médecine/sciences 2021 ; 37 : 197-200

Chroniques génomiques

AlphaFold
Un pas essentiel vers la fonction des protéines

Bertrand Jordan

UMR 7268 ADÉS, Aix-Marseille, Université/EFS/CNRS ;
CoReBio PACA, case 901,
Parc scientifique de Luminy,
13288 Marseille Cedex 09,
France.
brjordan@orange.fr

De la séquence à la structure tridimensionnelle

Les protéines sont un élément essentiel de la mécanique de la vie. Elles assurent les fonctions très diverses nécessaires à la vie d'un organisme. Leur structure est longtemps restée nébuleuse, mais le concept d'une chaîne linéaire d'acides aminés s'est progressivement imposé et a été définitivement prouvé par la première détermination d'une séquence protéique, celle de l'insuline achevée par Fred Sanger en 1952 [1]. Dans les années qui ont suivi, on a pris conscience que le repliement de cette chaîne, sa structure secondaire et tertiaire, était essentiel pour assurer la fonction de chaque protéine, gouverner ses interactions avec d'autres molécules et, parfois, former un site actif capable de catalyser une réaction chimique. La détermination de la structure tridimensionnelle des protéines (une fois connue leur séquence) devenait dès lors un objectif essentiel. Il fut atteint pour la première fois par l'équipe de Kendrew avec la publication, en 1958, de la structure tridimensionnelle de la myoglobine, obtenue par diffraction aux rayons X [2].

La voie était ainsi ouverte pour d'autres protéines – mais le chemin était long, difficile et le succès n'était pas garanti : il fallait disposer de la protéine pure en quantité importante (de l'ordre du gramme), parvenir à en dériver des cristaux stables et de taille suffisante, mesurer l'image de diffraction des rayons X par un tel cristal *(Figure 1)*, et calculer à partir de ces données la disposition dans l'espace des atomes de la protéine. Cette dernière étape nécessitait en général l'obtention de nouveaux cristaux à partir d'une forme modifiée de la protéine portant un atome lourd en un point bien défini, nécessaire pour choisir entre plusieurs structures possibles d'après le spectre de diffraction de la protéine non modifiée. Au total, l'entreprise durait des années, pouvait parfaitement échouer (certaines protéines ne cristallisent pas, ou forment des cristaux qui ne résistent pas à l'irradiation X), et exigeait des moyens de calcul à la limite des possibilités de l'époque. Des techniques plus rapides, la résonance magnétique nucléaire (RMN) [3] pour les peptides et les petites protéines, et plus récemment la cryomicroscopie électronique [4] pour les plus grosses protéines et les assemblages multimoléculaires ont permis des progrès significatifs mais encore très insuffisants face à l'avalanche de nouvelles séquences. L'irruption des techniques du génie génétique a permis en effet, à partir des années 1970, de « cloner » (isoler) et de séquencer de très nombreux gènes. La laborieuse obtention de la séquence d'une protéine était court-circuitée par le séquençage de plus en plus rapide du gène correspondant, et l'on se retrouvait ainsi en face de très nombreuses protéines de séquence connue… mais de fonction inconnue. La connaissance de leur structure dans l'espace, essentielle à leur fonction, devenait un enjeu majeur. Pour fixer les idées, on répertorie actuellement la séquence en acides aminés d'environ deux millions de protéines, mais on ne connaît la structure tridimensionnelle que pour 170 000 d'entre elles [5]. C'est déjà beaucoup, mais c'est à l'évidence insuffisant.

Une méthode qui permettrait de déterminer la structure tertiaire d'une protéine à partir de sa séquence primaire serait donc essentielle et constituerait un grand pas vers l'élucidation de la fonction de toutes ces nouvelles protéines révélées par le séquençage d'ADN. Cela devrait être possible puisque le repliement dans l'espace de la plupart des protéines se fait

spontanément lors de leur synthèse par les ribosomes qui lisent l'ARN messager[1] [6] (➜). Certaines d'entre elles peuvent même être dénaturées, par exemple par la chaleur ; elles se renaturent spontanément en retrouvant leur activité lorsqu'on refroidit la solution (cette démonstration valut son prix Nobel à Christian Anfinsen en 1961 [7]). Dès les années 1970, plusieurs équipes publièrent des programmes informatiques supposés effectuer ce type de prédiction [8], mais les résultats furent très décevants, aucun de ces algorithmes ne s'avérant avoir une réelle validité prédictive. Il allait falloir attendre le tournant du siècle pour voir apparaître des prédictions utilisables.

