médecine/sciences 2026 ; 42 : 621-9

Bio-informatique et analyses « omiques » : chronologie d'une nouvelle approche scientifique

Sylvère Bastien [1,2], Marouchka Gourinovitch[1], François Vandenesch [1,2], Karen Moreau[1]

Depuis les premiers débats des années 1940-60 sur la nature du support de l'information génétique – opposant protéines et ADN – jusqu'à l'essor des approches « omiques » désormais largement utilisées pour explorer la complexité des génomes et de leur expression, la biologie moléculaire a connu de profondes évolutions et véritables révolutions technologiques. Ces avancées ont permis d'acquérir une compréhension toujours plus fine et exhaustive du vivant, tout en faisant émerger de nouveaux défis liés au développement d'outils bio-informatiques capables d'analyser des volumes de données en constante croissance. En offrant un accès inédit à l'organisation et au fonctionnement des systèmes biologiques, elles ouvrent par ailleurs la voie à de nombreuses applications, notamment en recherche et en pratique médicale.

[1] CIRI, Centre international de recherche en infectiologie, Université de Lyon, Inserm U1111, Université Claude Bernard Lyon 1, CNRS UMR5308, ENS de Lyon, Lyon, France.
[2] Centre national de référence des staphylocoques, Institut des agents infectieux, Hospices Civils de Lyon, Lyon, France.
karen.moreau@univ-lyon1.fr

Des approches biochimiques aux approches « omiques »

Avant les années 1960, un débat majeur a structuré les fondations de la biologie moléculaire moderne : quel est le véritable support de l'information génétique ? Tandis que les protéines étaient longtemps considérées comme les candidates évidentes en tant que vecteurs de cette information, notamment en raison de leur diversité structurale, l'ADN, alors encore peu compris, n'apparaissait pas comme une molécule pouvant jouer ce rôle. Ce questionnement a profondément influencé l'émergence de deux disciplines complémentaires : la génétique et l'étude des protéines.

Sur le plan expérimental, les progrès en cristallographie aux rayons X ont permis une meilleure compréhension de la structure tridimensionnelle des protéines, et leur séquence est devenue accessible grâce à la mise au point de la méthode de dégradation développée par le biochimiste suédois Pehr Edman (1916-1977), permettant d'obtenir des fragments peptidiques à partir de protéines purifiées [1, 2]. Cette avancée a été complétée par les premières tentatives d'analyse bio-informatique des données protéiques, notamment grâce au logiciel Comprotein conçu par la bio-informaticienne américaine Margaret Dayhoff (1935-1983), en 1962 [3]. Ce programme, s'apparentant à une approche d'assemblage *de novo*, permettait de reconstruire des structures primaires (constituées par la séquence en acides aminés), à partir de séquences partielles, jetant ainsi les bases de la protéomique comparative. L'étude de l'hémoglobine a par exemple permis de comparer les séquences chez différentes espèces animales et d'en déduire des relations phylogénétiques et des mécanismes d'évolution convergente [4].

Les besoins croissants en analyse de séquences ont conduit également à des avancées méthodologiques importantes. Afin d'améliorer la comparaison entre séquences protéiques, deux chercheurs américains de l'université Northwestern (Chicago, États-Unis), Saul Needleman (1927-2019) et Christian Wunsch, développent en 1970 un algorithme d'alignement des séquences protéiques, prenant en compte les insertions et les délétions [5]. Cet outil a ouvert la voie aux alignements de séquences multiples, formalisés dans les années 1980 avec des outils tels que CLUSTAL, encore largement utilisés dans les pipelines bio-informatiques modernes [6, 7]. Dès les années 1970, l'introduction de l'électrophorèse bidimensionnelle par le biochimiste

américain Patrick O'Farrell, permettant de séparer et visualiser simultanément plusieurs centaines de protéines et leurs isoformes a marqué une étape supplémentaire [8]. Ces approches expérimentales, combinées à la cristallographie, ont posé les bases d'une analyse globale du répertoire protéique.

Parallèlement, le domaine de la génomique connaît également ses premières révolutions. La mise en évidence en 1953 de la structure en double hélice de l'ADN par James Watson (1928-2025), Francis Crick (1916-2004) et Rosalind Franklin (1920-1958) a constitué un tournant décisif puisqu'elle a permis de l'établir comme un support potentiel de l'hérédité [9, *m/s* 10]. Le déchiffrement du code génétique entre 1961 et 1966, récompensé par un prix Nobel de physiologie ou médecine décerné à Robert W. Holley (1922-1993), Har Gobind Khorana (1922-2011) et Marshall W. Nirenberg (1927-2010) en 1968, établit définitivement le lien fonctionnel entre l'ADN et les protéines, en montrant comment l'information génétique est traduite en acides aminés [11].

Créée en 1977, la méthode de séquençage développée par le biochimiste anglais Fréderick Sanger (1918-2013), robuste et accessible, deviendra la technique de référence pendant plusieurs décennies, et permettra le séquençage complet du génome du bactériophage ΦX174, une molécule d'ADN de 5 386 bases [12, 13]. Cet exploit marque une étape majeure, car il permet, pour la première fois, de relier une séquence génomique complète à l'ensemble des protéines qu'elle code, sans passer par leur purification. Cette avancée inaugure une nouvelle discipline : la génomique, qui permet de déterminer le contenu protéique à l'échelle d'un organisme entier, notamment via l'identification de nouveaux gènes, la comparaison de séquences génomiques interespèces, la reconstruction d'arbres phylogénétiques, et la prédiction des structures primaires de toutes les protéines codées par un génome donné.

Contrairement aux approches initiales, centrées sur l'étude d'une ou quelques protéines purifiées, l'essor du séquençage de l'ADN a facilité le développement de la génomique qui, en offrant une vision globale du répertoire protéique, a progressivement transformé les approches expérimentales et computationnelles. Cette transition a ainsi posé les bases de la biologie moléculaire moderne, pour laquelle l'analyse des séquences, protéiques ou nucléotidiques, est devenue centrale pour comprendre les mécanismes du vivant, et a conduit à l'avènement de la génomique fonctionnelle et de la protéomique systématique actuelles.

