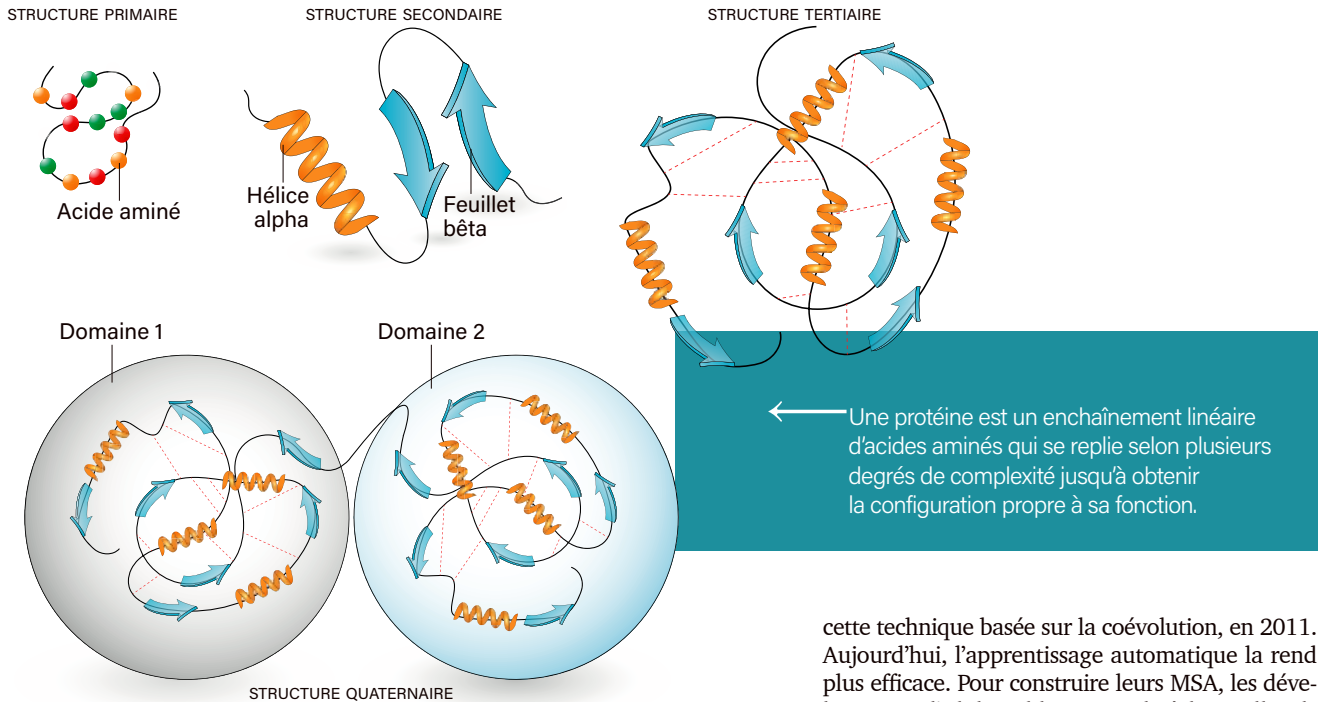




Mais ces algorithmes fonctionnaient le plus souvent au sein de logiciels plus larges, ce qui créait des risques d'erreurs. « Ils étaient en outre fragmentés, et les différents constituants communiquaient souvent mal », explique Mohammed AlQuraishi, de l'université Columbia, à New York. Ces difficultés sont levées avec des algorithmes dédiés qui gèrent l'ensemble du processus, de la séquence à la structure. C'est pourquoi John Jumper, le chercheur principal de DeepMind, à Londres, a abandonné AlphaFold après le CASP13 pour se consacrer à AlphaFold2.

L'ÈRE DES TRANSFORMEURS

Plusieurs aspects de ce programme reposent sur des techniques établies. Par exemple, l'algorithme commence par des alignements multiséquences (MSA), où la séquence d'une nouvelle protéine de structure inconnue est comparée à celles de protéines apparentées provenant d'autres espèces. En identifiant les acides aminés qui évoluent en parallèle, les algorithmes repèrent les plus susceptibles de s'associer dans la protéine repliée, en supposant qu'une modification de la séquence en un endroit a nécessité ailleurs une mutation compensatoire pour préserver la structure globale de la protéine.

Chris Sander, de l'institut de cancérologie Dana-Farber, à Boston, et Debora Marks, de l'université Harvard, à Cambridge, avaient mis au point

cette technique basée sur la coévolution, en 2011. Aujourd'hui, l'apprentissage automatique la rend plus efficace. Pour construire leurs MSA, les développeurs d'AlphaFold2 ont utilisé les milliards de séquences protéiques compilées par Martin Steinegger, de l'université de Séoul, en Corée du Sud, et Johannes Söding, de l'institut Max-Planck de chimie biophysique, à Göttingen, en Allemagne.

L'équipe de DeepMind a également imaginé des solutions innovantes. L'une d'elles consiste en des transformeurs, des outils de reconnaissance de motifs couramment utilisés dans l'analyse d'images et le traitement du langage naturel. Les transformeurs sont conçus pour repérer des modèles locaux (des chaînes de mots fréquentes ou des éléments visuels souvent associés, par exemple) qui guident l'interprétation des données. DeepMind les a adaptés aux interactions à longue portée entre acides aminés éloignés de la séquence et susceptibles d'être importantes dans la forme finale.