(➜) Voir la Nouvelle de R. Barouki, *m/s* n° 12, décembre 2002, page 1200

Une course entre « modéliseurs »

Pour aider à structurer la communauté intervenant dans ce domaine de recherche, quelques universitaires, rassemblés par John Moult (Université du Maryland), mirent sur pied au début des années 1990 un consortium appelé CASP (*Critical Assessment of protein Structure Prediction*)[2] destiné à comparer objectivement la qualité des algorithmes proposés par différentes équipes. Ce consortium [9] organise tous les deux ans une session pour laquelle quelques dizaines de séquences d'acides aminés sont fournies aux équipes participantes, qui doivent chacune au cours des semaines suivantes fournir les structures tridimensionnelles prédites par leur méthode. En fait, ces structures ont déjà été déterminées expérimentalement mais ne sont pas encore publiées ; elles sont naturellement inconnues des équipes participantes. À l'arrivée, on compare chaque structure théorique avec sa correspondante expérimentale, et on en déduit notamment le score d'un test de distance globale (le GDT pour *global distance test*) qui incorpore les distances, pour chaque atome, entre sa position déduite du modèle et celle indiquée par l'expérience. Un score de 100 indique une identité parfaite entre les deux structures et, compte tenu des incertitudes expérimentales, un score de 90 indique une prédiction équivalente à une détermination expérimentale. La première rencontre, CASP1, a eu lieu en 1994. Elle a impliqué 35 équipes qui ont fourni leurs prédictions pour 33 protéines. Les scores obtenus allaient de 40 ou 50 pour les protéines « faciles » (de petite taille) à 20 pour les protéines « difficiles ». Au fil des années, ces performances se sont améliorées *(Figure 2)* mais sans atteindre le graal d'une équivalence avec les données expérimentales, sauf pour les protéines les plus faciles.

C'est en 2018 qu'apparut, parmi les équipes participantes, une entreprise appelée *DeepMind*, une start-up britannique spécialisée dans les applications de l'intelligence artificielle, fondée en 2010 et rachetée par *Google* en 2017 tout en gardant son identité. La firme est connue pour son expertise en intelligence artificielle appliquée au raisonnement relationnel et prédictif, et son programme AlphaGo a battu les meilleurs joueurs mondiaux au jeu de Go en 2017 [10]. *DeepMind* a

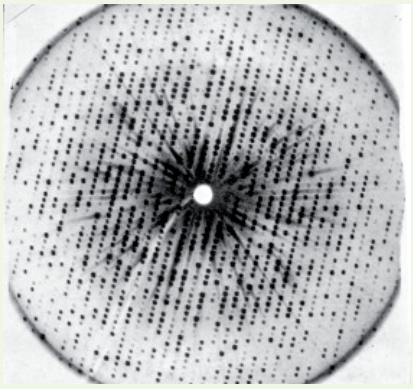

Figure 1. ***Image de diffraction de rayons X par un cristal de myoglobine*** (MRC *Laboratory of Molecular Biology*, Cambridge, GB).

mis au point un algorithme appelé Alphafold[3] qui, dès la session de 2018 de CASP (CASP13) s'est avéré plus performant que tous les systèmes rivaux, fournissant des prédictions dont le score GDT était supérieur d'une quinzaine de points à celui des autres [9]. L'édition 2020 de CASP (CASP14) a vu le triomphe de la nouvelle version d'AlphaFold [11], salué par un communiqué de presse du consortium CASP[4]. AlphaFold 2 *(Figure 3)* atteint un score d'environ 90, même pour des protéines « difficiles », surclassant très nettement tous ses concurrents, parmi lesquels on compte, outre de nombreux laboratoires universitaires, *Microsoft* et l'entreprise technologique chinoise *Tencent*. Mais le plus important est que, pour plus des deux tiers de la centaine de protéines testées au cours de CASP14, le score de 90 atteint par AlphaFold indique que cette prédiction est équivalente à une structure tridimensionnelle déterminée de manière expérimentale après des mois sinon des années de travail acharné...

Qu'y a-t-il sous le capot ?

Quel est le secret des performances d'AlphaFold ? *A priori* rien de vraiment exceptionnel. Comme beaucoup d'autres, cet algorithme fait un large usage des bases de données renfermant les structures tridimensionnelles de dizaines de milliers de protéines, et utilise des approches d'apprentissage (*deep learning*) pour améliorer ses capacités de prédiction. Le détail de la méthode sera publié (c'est une des conditions de la participation à CASP) ; pour le moment on ne dispose

[1] Dans quelques cas, une autre protéine appelée « protéine chaperonne » aide à donner sa forme à la protéine néo-synthétisée [6].

