L'intelligence artificielle pour prédire les structures des biopolymères

La fonction d'une protéine est étroitement liée à la façon dont ses atomes s'organisent dans l'espace, et le plus souvent à son association avec d'autres biomolécules, qu'il s'agisse d'autres protéines, d'acides nucléiques ou de petits ligands. Malgré les progrès techniques, la résolution expérimentale des structures tridimensionnelles de protéines ou d'assemblages macromoléculaires reste longue et difficile. En conséquence, le nombre de structures de macromolécules biologiques résolues expérimentalement (de l'ordre de 160 000) reste inférieur de plusieurs ordres de grandeur au nombre de séquences protéiques annotées dans les bases de données (plus de 200 millions).

Pourquoi est-il si important de connaître la structure tridimensionnelle des protéines ? D'abord, la structure est un outil pour les biologistes qui cherchent à disséquer le fonctionnement de certaines machineries protéiques : grâce à la connaissance ou à la prédiction de la structure des protéines et de leurs assemblages, on peut concevoir des mutations qui vont perturber spécifiquement certaines de leurs fonctions. La structure est également particulièrement intéressante pour le développement d'approches thérapeutiques, car elle donne les clés pour concevoir des composés synthétiques (molécules organiques ou petits biopolymères) capables de cibler spécifiquement une fonction ou une interaction. Un exemple emblématique de succès de la conception de médicaments basée sur la structure concerne plusieurs molécules inhibitrices de la protéase du VIH qui ont été autorisées comme traitement contre le SIDA dans les années 1990 [1].

Un problème de prédiction difficile

Les travaux du biochimiste Christian Anfinsen, lauréat du prix Nobel de chimie en 1972, ont suggéré dès les années 1960 que toute l'information nécessaire pour déterminer la structure d'une protéine est encodée dans sa séquence en acides aminés [2]. Partant de ce postulat fondamental de la biologie moléculaire, quoi de plus naturel que de tenter de déterminer le code qui permet, à partir de la séquence en acides aminés d'une protéine, de prédire comment celle-ci va se replier dans l'espace ?

Ce problème de prédiction de la structure tridimensionnelle des protéines s'est cependant avéré extrêmement difficile. Pendant plusieurs décennies, de nombreuses méthodes ont été développées pour tenter d'y répondre, avec un succès mitigé. Idéalement, on souhaiterait pouvoir déterminer la façon dont la protéine se replie en n'utilisant que les principes fondamentaux de la physique. Cependant, malgré de grandes avancées, liées en particulier à la conception de super-ordinateurs dédiés, ces méthodes dites *ab initio* restent à l'heure actuelle trop coûteuses pour être généralisables à l'échelle de toutes les séquences protéiques connues. Des approximations ou des heuristiques doivent donc être utilisées. Une autre grande famille d'approches utilise les structures déterminées expérimentalement comme point de départ pour prédire de nouvelles structures. C'est le cas par exemple des méthodes par fragments, qui assemblent *in silico* de courtes portions de structure connue, utilisées comme des briques de construction, mais aussi d'autres stratégies qui s'appuient sur les propriétés évolutives des protéines [3].

Le pouvoir de l'information évolutive

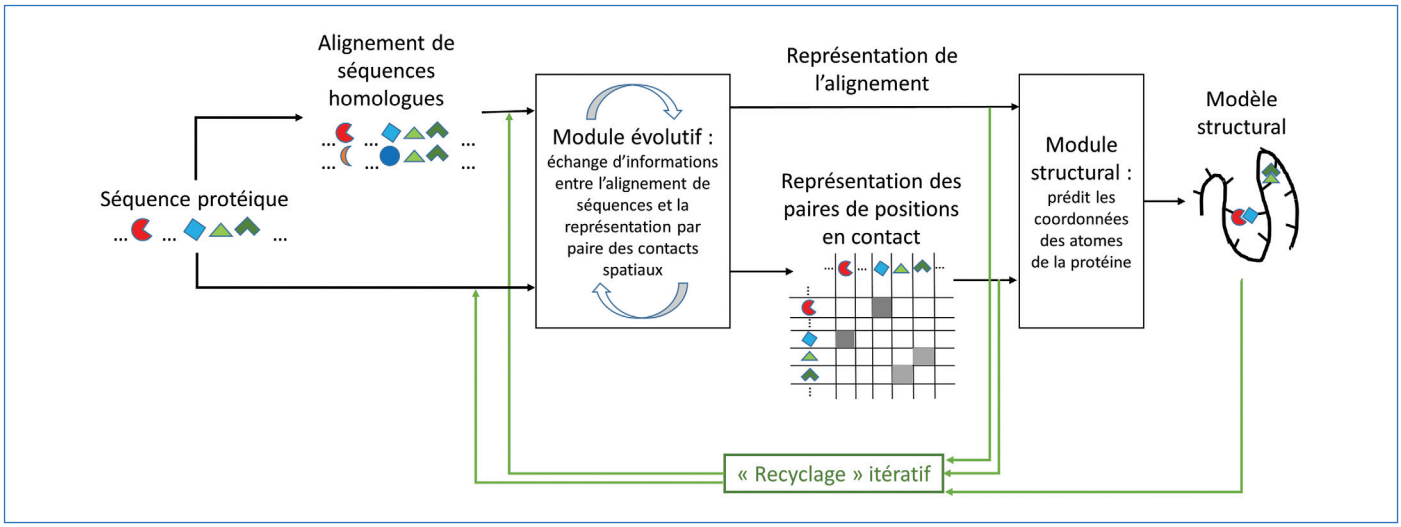
Les protéines subissent une pression évolutive liée à leur fonction. En pratique, des contraintes différentes s'appliquent aux positions de la séquence protéique en fonction de leur implication dans la structure tridimensionnelle de la protéine ou dans un assemblage. Par exemple, les positions enfouies au cœur de la structure, dont la modification pourrait induire une déstabilisation du repliement, sont plus conservées au cours de l'évolution que les positions de surface. Ces contraintes évolutives créent une empreinte qui permet de détecter les protéines ayant évolué à partir d'une même protéine ancestrale (on parle de protéines homologues) ; ces protéines ayant une origine évolutive commune adoptent une structure similaire, même si leur séquence en acides aminés a divergé. Par ailleurs, toute protéine subit des mutations au gré de l'évolution, mais les positions de la séquence qui sont en contact physique dans la structure vont être modifiées de façon corrélée afin de maintenir cette structure. Ainsi, si une position de la séquence protéique change de nature chimique, un changement à une position voisine va pouvoir créer une compensation permettant de conserver la structure tridimensionnelle. Par exemple, si un acide aminé chargé dans une protéine devient hydrophobe dans une protéine homologue, un acide aminé voisin de charge opposée peut lui aussi devenir hydrophobe pour maintenir un contact.