L'essor de la bio-informatique : de l'analyse de séquences à l'explosion des données génomiques et transcriptomiques

Dans la continuité des avancées en génomique et en protéomique, les années 1990 marquent un tournant décisif avec l'émergence de la bio-informatique (*Figure 1*). En réponse à l'explosion des données de séquençage, soutenue par les progrès technologiques et la structuration des connaissances biologiques, ce champ interdisciplinaire s'est développé, en premier lieu, grâce à l'apparition des premiers ordinateurs personnels, combinée à l'émergence de langages de programmation adaptés à l'analyse de données biologiques, tels que Fortran, C ou Perl, qui ont permis l'adoption des outils informatiques dans les laboratoires de biologie, notamment pour l'assemblage et l'annotation des génomes [14, 15]. Aujourd'hui, les langages Python et R sont des standards en bio-informatique pour le traitement, la visualisation et la modélisation des données. L'ouverture d'internet au grand public a ensuite permis le partage de données biologiques à grande échelle avec la mise en ligne de plusieurs bases de données, parmi lesquelles la librairie de données de séquences nucléotidiques, hébergée par le laboratoire européen de biologie moléculaire (EMBL-EBI, *European molecular biology laboratory-European bioinformatics institute*) [16], intégrant notamment la base de données SWISS-PROT [17], disponible à partir de 1993. L'année suivante, le NCBI (*National center for biotechnology information*) lance son propre portail, incluant l'outil BLAST (*basic local alignment search tool*) qui devient rapidement un incontournable pour la recherche de similarités entre séquences [18]. Ces avancées s'inscrivent dans une démarche collaborative globale, et, dès 1987, un effort de standardisation des formats de données entre les principales banques de séquences – GenBank (États-Unis), EMBL (Europe), et DDBJ (*DNA data bank of Japan*) – aboutit à un consensus international, aujourd'hui représenté par l'INSDC (*International nucleotide sequence database collaboration*) facilitant ainsi la compatibilité et l'interopérabilité des données à travers les différents dépôts [19, 20].

En 1995, un jalon important est franchi avec le séquençage complet du génome de *Haemophilus influenzae*[1] (1,83 mégabases) par l'Institut pour la recherche génomique (TIGR, *the Institute for genomic research*), créé par Craig Venter, ouvrant la voie à des projets de plus grande envergure, tels que le projet du génome humain, lancé dans les années 1990 par le DOE (*Department of energy*), suivi par le NIH (*National institutes of health*) aux États-Unis. Ce projet a mobilisé un vaste consortium international sur plus d'une décennie nécessitant un investissement estimé à plusieurs milliards de dollars. Il marque le début de la génomique à très grande échelle et révèle le besoin de technologies capables de générer et d'analyser des volumes massifs de données (*Figure 2*). C'est dans ce contexte qu'au début des années 2000, l'émergence de techniques de séquençage de seconde génération (NGS, *next-generation sequencing*) avec notamment la plateforme 454

[1] *Haemophilus influenzae*, autrefois appelé bacille de Pfeiffer, est une bactérie strictement humaine de la famille des Pasteurellaceae et de la classe des Gammaproteobacteria. Richard Pfeiffer (1858-1945) a été le premier à les décrire en 1892 lors de la pandémie de grippe (influenza) de 1889-1892 (ndlr).

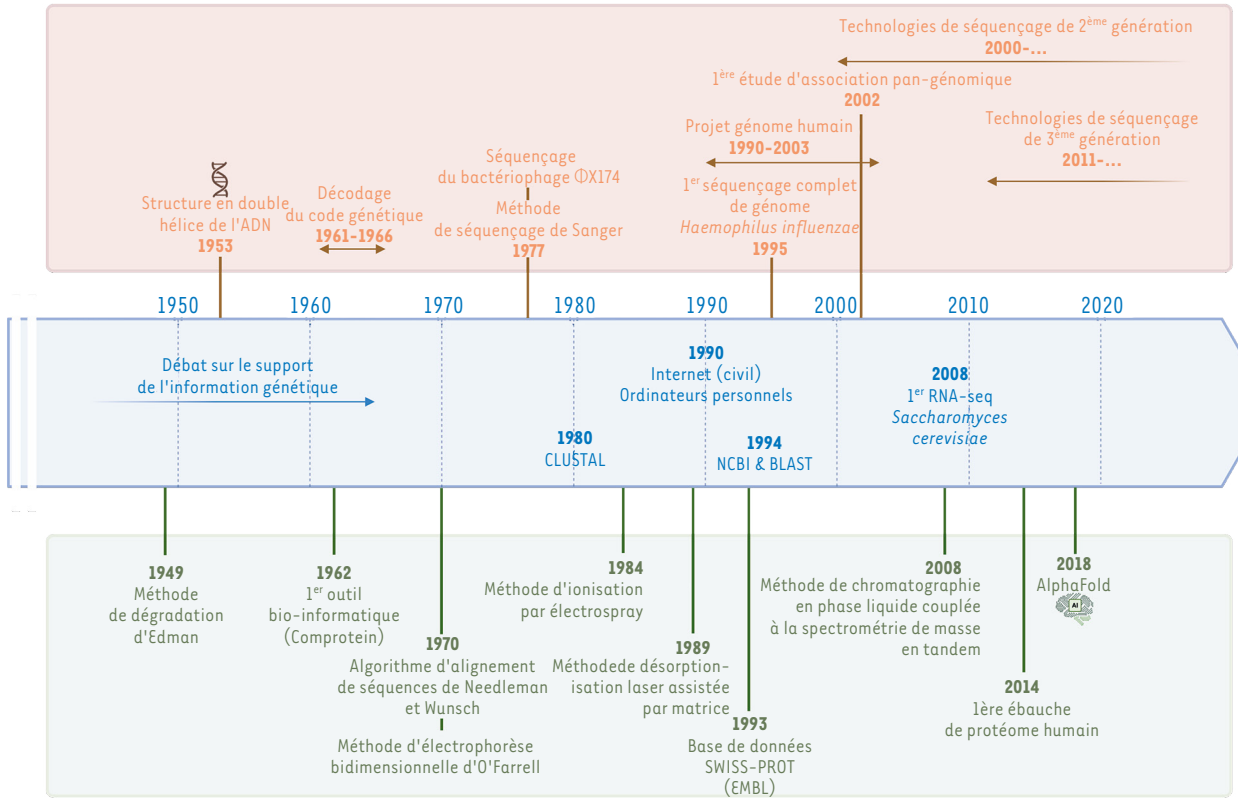

Figure 1. Frise chronologique de différents évènements et avancées techniques, expérimentales et analytiques, depuis les prémices des analyses génétiques et protéiques jusqu'à l'intégration de l'intelligence artificielle comme outil au service de la biologie. Figure réalisée avec BioRender.