[2] https://predictioncenter.org/

[3] https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

[4] https://predictioncenter.org/casp14/doc/CASP14_press_release.html

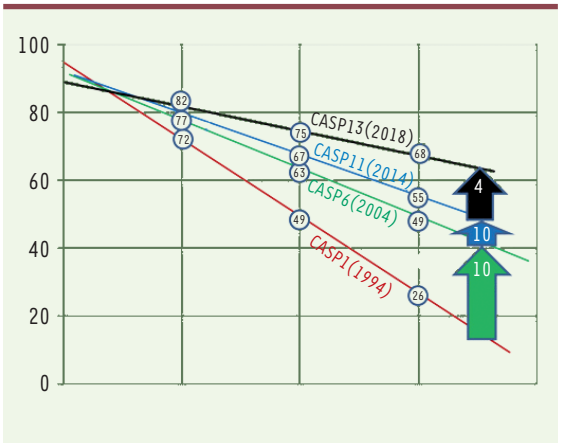

Figure 2. ***Meilleures performances (score GDT) lors des CASP successifs, de 1994 à 2018.*** Échelle verticale : score GDT ; échelle horizontale : difficulté de la prédiction, protéines « faciles » à gauche. Les chiffres dans les cercles indiquent le score médian pour l'édition de CASP et la classe de protéines considérées. Les chiffres dans les flèches sont exprimés en années. Extrait partiel et remanié de la page d'accueil du site CASP (https://predictioncenter.org).

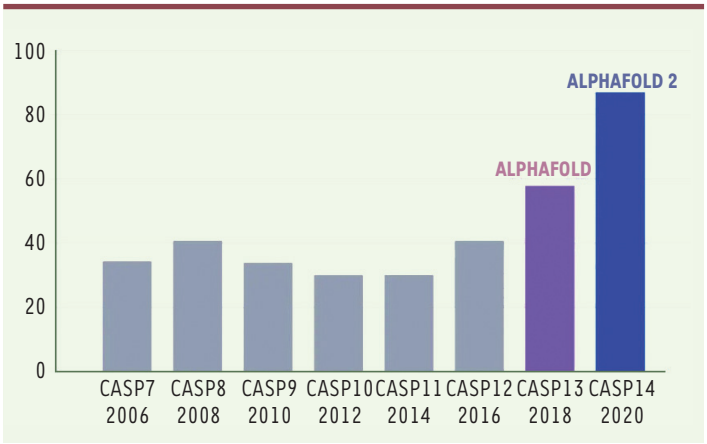

Figure 3. ***Échelle verticale : score GDT médian*** (pour l'ensemble des concurrents et des structures prédites) de 2006 (CASP7) à 2016 (CASP12), et score médian pour les prédictions d'AlphaFold en 2018 et 2020. Le plafonnement des scores médians entre 2006 et 2016 résulte du choix de séquences de plus en plus « difficiles » pour les CASP successives, alors même que des progrès significatifs étaient enregistrés (voir *Figure 2*). Documentation DeepMind (https://deepmind.com/research/case-studies/alphaFold).

que d'un *abstract* (un résumé) très cryptique pour les non-spécialistes, publié en préparation de CASP14 [12] et dont vingt-neuf des trente auteurs appartiennent à l'entreprise *DeepMind*. La puissance de calcul mise en œuvre n'est pas gigantesque : un *cluster* de 128 microprocesseurs travaillant en parallèle, un ensemble performant mais qui n'a rien d'un supercalculateur, détermine en quelques heures la structure tridimensionnelle d'une protéine. C'est le talent des concepteurs de l'algorithme qui a fait la différence, la manière de combiner les informations tirées des bases de données, l'appartenance ou la parenté avec une famille de protéines connue, la façon de combiner interactions locales et relations à distance, sans oublier la conception du système d'apprentissage permettant à l'algorithme de se perfectionner (en prédisant la structure de protéines déjà connues). Il faut dire que l'équipe avait déjà montré son expertise avec le jeu de Go [10] ...

Une avancée de premier ordre (et un plan social pour les cristallographes ?)