Ces contraintes évolutives sont utilisables pour détecter les positions importantes pour la structuration d'une protéine, et même prédire les contacts formant sa structure tridimensionnelle, en considérant la variation corrélée des positions en contact au cours de l'évolution. Ces principes de co-variation sont connus depuis plusieurs dizaines d'années, mais l'efficacité de leur mise en application n'a été démontrée que dans la dernière décennie, grâce à l'abondance d'information de séquence disponible et à l'utilisation de méthodes statistiques empruntées à la physique, puis de méthodes d'apprentissage automatique.

Le défi international bisannuel CASP (« critical assessment of protein structure prediction ») permet aux équipes de recherche développant des méthodes de prédiction de se confronter à l'aveugle à des cibles protéiques dont la structure expérimentale n'est pas encore publique. Une catégorie particulièrement intéressante de cibles CASP concerne les protéines ne ressemblant à aucune structure connue expérimentalement, ou pour lesquelles le lien évolutif avec une structure connue est trop ténu pour être détecté. En 2016, la prédiction de ces cibles a connu un progrès important grâce à la combinaison des principes de co-variation et des méthodes d'apprentissage automatique [4].

L'avènement de l'apprentissage automatique et profond

Depuis, les méthodes d'apprentissage automatique ont permis de faire des bonds de géant dans la performance des prédictions de structure des protéines. Elles ont pour cela dû relever plusieurs défis, dont le choix des modes de représentation de ces biopolymères complexes et la relative rareté des données d'apprentissage par rapport à d'autres domaines comme la vision par ordinateur ou la traduction automatique, où des millions d'exemples sont disponibles. En 2018, CASP a connu une première révolution avec l'entrée en jeu de l'algorithme AlphaFold développé par l'entreprise DeepMind, qui a démontré sa supériorité sur toutes les méthodes développées par la communauté académique [5]. Fin 2020, la nouvelle version d'AlphaFold a atteint des performances meilleures encore, très proches de celles des méthodes expérimentales [6]. En seulement quelques années, l'entreprise a réussi à développer des architectures originales de réseaux de neurones profonds où chaque protéine est représentée sous forme d'un graphe, dont les nœuds sont les résidus d'acides aminés et les arêtes connectent les résidus proches dans l'espace (voir *figure*). Le modèle est entraîné sur les structures de protéines résolues expérimentalement. Une approche itérative permet d'améliorer progressivement les prédictions et les



Représentation schématique simplifiée du fonctionnement de l'algorithme AlphaFold. Pour prédire la structure d'une protéine, on part de sa séquence pour construire un alignement de séquences homologues. Cet alignement est utilisé par un premier module fondamental d'apprentissage profond pour produire deux représentations : une de l'alignement et une des paires de positions en contact spatial (en niveaux de gris). Ces deux représentations sont utilisées par un second module d'apprentissage profond qui prédit les coordonnées des atomes de la protéine. Les représentations et les coordonnées prédites sont réinjectées dans le premier module de façon itérative pour améliorer les prédictions. D'après [6, *figure 1*] utilisée sous licence CC-BY-4.0 (https://creativecommons.org/licenses/by/4.0).