(Roche) fondée sur le principe du pyroséquençage[2] permet pour la première fois de séquencer des millions de molécules d'ADN en parallèle [21, 22]. Pour accompagner ces volumes de données, la société Roche développe également un outil d'assemblage nommé Newbler. Rapidement, d'autres plateformes prennent le relais, en particulier Illumina, qui devient la référence du marché grâce à sa capacité de production massive et sa précision [23]. L'arrivée de ces technologies entraîne une réduction exponentielle du coût par base séquencée, tout en améliorant la qualité des données et leur vitesse d'acquisition.

Par la suite, des techniques de troisième génération voient le jour. Elles permettent le séquençage en molécule unique, de façon continue et en temps réel, via des méthodes comme SMRT (*single molecule real time*) de Pacific Biosciences, ou le séquençage par nanopore *d'Oxford Gene technologies* (*Figure 2*) [*m/s* 24]. Cette course technologique a profondément bouleversé le paysage de la génomique, en rendant possible le séquençage du génome entier pour un large éventail d'organismes, mais aussi en posant de nouveaux défis liés à la gestion, au stockage et à l'analyse des données générées.

Du côté de la protéomique, un véritable essor a été rendu possible dans les années 1990 grâce aux avancées de la spectrométrie de masse (MS, *mass spectrometry*) *via* le développement des sources d'ions avec l'appui des techniques de désorption-ionisation laser[3], assistée par matrice (MALDI, *matrix-assisted laser desorption ionisation*) et d'ionisation par électrospray (ESI, *electrospray ionization*), qui ont facilité l'analyse précise de mélanges complexes de protéines, des avancées valorisées par un prix Nobel de chimie en 2002, attribué à John Fenn (1917-2010) et Koichi Tanaka [25]. À partir des années 2000, les plateformes de spectrométrie de masse à haute résolution couplées à la chromatographie liquide (LC-MS/MS) sont devenues le cœur de la protéomique moderne, permettant l'identification et la quantification de milliers de protéines, et générant ainsi de grandes quantités de données (*Figure 3*) [26].

[2] Le pyroséquençage est une technique de séquençage de l'ADN qui identifie les nucléotides incorporés au fur et à mesure de la synthèse du brin d'ADN grâce à la détection d'un signal lumineux.

[3] La désorption-ionisation laser est une méthode utilisée en spectrométrie de masse dans laquelle l'énergie d'un laser provoque la désorption et l'ionisation des molécules présentes dans l'échantillon, permettant ensuite leur séparation et leur identification selon leur rapport masse/charge.

