

On a longtemps pensé que les langues
étaient des systèmes symboliques,
et que traduire revenait à effectuer des calculs
sur des symboles

Certaines langues sont-elles mieux traduites que d'autres ?

La plupart des traducteurs automatiques en ligne grand public proposent une centaine de langues, qui sont toutes des langues écrites (sur les quelque 4000 qui disposent d'une forme d'écriture plus ou moins développée). Celles pour lesquelles la traduction fonctionne le mieux, une vingtaine, sont celles pour lesquelles on a le plus de données – la taille du corpus est le premier critère. Et la traduction est meilleure si l'anglais est la langue source ou la langue cible. Cette langue présente deux avantages : la masse des données disponibles, mais aussi sa faible complexité morphologique, c'est-à-dire que la forme des mots change très peu : au présent simple le verbe reçoit un *s* à la troisième personne du singulier, le passé se forme souvent en *ed*, etc., mais c'est peu, comparé au français. En allemand, la forme du mot varie selon sa fonction dans la phrase, ce qui veut dire qu'il faut déterminer sa fonction syntaxique avant de traduire. D'autres langues agrègent encore plus d'informations sur une seule forme graphique, par exemple, en finnois, le nombre, le possessif, etc. Produire la bonne forme est dès lors très difficile, alors qu'en anglais beaucoup d'informations passent par des « petits » mots isolés (préposition, déterminant, possessif) : vous pouvez tout décomposer, et la traduction mot à mot ne marche pas si mal.

Quel est l'avenir de la traduction automatique ?

Pour la recherche, l'un des enjeux importants est de mieux traduire les langues pour

lesquelles on a le moins de données, voire celles pour lesquelles on n'a même pas deux corpus qui peuvent être « alignés » au niveau des phrases, c'est-à-dire pour chaque phrase dans la langue source, une phrase en regard dans la langue cible, afin d'acquérir les connaissances permettant de passer d'une langue à l'autre. Pour les opérateurs de télécommunication, l'enjeu est plus économique : il s'agit de traduire de l'oral vers l'oral afin de pouvoir converser de manière transparente, même sans connaître la langue de l'autre. En pratique, la parole est d'abord transcrite en texte, lequel est traduit, puis retranscrit à l'oral. Les performances s'améliorent et on peut imaginer que, dans quelques années, on utilisera ce type de technologie dans nos systèmes de communication à distance, en visio.

Est-ce la fin du métier de traducteur ?

Il est sûr que le métier change. Par exemple, les sites d'information sur le web proposent désormais des vidéos dont les sous-titres sont souvent traduits automatiquement. Le résultat de la traduction par le programme est juste retouché au minimum lors d'une phase de post-édition : le traducteur relit le texte produit, regarde l'original et ne corrige que les contresens. Tant pis si, esthétiquement, le français n'est pas idéal. Même si la traduction automatique va ouvrir de nouveaux marchés (pour les revues de presse, le journalisme ou l'édition par exemple), il est clair qu'elle propose un service plus rapide et moins coûteux, qui conduit dans certains cas à des remplacements purs et simples.

