, ce qui nous a le plus intéressés, c'est comment devenir bioinformaticien car cela pourrait nous inétresser plus tard. Lola aime le sport ; Corentin et Tarik aiment le foot et jouent dans un club.

YLIES, JANAËLLE, EVA, 14-15 ANS

Nous aimons la science et nous intéressons au quotidien de la vie. Dans cet article, ce qui nous a le plus intéressées, c'est comment devenir bioinformaticien car cela pourrait nous inétresser plus tard.

PAUL, 12 ANS

Je m'appelle Paul, j'ai 12 ans et j'aime le tennis et le parkour. J'ai une chienne qui s'appelle Cute (mignonne en anglais). J'ai passé 2 ans en Australie et le reste de ma vie en France. Carlos Alcaraz et Rafael Nadal sont pour moi les meilleurs joueurs de tennis du monde.

ADRIEN, 11 ANS

Je m'appelle Adrien et j'ai 11 ans. J'aime le handball

MARC, 11 ANS

Je m'appelle Marc, j'ai 11 ans et j'aime le sport. C'est ma troisième année en France. Dans mon école précédente, j'ai fait du karaté et joué aux échecs.

ARTICLE ORIGINAL (VERSION ANGLAISE)

SOUMIS le 24 juillet 2023 ; ACCEPTÉ le 15 janvier 2024.
PUBLIÉ EN LIGNE le 31 janvier 2024.

ÉDITION : Ornella Cominetti

MENTORS SCIENTIFIQUES : Renee WY Chan et Anita Singh

CITATION : Fruggiero C, Aufiero G and D'Agostino N (2024) Bioinformatics: Analyzing Data From Living Things. *Front. Young Minds.* 12:1266091. doi: 10.3389/frym.2024.1266091

DÉCLARATION DE CONFLIT D'INTÉRÊTS : Les auteurs déclarent que les travaux de recherche ont été menés en l'absence de toute relation commerciale ou financière pouvant être interprétée comme un conflit d'intérêt potentiel.

JEUNES EXAMINATEURS

ADI, 12 ANS

Je suis passionné de Lego et aime les sciences et les mathématiques. J'aime aussi faire des impressions 3D d'avions et d'engrenages. La chose la plus amusante pour moi, quand je ne code pas ou n'imprime pas en 3D, est de lire sur les nouveaux sujets en science et technologie. J'aime me plonger dans des projets et des vidéos sur ces sujets qui permettent d'en apprendre davantage d'une manière amusante. Je joue des instruments de percussion et je suis en train d'apprendre à jouer du saxophone ténor.

DIYA, 12 ANS

Je suis une jumelle fière qui aime écouter de la musique et apprécie l'art. Je suis passionnée par l'aide aux personnes qui n'ont pas accès aux ressources permettant d'améliorer l'apprentissage des STIM (Sciences, Technologie, Ingénierie, Mathématiques). Chaque année, je participe à l'organisation d'un camp d'été pour ingénieurs et je visite des pays où je peux partager mes connaissances. Je suis également danseuse et j'aime passer du temps avec mon chien.

JOHNSON, 14 ans

J'ai commencé à travailler comme jeune examinateur pour cette revue à l'âge de 10 ans. Actuellement, je suis un élève de 10ème année dans la filière scientifique et je suis également YouTuber. Mon expérience en tant que jeune examinateur d'articles scientifiques a élargi mes perspectives. Il est étonnant de voir comment les hypothèses scientifiques se traduisent en résultats expérimentaux de la vie réelle.

AUTEURS

CARMINE FRUGGIERO

Carmine Fruggiero a obtenu un diplôme en biotechnologie agro-environnementale et alimentaire à l'Université Federico II de Naples avec mention très bien. Il est actuellement doctorant en bio-informatique.

GAETANO AUFIERO

Gaetano Aufiero a obtenu un diplôme en biotechnologie agro-

environnementale et alimentaire à l'Université Federico II de Naples avec mention très bien. Il est actuellement doctorant et travaille sur l'analyse des bioséquences.

NUNZIO D'AGOSTINO

Nunzio D'Agostino est professeur associé de bioinformatique et de génomique au Département des sciences agricoles de l'Université Federico II de Naples. Son principal domaine de recherche est la bioinformatique appliquée à l'étude des génomes végétaux et à l'amélioration génétique des espèces végétales. Son activité de recherche se concentre en particulier sur l'analyse des données de séquençage de nouvelle génération et sur le développement de stratégies, de méthodes et d'outils pour la gestion des données -omiques. *nunzio.dagostino@unina.it