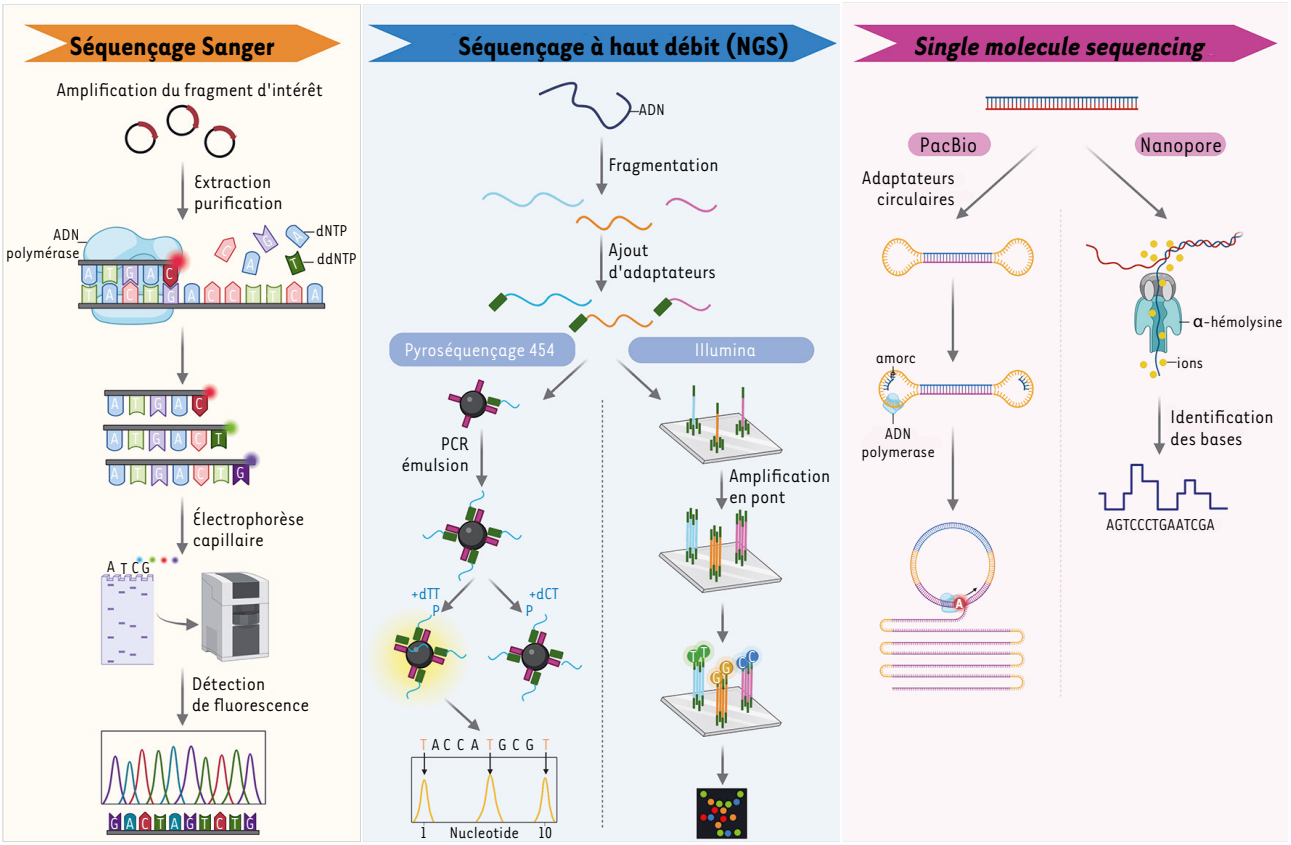

Figure 2. Principales approches pour le séquençage de l'ADN. Le séquençage par la méthode de Sanger repose sur une amplification préalable du gène d'intérêt, par PCR (*polymerase chain reaction*) ou par clonage. L'ADN est incubé avec une ADN polymérase, des désoxynucléotides (dNTP) et des didésoxynucléotides (ddNTP) marqués par un fluorophore et dépourvus du groupement 3'OH, agissant comme terminateurs de chaîne. Quatre réactions parallèles, contenant chacune l'un des quatre ddNTP, produisent des fragments de différentes tailles, séparés par électrophorèse et détectés par fluorescence pour reconstituer la séquence. Les technologies de séquençage à haut débit reposent sur la fragmentation de l'ADN et l'ajout d'adaptateurs pour une amplification préalable. Dans le pyroséquençage, les fragments sont amplifiés sur des billes par PCR en émulsion ; chaque microgouttelette contient ainsi une bille et un fragment, amplifié de manière clonale. Les nucléotides sont ajoutés séquentiellement et leur incorporation libère du pyrophosphate (PPi), qui est converti en ATP, puis utilisé par la luciférase pour produire un signal lumineux, permettant la lecture de la séquence. Dans le séquençage Illumina, les fragments s'hybrident à des amorces fixées sur une surface et sont amplifiés par « amplification en pont » pour former des clusters de copies identiques d'un même fragment. Le séquençage par synthèse incorpore un à un des nucléotides marqués, et la fluorescence émise après chaque incorporation permet de reconstituer la séquence de millions de fragments en parallèle. Contrairement aux approches précédentes, les technologies de troisième génération ne nécessitent pas d'amplification préalable de l'ADN : la méthode PacBio utilise des adaptateurs circulaires et séquence plusieurs fois la même molécule pour corriger les erreurs. Le séquençage par nanopore détecte les variations de potentiel ionique générées lors du passage de chaque base à travers un nanopore, permettant de déterminer directement la séquence. Figure réalisée avec BioRender.

De l'explosion des données à l'intelligence artificielle : structurer et exploiter l'information

Face à l'explosion du volume de données générées, des plateformes de dépôt et de partage structurées sont créées, comme la SRA (*sequence read archive*) [27] ou l'ENA (*european nucleotide archive*) [28]. Celles-ci intègrent progressivement des standards de métadonnées afin de garantir la traçabilité, la qualité et la réutilisabilité des données, en accord avec les principes FAIR (*findable, accessible, interoperable, reusable*)[4] [29]. En parallèle,

[4] Les principes FAIR, formalisés en 2016 par Wilkinson *et al.*, visent à améliorer la gestion et le partage des données scientifiques en les rendant faciles à trouver (*findable*), accessibles (*accessible*), interopérables (*interoperable*) et réutilisables (*reusable*). Ils ont été conçus dans le contexte du développement de la science fondée sur les données, avec une attention particulière à la réutilisation par les chercheurs mais aussi par les outils informatiques.