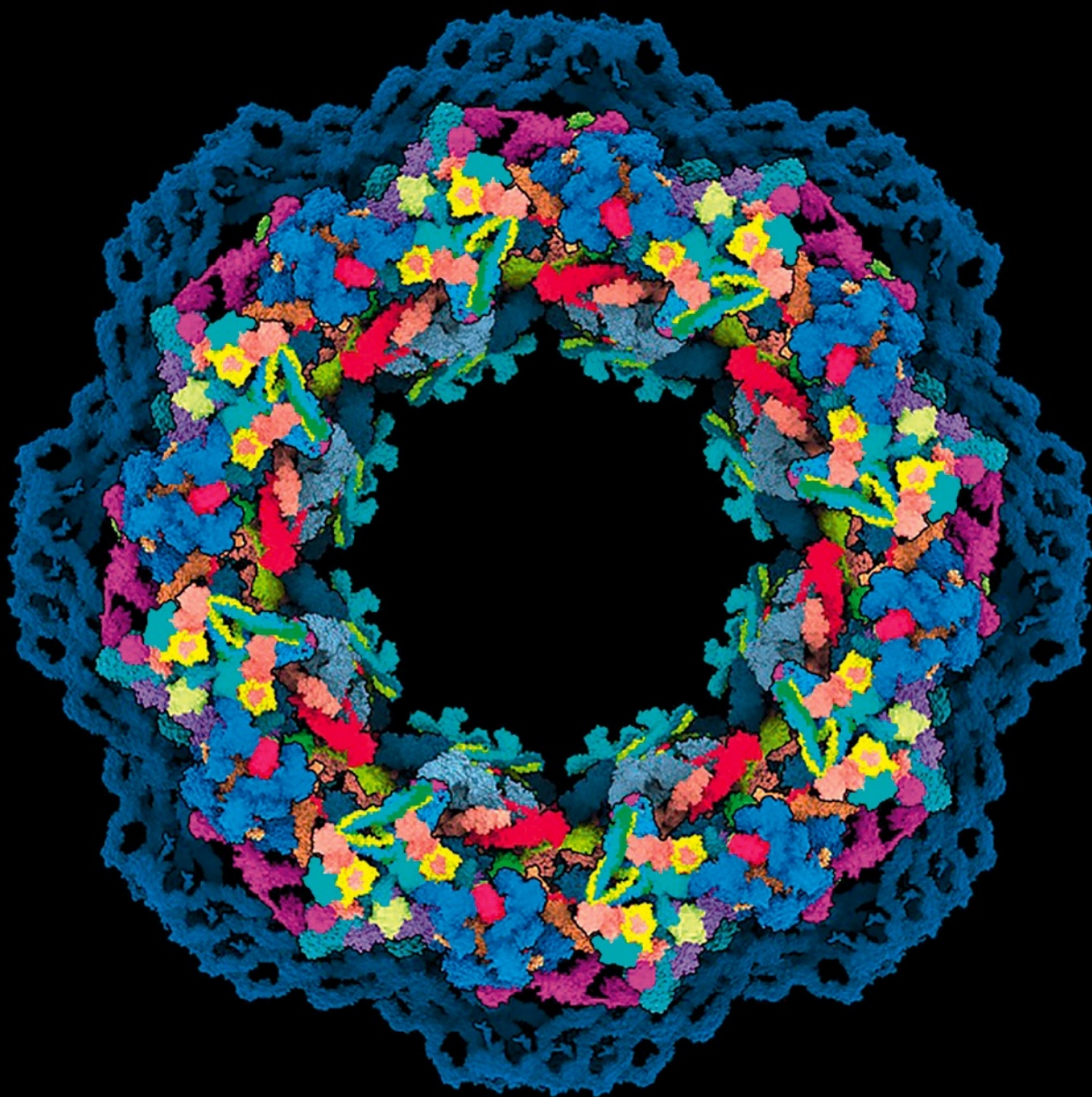
Propos recueillis par Olivier Voizeux

Les algorithmes d'apprentissage profond prédisent désormais la forme 3D d'une protéine à partir de sa séquence linéaire. Une révolution pour la biologie!

28

Les protéines, nouveau terrain de jeu de l'IA

Michael Eisenstein



La structure de ce complexe protéique
(un pore nucléaire humain)
a été déterminée à l'aide d'AlphaFold2
et de données structurales.

«L'un des plus grands mystères de la biologie est en grande partie résolu par l'IA», pour la BBC. La «réalisation la plus importante de l'IA», selon Forbes. Rarement un logiciel scientifique aura fait l'objet de titres aussi sensationnels. De quoi parle-t-on ? D'un nouveau jeu vidéo ? Non, du programme AlphaFold2 mis au point par DeepMind, la branche de Google consacrée à l'intelligence artificielle, et dédié à l'un des problèmes les plus épineux de la biologie : prédire la structure tridimensionnelle, et donc fonctionnelle, d'une protéine à partir de sa séquence linéaire en acides aminés. L'association américaine pour l'avancement des sciences en fait une des «percées de l'année 2021».

Auparavant, les chercheurs déterminaient la structure 3D des protéines grâce à des techniques expérimentales lourdes, comme la cristallographie à rayons X et la cryomicroscopie électronique (cryo-EM). Les informaticiens ont fait des progrès constants en développant des logiciels qui complètent ces méthodes, mais malgré ces avancées, on ne disposait toujours pas d'informations structurales pour environ 4800 protéines humaines : AlphaFold2 a réduit ce nombre à 29...

Le programme a été dévoilé en novembre 2020 lors de la 14^e édition d'une compétition biennale, le CASP (*Critical assessment of structure prediction*). L'objectif : mettre au défi biologistes et algorithmes avec des séquences de protéines dont les structures ont été déterminées expérimentalement sans être rendues publiques. Le logiciel de DeepMind a survolé la compétition.

«Au CASP14, il a livré la structure d'environ deux tiers des protéines proposées, dont certaines difficiles, avec une précision de type expérimentale», explique Yang Zhang, de l'université du

— En bref —

> La fonction d'une protéine, un enchaînement d'acides aminés, dépend de sa structure tridimensionnelle.

> Pour la déterminer, on a longtemps eu recours à des études expérimentales complexes, comme la cristallographie à rayons X et la cryomicroscopie électronique.

> Depuis peu, l'intelligence artificielle fondée sur l'apprentissage profond a révolutionné la discipline.