Le succès d'AlphaFold implique-t-il l'abandon des approches expérimentales ? Non, bien sûr, mais il va profondément modifier le paysage. On va pouvoir définir rapidement, et à peu de frais, la structure tridimensionnelle probable de toute nouvelle protéine, ce qui servira de guide aux travaux visant à préciser tel ou tel point, à étudier les changements induits par la formation de complexes multimoléculaires ou encore par l'insertion de la protéine dans une membrane. C'est probablement la résonance magnétique nucléaire et la cryomicroscopie qui seront le plus mises à contribution ; la cristallographie classique, en revanche, va beaucoup perdre de son importance, et la difficile obtention de cristaux d'une protéine ne sera plus le passage obligé pour accéder à sa structure dans l'espace.

Mais prenons un peu de recul. La capacité de déduire la forme d'une protéine de sa séquence en acides aminés constitue une avancée fondamentale, comparable en importance à la découverte de la double hélice de l'ADN en 1953. On savait depuis cinquante ans qu'il devait être possible de faire une telle prédiction (puisque la plupart des protéines se replient toutes seules), mais la complexité de ces molécules et la difficulté du problème semblaient renvoyer sa solution à un avenir lointain. Comme bien des scientifiques de ce domaine, Janet Thornton (ex-directrice de l'*European Bioinformatic Institute*, Hinxton, Grande-Bretagne) pensait que ce problème ne serait pas résolu de son vivant[5]... Nous y sommes. La structure tridimensionnelle est un élément essentiel de la fonction d'une protéine, et nous ignorons encore la fonction de milliers de protéines humaines – sans parler des centaines de milliers trouvées (et séquencées) dans d'autres organismes. Lors de l'émergence d'un nouveau pathogène, ce type d'information peut considérablement accélérer la mise au point d'inhibiteurs spécifiques de ces protéines impliquées dans son ciblage cellulaire ou sa multiplication. Les changements de conformation induits par certaines mutations, insertions ou délétions – notamment dans le domaine du cancer – pourront aussi être prédits et servir de base à des applica-

[5] *Would not get solved in my lifetime* [11].

tions thérapeutiques. Bref, cette avancée va profondément modifier le paysage de la recherche biologique et faciliter le passage de la structure à la fonction, qui est essentiel pour notre compréhension du monde vivant. ◊

SUMMARY

A giant step towards elucidating the function of proteins

Prediction of the 3D structure of a protein from its amino-acid sequence should in principle be possible, since most proteins fold spontaneously on synthesis. However attempts over the last 50 years have not been successful because of the complexity of the problem, although there has been some progress. Recently, an algorithm developed by the DeepMind company has achieved excellent prediction accuracy, thanks to skilled use of artificial intelligence, deep learning and extensive protein structure databases. This success has profound implications and will tremendously accelerate the elucidation of protein functions, with many applications in basic and applied research. ◊

LIENS D'INTÉRÊT

L'auteur déclare n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

RÉFÉRENCES

1. Sanger, F. Chemistry of Insulin: determination of the structure of insulin opens the way to greater understanding of life processes. *Science* 1959 ; 129 : 1340-44.
2. Kendrew JC, Bodo G, Dintzis HM, *et al*. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 1958 ; 181 : 662-6.
3. Loquet A, Gardiennet C, Bockmann A. Protein 3D structure determination by high-resolution solid-state NMR. *CR Chimie* 2010 ; 13 : 423-30.
4. Fujii T, Iwane AH, Yanagida T, Namba K. Direct visualization of secondary structures of F-actin by electron cryomicroscopy. *Nature* 2010 ; 467 : 724-8.
5. Service RF. "The game has changed". AI triumphs at protein folding. *Science* 2020 ; 37 : 1144-5.
6. Barouki R. Les chaperons de l'évolution. *Med Sci (Paris)* 2002 ; 18 : 1200-1.
7. Haber E, Anfinsen CB. Regeneration of enzyme activity by air oxidation of reduced subtilisin-modified ribonuclease. *J Biol Chem*. 1961 ; 236 : 422-4.
8. Sternberg MJ, Thornton JM. Prediction of protein structure from amino acid sequence. *Nature* 1978 ; 271 :15-20.
9. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-round XIII. *Proteins* 2019 ; 87 : 1011-20.
10. Silver D, Schrittwieser J, Simonyan, K, *et al.* Mastering the game of Go without human knowledge. *Nature* 2017 ; 550 : 354-9.
11. Callaway E. It will change everything: deepmind's AI makes gigantic leap in solving protein structures. *Nature.* 2020 ; 588 : 203-4.
12. Jumper J, Evans R, Pritzel A, *et al.* High accuracy protein structure prediction using deep learning. In *Fourteenth critical assessment of techniques for protein structure prediction* (abstract book). CASP14, 30 November-4 December 2020, p. 22-5. https://predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf

TIRÉS À PART

B. Jordan