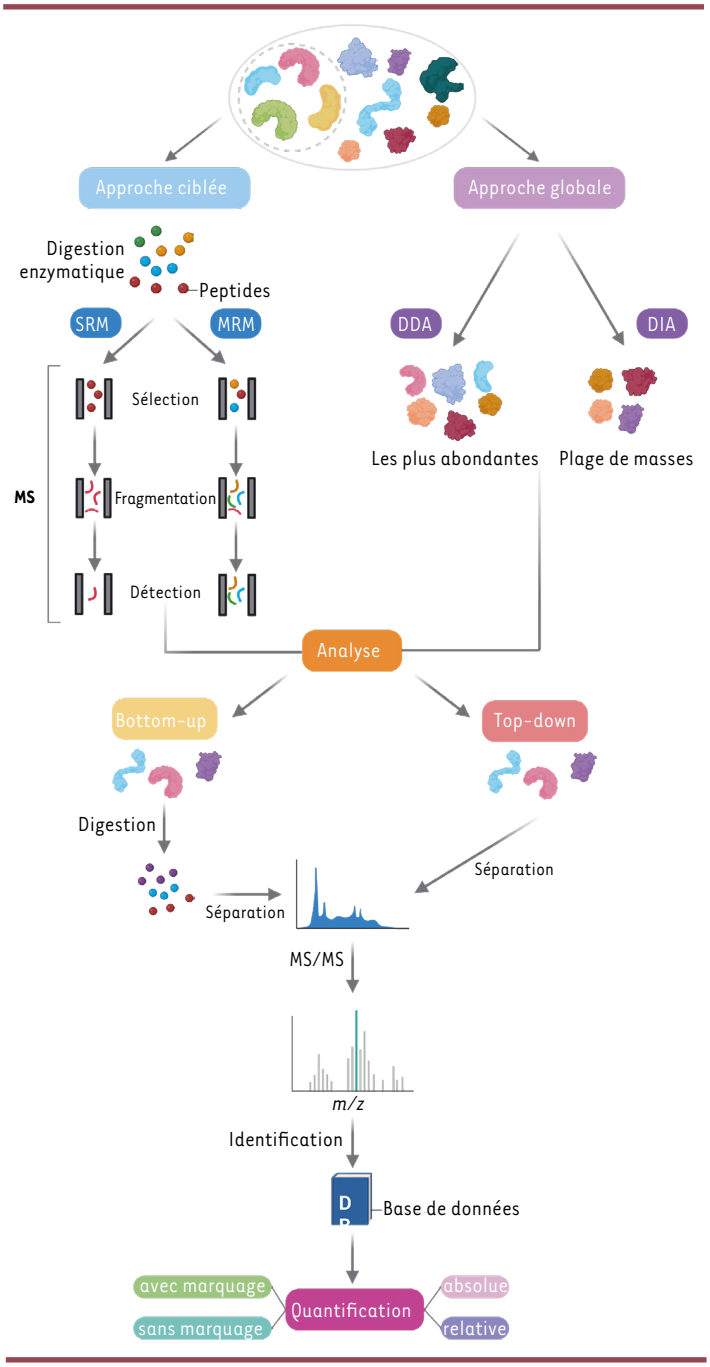

Figure 3. Principales stratégies d'acquisition et de quantification des protéines en spectrométrie de masse. Deux modes d'acquisition permettent une analyse globale : la DDA (*data-dependent acquisition*), qui cible les molécules les plus abondantes, et la DIA (*data-independent acquisition*), qui fragmente systématiquement tous les ions présents dans des fenêtres de masse prédéfinies pour une couverture plus complète. Les approches ciblées, telles que la SRM/MRM (*selected/multiple reaction monitoring*), sont privilégiées pour l'analyse de protéines connues, en raison de leur rapidité et de leur reproductibilité. Deux stratégies analytiques se distinguent : l'approche « *top-down* », appliquée aux protéines intactes, et l'approche « *bottom-up* », basée sur l'analyse de peptides issus de la digestion enzymatique. La quantification peut être réalisée sans marquage (comparaison directe entre échantillons) ou avec marquage isotopique, offrant une meilleure précision. Elle peut être relative ou absolue, cette dernière reposant sur des peptides de référence marqués (peptides AQUA). Figure réalisée avec BioRender.

la multiplicité des méthodes d'analyse et la complexité croissante des pipelines bio-informatiques exigent la mise en place de cadres d'évaluation rigoureux. La recherche en bio-informatique s'oriente alors vers une plus grande rigueur méthodologique. Cela se traduit par une attention portée à la reproductibilité des analyses, au suivi des versions des outils employés, et par l'usage de *pipelines* automatisés développés dans des langages comme Nextflow [30] ou Snakemake [31]. Ces *pipelines* s'appuient fréquemment sur des systèmes comme Docker [32] ou Singularity [33], permettant d'empaqueter l'ensemble des outils et de leurs dépendances[5] dans un conteneur isolé pouvant être exécuté sur n'importe quelle machine. En parallèle, l'accroissement des besoins en ressources de calcul a conduit au développement de centres de calcul haute performance (HPC, *high-performance computing*). Des plateformes, comme celles fournies par Amazon Web Services[6], Microsoft Azure[7] ou la SFBI (*Société française de bioinformatique*)[8], proposent des solutions d'analyse à grande échelle, avec un recours croissant au calcul sur carte graphique (GPU, *graphics processing unit*) et à l'apprentissage automatique pour traiter les grands jeux de données « omiques ».

Entre 2018 et 2021, le développement puis le déploiement d'AlphaFold par la société DeepMind a marqué un tournant majeur en biologie et particulièrement en protéomique, en intégrant l'intelligence artificielle comme outil pour le traitement de jeux de données « omiques » de plus en plus volumineux [*m/s* 34, *m/s* 35, 36]. Cet algorithme basé sur des réseaux de neurones profonds a révolutionné la prédiction de la structure tridimensionnelle des protéines à partir de leur seule séquence en acides aminés, en atteignant une précision comparable à celle des méthodes expérimentales, comme la cristallographie aux rayons X ou la cryo-microscopie électronique, tout en réduisant drastiquement le temps

[5] Une dépendance correspond à une composante externe à l'outil mais nécessaire à son bon fonctionnement.
[6] https://aws.amazon.com/
[7] https://azure.microsoft.com
[8] https://www.sfbi.fr/

et les coûts nécessaires [37, *m/s* 38]. L'ouverture publique de la base de données AlphaFold mis à disposition de la communauté scientifique des millions de structures protéiques prédites a accéléré considérablement la recherche en biologie structurale, en facilitant l'exploration fonctionnelle des protéines, ainsi que la compréhension de maladies liées à des mutations et le développement de nouvelles approches en conception de médicaments [35]. AlphaFold illustre ainsi l'impact transformateur des approches d'apprentissage automatique sur la bio-informatique moderne, en reliant directement séquence et fonction biologique. Toutefois, ses prédictions demeurent moins fiables pour certaines protéines, comme les complexes multidomaines ou les régions intrinsèquement désordonnées, rappelant l'importance de combiner modèles computationnels et validations expérimentales.

Quelques exemples d'applications

La génomique et les études d'association pangénomique

La génomique, entendue comme l'étude exhaustive du génome et de son organisation au-delà de la simple description des séquences des gènes, a profondément transformé les sciences biomédicales contemporaines, puisqu'elle vise à comprendre l'organisation, la fonction, et l'évolution du matériel génétique dans son intégralité. L'achèvement en 2003 du projet Génome humain issu du séquençage de 1 000 génomes a marqué un tournant décisif, en offrant une compréhension approfondie des variations génétiques entre les individus [39]. Cela a ouvert la voie à une médecine personnalisée, où le profil génétique d'une personne peut être utilisé pour établir un diagnostic précis, proposer des mesures préventives adaptées et choisir des traitements sur mesure. Ce tournant permet une approche de la santé beaucoup plus individualisée. Parmi les applications les plus marquantes, les études d'association pangénomique (GWAS, *genome-wide association studies*) occupent une place centrale. Ces analyses consistent à rechercher, sans hypothèse préalable, des associations statistiques entre des polymorphismes nucléotidiques simples (SNPs, *single nucleotide polymorphism*) et des traits complexes, tels que des maladies ou des caractéristiques phénotypiques. L'émergence des techniques de séquençage Illumina, puis Nanopore et PacBio, a largement soutenu ces avancées, notamment pour des génomes complexes où les lectures longues facilitent la résolution des régions répétées. Les GWAS ont d'abord été appliqués chez l'être humain, révélant en 2002 que le gène de la lymphotoxine-α serait associé à une susceptibilité à l'infarctus du myocarde [40]. Depuis, des milliers de locus ont été associés à des pathologies, comme dans le diabète de type 2 où une méta-analyse sur plus de 898 000 individus a mis en évidence 143 variants impliquant notamment *CAMK1D* et *TP53INP1* [41], ou dans la schizophrénie où 108 locus liés à la neurotransmission et au développement neuronal ont été identifiés [42]. Le consortium GIANT (*genetic investigation of anthropometric traits*) a par ailleurs montré que la taille humaine est influencée par plus de 12 000 variants regroupés dans 7 000 locus, expliquant environ 40 % de la variance phénotypique chez les Européens [43]. Certains GWAS ont également intégré une dimension infectieuse ; une étude portant sur plus de 50 000 patients a mis en évidence une association entre des variants génétiques de classe II du complexe majeur d'histocompatibilité HLA (*human leukocyte antigen*) et la susceptibilité à l'infection par *Staphylococcus aureus* [44]. L'approche a aussi été transposée au génome microbien : chez *S. aureus*, elle a permis d'associer des mutations du gène *mprF* à la résistance à la daptomycine [45], ainsi que des polymorphismes de la protéine foldase PrsA à des signatures d'adaptation en milieu hospitalier [46].