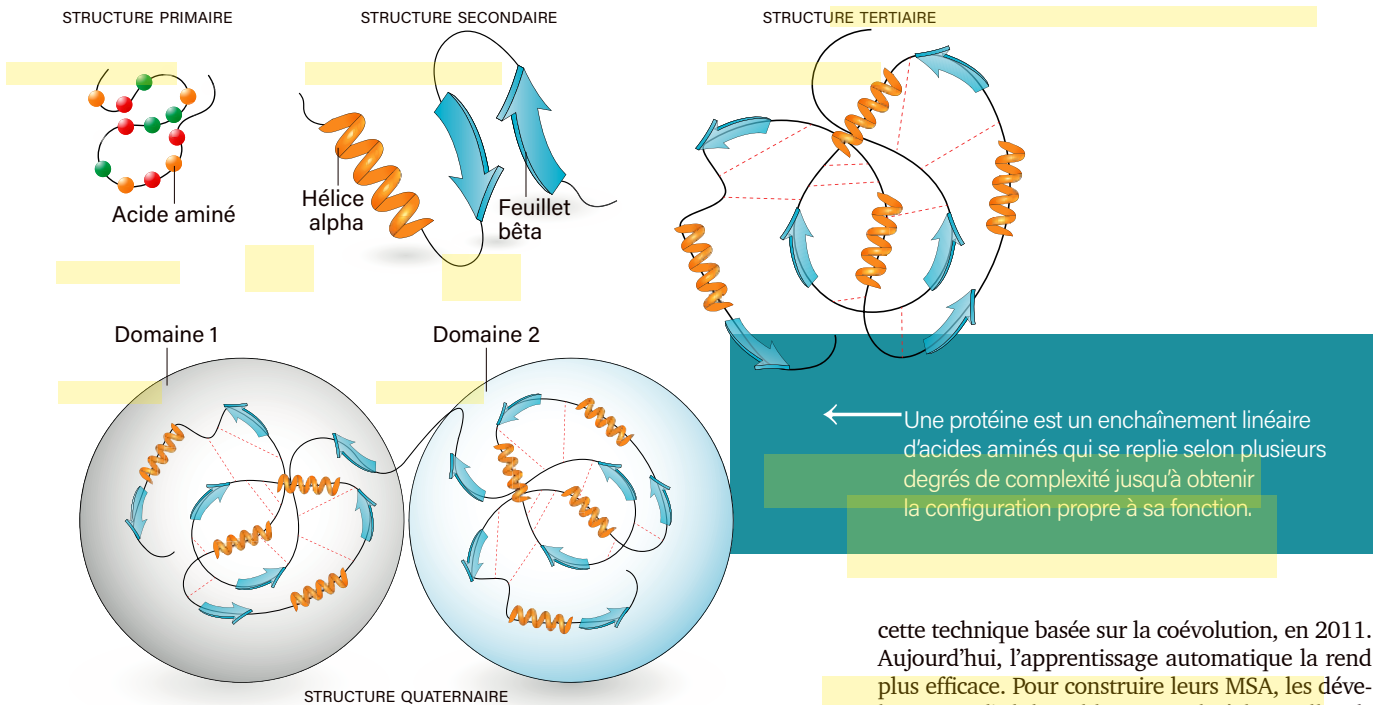
> Le logiciel AlphaFold2, conçu par Google, est à ce jour le programme le plus efficace, mais quelques obstacles, notamment la dynamique des protéines, restent à franchir.

Michigan, dont l'algorithme figurait parmi les finalistes. Depuis, plusieurs publications ont confirmé l'efficacité d'AlphaFold2. Le biologiste y voit une preuve éclatante de la puissance de l'apprentissage profond, mais une solution encore partielle au problème du repliement des protéines, car certains aspects restent inaccessibles. Comment fonctionnent ces algorithmes ?

«ENSEIGNEMENT SUPÉRIEUR»

Le processus intègre des phases d'apprentissage automatique durant lesquelles des réseaux neuronaux sont formés pour reconnaître et interpréter des modèles dans les données. «Ces modèles n'essaient pas de prédire la structure en une seule fois, explique David Baker, de l'université de Washington, à Seattle. Ils s'apparentent davantage à une simulation physique où les modèles apprennent à faire les bons mouvements pour améliorer la structure.» Entraînés avec de grandes quantités de données expérimentales annotées, ces algorithmes identifient progressivement des liens entre la séquence et la structure.

Ces cinq dernières années, les progrès ont été importants. La première itération d'AlphaFold a remporté le CASP13, en 2018, mais ses performances étaient loin de celles de son successeur en 2020. Plusieurs laboratoires universitaires ont par la suite développé des programmes basés sur l'apprentissage profond surpassant la première génération d'AlphaFold, notamment D-I-TASSER du laboratoire de Yang Zhang, trRosetta du groupe de David Baker, et RaptorX, développé par Jinbo Xu et son équipe de l'institut de technologie Toyota, à Chicago.



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 éparpillées sur l'architecture du logiciel pour