AlphaFold2 traite le repliement des protéines sous plusieurs angles simultanément, et crée en parallèle plusieurs représentations de la structure possible. Celles-ci sont ensuite comparées de façon à affiner le processus de modélisation lors des itérations suivantes. Pour ce faire, John Jumper et ses collègues ont conçu une architecture de réseau neuronal qui favorise un échange d'informations fluide et efficace entre les composants du logiciel.

En raison du décalage entre les débuts publics d'AlphaFold2 et la parution des articles le décrivant, et de l'incertitude des universitaires quant à la diffusion de tous les détails, David Baker et Minkyung Baek, lui aussi de l'université de Washington, ont travaillé à partir d'informations éparses sur l'architecture du logiciel pour

développer leur propre version, RoseTTAFold. Celle-ci utilise de nombreuses stratégies identiques à celles d'AlphaFold2, mais avec quelques particularités.

«Au moment où nous l'avons mis à disposition, c'était de loin le meilleur outil de prédiction de structure, mais il n'est pas aussi bon qu'AlphaFold2», déclare David Baker. Il souligne que DeepMind est une entité privée disposant, contrairement à la plupart des laboratoires universitaires, d'énormes ressources. Le succès d'AlphaFold2 «tient beaucoup à l'argent mis par Google, confirme Amelie Stein, de l'université de Copenhague, mais aussi aux compétences des nombreux experts réunis».

Depuis la sortie d'AlphaFold2 en juillet 2021, les laboratoires se sont rués dessus. Le logiciel est simple d'utilisation, bien qu'il nécessite plusieurs téraoctets d'espace disque pour les bases de données et plusieurs unités de traitement graphique (GPU) pour effectuer les analyses. «Les calculs sur une seule structure sont plutôt faciles», explique Arne Elofsson, de l'université de Stockholm. Mais les analyses de l'ensemble des protéines d'un organisme, ou protéome, sont encore hors de portée de la plupart des laboratoires universitaires.

Pour pallier cette difficulté, Martin Steinegger et ses collègues ont mis au point ColabFold, un système dématérialisé qui fait fonctionner

AlphaFold2 n'est pas encore adapté pour comprendre l'effet des mutations sur la structure des protéines

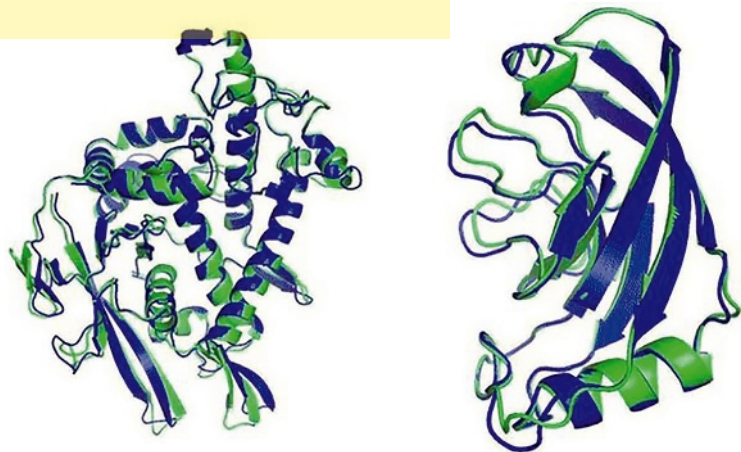
AlphaFold2 et RoseTTAFold en utilisant des bases de données distantes et la puissance de calcul fournie par Google. L'interface web est relativement simple et permet d'optimiser l'expérience.

Même l'équipe de DeepMind a été surprise par les performances d'AlphaFold2 lors du CASP14. Depuis, de nombreuses démonstrations des capacités du logiciel ont été apportées, mais également de ses limites. Dans une étude publiée parallèlement à l'article décrivant l'algorithme, l'équipe de Google a appliqué AlphaFold2 à un ensemble de données comprenant 98,5% du protéome humain. L'algorithme utilise un test de différence de distance locale prédite (pLDDT, pour *predicted local distance difference test*) pour quantifier sa confiance dans la position et l'orientation qu'il propose pour un acide aminé particulier dans la protéine réelle. Finalement, 36% de tous les résidus aminés du protéome ont pu être analysés avec un degré de confiance très élevé.

En août 2021, le groupe d'Alfonso Valencia, du Centre de supercalcul de Barcelone, en Espagne, a conclu de son côté qu'AlphaFold2 avait fait passer de 31 à 50% la proportion d'acides aminés des protéines humaines cartographiables avec précision.

SANS FAUX PLI

Yang Zhang s'attend à ce que le logiciel fasse un travail rapide sur les «fruits les plus faciles à cueillir» du protéome. Mais l'analyse de nombreuses autres protéines relève du défi, notamment celles qui sont composées de plusieurs unités fonctionnelles indépendantes (des domaines) reliées de façon flexible. Ici, les domaines individuels sont



Dans la détermination des structures de protéines, l'intelligence artificielle (*en bleu*) fait désormais presque jeu égal avec les études expérimentales (*en vert*).