La transcriptomique

La transcriptomique est l'étude systématique de l'ensemble des ARN transcrits à partir du génome. À la différence de la séquence du génome, peu changeante, le transcriptome est dynamique et sensible aux conditions environnementales, offrant un aperçu de la régulation des gènes. Il inclut les ARNm mais aussi les ARN non codants, les ARN de transfert et les ARN ribosomiques, et reflète une étape intermédiaire de l'expression qui ne préjuge pas de la traduction protéique. Jusqu'aux années 2000, l'étude du transcriptome reposait principalement sur les puces à ADN, qui restaient limitées du fait de l'utilisation des sondes prédéfinies. L'avènement du séquençage à haut débit a conduit au développement des techniques de RNA-seq[9] et la première publication démontrant son efficacité concernait la levure *Saccharomyces cerevisiae* en 2008 [47]. D'autres études pilotes, publiées presque simultanément, concernaient le nématode, la souris et l'humain, confirmant le potentiel de cette technique [48, 49]. L'un des avantages majeurs du RNA-seq réside dans sa capacité à explorer la complexité de l'épissage alternatif, qui augmente considérablement la diversité protéique au-delà du nombre de gènes codants. Des travaux pionniers ont ainsi montré que plus de 90 % des gènes humains subissent un épissage alternatif, jouant un rôle crucial dans le développement, la différenciation cellulaire et la pathogenèse de multiples maladies [50, 51]. Par ailleurs, il a conduit à la découverte et à la caractérisation de nouveaux ARN non codants (lncRNA [long non-coding RNA], miRNA [micro RNA], circRNA [circular RNA]), essentiels dans la régulation fine de l'expression génique, apportant ainsi un éclairage nouveau sur la plasticité des réponses [52].

Aujourd'hui, la transcriptomique évolue vers des approches encore plus résolutives. Le *single-cell* RNA-seq (scRNA-seq) permet de caractériser l'expression génétique en cellule unique, révélant l'hétérogénéité jusque-là invisible au sein des tissus. Ceci a permis la cartographie des types cellulaires du cerveau humain ou encore la description des trajectoires de différenciation

[9] Technique de séquençage à haut débit qui permet d'identifier et de quantifier l'ensemble des ARN présents dans une cellule, un tissu ou un organisme à un moment donné.

des cellules immunitaires [53, 54]. Durant la pandémie de COVID-19, il a permis de caractériser l'hétérogénéité des réponses immunitaires et d'identifier les signatures spécifiques des cas sévères [55, 56]. Le RNA-seq spatial représente un autre développement récent, combinant données transcriptomiques et localisation histologique, afin de replacer les profils d'expression dans leur contexte anatomique. De nouvelles extensions incluent le *multi-omic single-cell* comme le CITE-seq (*Cellular indexing of transcriptomes and epitopes*), combinant transcriptome et protéome, ainsi que le RNA *velocity*, permettant de suivre la dynamique d'épissage des ARN [57].

La protéomique

La protéomique désigne l'étude systématique de l'ensemble des protéines exprimées dans une cellule, un tissu ou un organisme, dans un état donné. Elle se distingue par sa capacité à révéler des phénomènes invisibles aux autres approches « omiques », en apportant une analyse quantitative et qualitative des protéines réellement présentes dans la cellule et qui sont responsables du phénotype. Elle s'impose ainsi comme un pilier incontournable des sciences biomédicales modernes, en offrant une vision plus fonctionnelle et dynamique des processus biologiques complexes. En effet, l'abondance des ARNm ne reflète pas toujours celle des protéines correspondantes, en raison de la régulation post-transcriptionnelle, de la dégradation différentielle des protéines ou encore des modifications post-traductionnelles. Toutefois, la complexité des protéines, amplifiée par les isoformes produites par épissage alternatif, et les modifications post-traductionnelles, constitue un obstacle majeur à cette approche, car elle rend difficile une couverture exhaustive. De plus, la dynamique du protéome varie selon le type cellulaire, le stade de développement ou l'état pathologique, et les protéines de faible abondance, souvent cruciales, restent difficiles à détecter.

Les applications de la protéomique sont multiples, allant de la recherche fondamentale à la médecine translationnelle. Dès 2014, deux équipes ont publié une ébauche du protéome humain, couvrant plus de 18 000 protéines, et des bases de données, comme le « *human proteome project* » ambitionnent désormais de cataloguer l'ensemble des protéines humaines, à l'image du projet génome humain pour l'ADN [58, 59]. En cancérologie, le consortium CPTAC (*clinical proteomic tumor analysis consortium*) a caractérisé les signatures protéiques de nombreux types tumoraux (ovaire, sein [60, 61]), révélant des biomarqueurs diagnostiques et des cibles thérapeutiques potentielles. Dans les maladies neurodégénératives, la protéomique a permis d'identifier des agrégats protéiques pathologiques, tels que les accumulations de protéines Tau et d'α-Synucléine, associées respectivement aux maladies d'Alzheimer et de Parkinson [62, 63]. En microbiologie, une étude a montré que l'exposition de *P. aeruginosa* aux produits sécrétés par *S. aureus* entraîne des changements protéomiques majeurs, affectant la motilité et la sensibilité aux antibiotiques [64].