modèles structuraux obtenus s'accompagnent d'une estimation de confiance à l'échelle de chaque acide aminé.
L'information évolutive, fournie à AlphaFold sous forme d'un alignement de séquences homologues, reste clé pour la prédiction, mais elle est exploitée de façon extrêmement efficace, de sorte que l'applicabilité a bondi : les prédictions de structure sont accessibles même pour des protéines pour lesquelles on ne peut identifier que quelques dizaines de séquences ayant la même origine évolutive. Ainsi, la structure de la protéine ORF8 du SARS-CoV-2, qui constituait une cible CASP extrêmement difficile, a été très bien prédite par AlphaFold malgré l'identification de seulement 24 séquences homologues [7].
Ces résultats s'appuient sur des dizaines d'années de travail, à la fois pour la résolution expérimentale de structures, sans lesquelles aucun apprentissage ne serait possible, et pour la progression méthodologique. D'autres méthodes prédictives continuent d'être développées par la communauté académique, sans atteindre pour l'instant la précision d'AlphaFold [8].

Vers la prédiction des assemblages de biopolymères

Depuis juillet 2021, la méthode AlphaFold est disponible publiquement sous forme d'un programme de prédiction et d'une base de données contenant les structures prédites pour les protéines d'une vingtaine d'espèces particulièrement étudiées, dont l'humain [9]. Ces modèles structuraux vont avoir de nombreuses applications en biologie. Du côté des développements méthodologiques, la résolution de la prédiction de la structure des protéines ouvre la voie à des questions plus difficiles : prédiction de structure des autres biopolymères, tels que les acides nucléiques, aspects dynamiques du repliement, conception et ingénierie de biopolymères. La prédiction structurale des assemblages de biopolymères peut à la fois directement exploiter les modèles disponibles pour les molécules individuelles, utilisés comme briques de construction, et bénéficier elle-même d'une prise en compte soigneuse de l'information évolutive [10-11]. Ce domaine est donc aussi particulièrement susceptible de bénéficier de la puissance de ces avancées. Après des travaux de la communauté montrant une capacité insoupçonnée d'AlphaFold à prédire les structures d'assemblages protéiques sans les avoir explicitement intégrées à l'apprentissage [12], un modèle en cours de développement nommé AlphaFold-Multimer, spécialisé pour les assemblages, montre des performances encore plus impressionnantes [13]. Les applications futures pourraient inclure la prédiction des complexes de protéines avec des petites molécules susceptibles d'être utilisées comme médicaments. Enfin, plus généralement, les avancées liées aux méthodes d'apprentissage profond touchent également d'autres domaines de la chimie et de la biologie, tels que l'interprétation d'images pour le diagnostic médical ou encore la conception de nouveaux matériaux.

[1] A. Wlodawer, J. Vondrasek, Inhibitors of HIV-1 protease: a major success of structure-assisted drug design, *Ann. Rev. Biophys. Biomol. Struct.*, **1998**, *27*, p. 249-284.
[2] C. Anfinsen, Nobel lecture, **1972**, www.nobelprize.org/prizes/chemistry/1972/anfinsen/lecture
[3] V.A. Jisna, P.B. Jayaraj, Protein structure prediction: conventional and deep learning perspectives, *Protein J.*, **2021**, *40*, p. 522-544.
[4] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, A.M. Bonvin, Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age, *Proteins*, **2018**, *86*, sup. 1, p. 51-66.
[5] A.W. Senior, R. Evans *et al.*, Improved protein structure prediction using potentials from deep learning, *Nature*, **2020**, *577*, p. 706-710.
[6] J. Jumper, R. Evans, D. Hassabis *et al.*, Highly accurate protein structure prediction with AlphaFold, *Nature*, **2021**, *596*, p. 583-589.
[7] J. Jumper, R. Evans *et al.*, Applying and improving AlphaFold at CASP14, *Proteins*, **2021**, *89*, p. 1711-21.
[8] M. Baek, F. DiMaio *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network, *Science*, **2021**, *373*, p. 871-876.
[9] K. Tunyasuvunakool, J. Adler, D. Hassabis *et al.*, Highly accurate protein structure prediction for the human proteome, *Nature*, **2021**, *596*, p. 590-596.
[10] J. Andreani, C. Quignot, R. Guerois. Structural prediction of protein interactions and docking using conservation and coevolution, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **2020**, *10*, e1470.
[11] C. Quignot, P. Granger, B. Chacon, R. Guerois, J. Andreani, Atomic-level evolutionary information improves protein-protein interface scoring, *Bioinformatics*, **2021**, *37*, p. 3175-81.
[12] I.R. Humphreys, J. Pei *et al.*, Computed structures of core eukaryotic protein complexes, *Science*, **2021**, *374*, http://doi.org/doi:10.1126/science.abm4805
[13] R. Evans, M. O'Neill *et al.*, Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, **2021**, http://doi.org/10.1101/2021.10.04.463034

Cette fiche a été réalisée par **Jessica ANDREANI**, chercheure, I2BC, Institut des Sciences du Vivant Frédéric Joliot, CEA, Gif-sur-Yvette (jessica.andreani@cea.fr).

Les fiches « Un point sur » sont coordonnées par Jean-Pierre FOULON (jpfoulon@wanadoo.fr). Elles sont regroupées et en téléchargement libre sur www.lactualitechimique.org.