Les analyses protéomiques nécessitent des instruments coûteux et une expertise bio-informatique avancée pour l'interprétation des données. Mais aujourd'hui, la protéomique sur cellule unique, encore émergente, vise à surmonter la limite de l'hétérogénéité cellulaire et à analyser les protéines à l'échelle d'une seule cellule. En faisant face à ces défis techniques et analytiques, ces avancées ouvrent la voie à des applications cliniques de plus en plus tangibles, qu'il s'agisse de biomarqueurs diagnostiques, de stratification des patients ou de découverte de nouvelles cibles thérapeutiques.

Conclusion et perspectives

Ces cinquante dernières années ont été marquées par des évolutions et révolutions technologiques qui ont conduit à l'exploration massive des génomes et de leur expression. L'augmentation continue du volume de données générées soulève encore aujourd'hui des défis majeurs. La question du stockage, de la sauvegarde, du transfert de données ou du coût de calcul informatique pour le traitement de ces données reste problématique, d'autant plus dans un contexte d'attention croissante aux enjeux de développement durable. Par ailleurs, la normalisation des méthodes pour la génération de ces données, leur annotation ou leur traitement informatique et statistique reste essentielle afin de favoriser leur partage au sein de la communauté scientifique et leur exploitation à plus ou moins long terme et dans différents contextes scientifiques. C'est ainsi que les principes FAIR ont été définis pour favoriser le partage et l'exploitation optimale des nombreuses données générées. Ces contraintes rappellent que l'essor de la bio-informatique ne repose pas uniquement sur des progrès techniques, mais aussi sur la capacité à gérer la complexité scientifique et organisationnelle qu'ils engendrent. Ces avancées favorisent toutefois l'émergence d'une culture bio-informatique communautaire, avec des plateformes d'entraide et de documentation comme Stack Overflow[10], Biostars[11] ou encore des applications web interactives destinées aux biologistes non spécialistes de la bio-informatique. Enfin, cette période voit l'apparition de nombreuses disciplines regroupées sous le terme « omiques », telles que l'épigénomique, la métagénomique, la transcriptomique, ou la métabolomique, qui visent à modéliser des systèmes biologiques complexes en intégrant plusieurs niveaux d'information dans une approche systémique. En définitive, les approches « omiques » ne se réduisent plus à une simple accumulation de données, mais constituent désormais des outils pour répondre à des questions biologiques précises, élargissant ainsi considérablement la compréhension des systèmes biologiques. Un des défis majeurs à relever aujourd'hui est l'intégration de ces

[10] https://stackoverflow.com/
[11] https://www.biostars.org/

différentes approches pour générer une compréhension fine des mécanismes reliant les divers niveaux d'expression du génome. De nouvelles méthodes d'analyse doivent être imaginées pour cela.

Les approches « omiques » jouent également un rôle central dans la recherche biomédicale et la médecine en ouvrant la voie à des applications médicales personnalisées, et sont d'ores et déjà appliquées en oncologie, dans les maladies inflammatoires et auto-immunes, ou encore pour le développement de biomarqueurs diagnostiques. Pour autant, la causalité biologique des associations entre les analyses « omiques » identifiées et les maladies étudiées demeure difficile à établir sans une validation fonctionnelle, et les défis éthiques liés à leur utilisation en médecine prédictive ou en santé publique rappellent la nécessité d'un usage encadré. ◊

SUMMARY

Bioinformatics and omics analyses: chronology of a new scientific approach

From the early debates of the 1940s–1970s on the nature of the genetic information carrier — opposing proteins and DNA — to the rise of "omics" approaches now widely used to explore the complexity of genomes and their expression, molecular biology has undergone profound developments and true technological revolutions. These advances have enabled an increasingly precise and comprehensive understanding of living systems, while also creating new challenges related to the development of bioinformatics tools capable of analyzing ever-growing volumes of data. By providing unprecedented access to the organization and functioning of biological systems, they also pave the way for numerous applications, particularly in medical research. ◊

LIENS D'INTÉRÊT

Les auteurs déclarent ne pas avoir de lien d'intérêt.

RÉFÉRENCES

1. Edman P, Begg G. A protein sequenator. In : Liébecq C, ed. *European Journal of Biochemistry*. Berlin, Heidelberg : Springer Berlin Heidelberg, 1967 : 80–91.
2. Edman P. A method for the determination of amino acid sequence in peptides. *Arch Biochem* 1949 ; 22 : 475.
3. Dayhoff MO, Ledley RS. COMPROTEIN, a computer program to aid primary protein structure determination. *AFIPS (Fall)*1, 1962 : 262.
4. Goodman M, Moore GW. Phylogeny of hemoglobin. *Syst Zool*1973 ; 22 : 508.
5. Gauthier J, Vincent AT, Charette SJ, *et al.* A brief history of bioinformatics. *Brief Bioinform* 2019 ; 20 : 1981–96.
6. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Bio*1970 ; 48 : 443–53.
7. Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 1988 ; 73 : 237–44.
8. O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* 1975 ; 250 : 4007–21.
9. Watson JD, Crick FHC. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 1953 ; 171 : 737–8.
10. Kahn A. L'hélice de la vie. *Med Sci (Paris)* 2003 ; 19 : 491–5.
11. Crick FHC. The origin of the genetic code. *J Mol Biol* 1968 ; 38 : 367–9.
12. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U.S.A.* 1977 ; 74 : 5463–7.
13. Sanger F, Air GM, Barrell BG, *et al.* Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* 1977 ; 265 : 687–95.
14. Christiansen T, Wall L, Orwant J. *Programming Perl: Unmatched power for text processing and scripting*. O'Reilly Media Inc, 4th edition. 2012 ; 1184 p.
15. Kernighan BW, Ritchie DM. *The C programming language*. 28. print. Englewood Cliffs, NJ : Prentice-Hall, 1991 : 228 p.
16. Stoesser G, Sterk P, Tuli MA, *et al.* The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* 1997 ; 25 : 7–13.
17. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res* 1997 ; 25 : 31–6.
18. Camacho C, Coulouris G, Avagyan V, *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009 ; 10 : 421.
19. Benson DA, Boguski MS, Lipman DJ, *et al.* GenBank. *Nucleic Acids Res* 1997 ; 25 : 1–6.
20. Karsch-Mizrachi I, Takagi T, Cochrane G, *et al.* The international nucleotide sequence database collaboration. *Nucleic Acids Res* 2018 ; 46 : D48–51.
21. Margulies M, Egholm M, Altman WE, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005 ; 437 : 376–80.
22. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016 ; 17 : 333–51.
23. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. *MolCell* 2015 ; 58 : 586–97.
24. Montel F. Séquençage de l'ADN par nanopores: Résultats et perspectives. *Med Sci (Paris)* 2018 ; 34 : 161– 5.
25. Whitehouse CM, Dreyer RN, Yamashita M, *et al.* Electrospray interface for liquid chromatographs and mass spectrometers. *Anal Chem* 1985 ; 57 : 675–9.
26. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003 ; 422 : 198–207.
27. Leinonen R, Sugawara H, Shumway M, *et al.* The sequence read archive. *Nucleic Acids Res* 2011 ; 39 : D19–21.
28. Leinonen R, Akhtar R, Birney E, *et al.* The european nucleotide archive. *Nucleic Acids Res* 2011 ; 39 : D28–31.
29. Wilkinson MD, Dumontier M, Aalbersberg IjJ, *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016 ; 3 : 160018.
30. Di Tommaso P, Chatzou M, Floden EW, *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017 ; 35 : 316– 9.
31. Mölder F, Jablonski KP, Letcher B, *et al.* Sustainable data analysis with Snakemake. *F1000Res* 2021 ; 10 : 33.
32. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J* 2014 ; 2.
33. Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. *PLoS ONE* 2017 ; 12 : e0177459.
34. Taly A, Verger A. Prédiction de structures biomoléculaires complexes par *AlphaFold 3*. *Med Sci (Paris)* 2024 ; 40 : 725–7.
35. Jordan B. AlphaFold : un pas essentiel vers la fonction des protéines : Chroniques génomiques. *Med Sci (Paris)* 2021 ; 37 : 197–200.
36. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021 ; 596 : 583– 9.
37. Varadi M, Anyango S, Deshpande M, *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022 ; 50 : D439–44.
38. Aguéro-Pizzolo S, Bettler E, Gouet P. Prix Nobel de chimie 2024: David Baker, Demis Hassabis et John M. Jumper - La révolution de l'intelligence artificielle en biologie structurale. *Med Sci (Paris)* 2025 ; 41 : 367– 73.
39. International human genome sequencing consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004 ; 431 : 931– 45.
40. Ozaki K, Ohnishi Y, Iida A, *et al.* Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat Genet* 2002 ; 32 : 650–4.
41. Mahajan A, Taliun D, Thurner M, *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet* 2018 ; 50 : 1505– 13.
42. Schizophrenia working group of the psychiatric genomics consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014 ; 511 : 421– 7.
43. Yengo L, Vedantam S, Marouli E, *et al.* A saturated map of common genetic variants associated with human height. *Nature* 2022 ; 610 : 704– 12.
44. Moreau K, Clemenceau A, Le Moing V, *et al.* Human genetic susceptibility to native valve staphylococcus aureus endocarditis in patients with S. aureus bacteremia: genome-wide association study. *Front Microbiol* 2018 ; 9.
45. Weber RE, Fuchs S, Layer F, *et al.* Genome-wide association studies for the detection of genetic variants associated with daptomycin and ceftaroline resistance in Staphylococcus aureus. *Front Microbiol* 2021 ; 12 : 639660.
46. Young BC, Wu C-H, Charlesworth J, *et al.* Antimicrobial resistance determinants are associated with Staphylococcus aureus bacteraemia and

adaptation to the healthcare environment: a bacterial genome-wide association study. *Microb Genom* 2021 ; 7 : 000700.
47. Nagalakshmi U, Wang Z, Waern K, *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008 ; 320 : 1344– 9.
48. Mortazavi A, Williams BA, McCue K, *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008 ; 5 : 621– 8.
49. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009 ; 10 : 57–63.
50. Pan Q, Shai O, Lee LJ, *et al.* Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008 ; 40 : 1413–5.
51. Sultan M, Schulz MH, Richard H, *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 2008 ; 321 : 956– 60.
52. Cabili MN, Trapnell C, Goff L, *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 2011 ; 25 : 1915–27.
53. Chen X, Huang Y, Huang L, *et al.* A brain cell atlas integrating single-cell transcriptomes across human brain regions. *Nat Med* 2024 ; 30 : 2679– 91.
54. Wang Y, Li R, Tong R, *et al.* Integrating single-cell RNA and T cell/B cell receptor sequencing with mass cytometry reveals dynamic trajectories of human peripheral immune cells from birth to old age. *Nat Immunol* 2025 ; 26 : 308– 22.
55. Wilk AJ, Rustagi A, Zhao NQ, *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat Med* 2020 ; 26 : 1070– 6.
56. Ren X, Wen W, Fan X, *et al.* COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* 2021 ; 184 : 1895- 913.
57. La Manno G, Soldatov R, Zeisel A, *et al.* RNA velocity of single cells. *Nature* 2018 ; 560 : 494– 8.
58. Kim M-S, Pinto SM, Getnet D, *et al.* A draft map of the human proteome. *Nature* 2014 ; 509 : 575– 81.
59. Wilhelm M, Schlegl J, Hahne H, *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* 2014 ; 509 : 582– 7.
60. Krug K, Jaehnig EJ, Satpathy S, *et al.* Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell* 2020 ; 183 : 1436- 56..
61. McDermott JE, Arshad OA, Petyuk VA, *et al.* Proteogenomic characterization of ovarian HGSC implicates mitotic kinases, replication stress in observed chromosomal instability. *Cell Rep Med* 2020 ; 1 : 100004.
62. Sarkar S, Murphy MA, Dammer EB, *et al.* Comparative proteomic analysis highlights metabolic dysfunction in α-synucleinopathy. *NPJ Parkinsons Dis* 2020 ; 6 : 40.
63. Yarbro JM, Shrestha HK, Wang Z, *et al.* Proteomic landscape of Alzheimer's disease: emerging technologies, advances and insights (2021 – 2025). *Mol Neurodegeneration* 2025 ; 20 : 83.
64. Reigada I, San-Martin-Galindo P, Gilbert-Girard S, *et al.* Surfaceome and Eexoproteome dynamics in dual-species Pseudomonas aeruginosa and Staphylococcus aureus Biofilms. *Front Microbiol* 2021 ; 12 : 672975.

TIRÉS À PART
K. Moreau