COMMUNIQUÉ DE PRESSE

Vers une conception de nouvelles protéines par apprentissage statistique

Paris, le 12 octobre 2021

Une équipe de recherche de l'Institut de Biologie Paris-Seine (Sorbonne Université / CNRS)[1], en collaboration avec l'École normale supérieure - PSL et l'École polytechnique de Turin a proposé une nouvelle approche très efficace de modélisation informatique de protéines dite "générative". Publiée dans la revue *Nature Communications* le 4 octobre 2021, cette approche permet de concevoir des séquences protéiques artificielles statistiquement équivalentes aux séquences naturelles, une caractéristique de grand intérêt dans le domaine du « *protein design* ». Cette publication est la première à bénéficier du soutien financier de l'initiative *i-Bio* de l'Alliance Sorbonne Université pour la promotion de l'interdisciplinarité dans la recherche biologique.

Au cours de l'évolution, les protéines explorent l'espace des séquences fonctionnelles. L'interaction entre mutations aléatoires du génome et sélection naturelle des organismes a permis l'apparition de milliers de protéines ayant des séquences d'acides aminés distinctes, mais des fonctions biologiques ou des structures tridimensionnelles équivalentes. Grâce aux techniques modernes de séquençage des génomes, de plus en plus de ces séquences sont connues. La base de données Uniprot rassemble, par exemple, plus de 200 millions de séquences distinctes, mais seulement environ 0,25 % de ces séquences ont une structure ou une fonction connue expérimentalement.

Les approches informatiques basées sur la science des données, la physique statistique et ou l'intelligence artificielle gagnent rapidement en importance pour explorer cette richesse croissante de données et en extraire des informations biologiques. Récemment, un exemple impressionnant a été donné par AlphaFold soutenu par Google Deepmind, qui arrive à prédire les structures des protéines à partir des séquences avec une précision sans précédent. Dans ce contexte, les modèles dits "génératifs" suscitent également un intérêt croissant, de par leur capacité à générer de manière computationnelle des séquences artificielles d'acides aminés statistiquement équivalentes à leurs homologues naturels. Il a récemment été démontré que la modélisation générative offre un nouveau paradigme pour concevoir et optimiser de nouvelles protéines en utilisant les bases de données existantes, avec des enjeux économiques importants.

Une équipe de recherche de l'Institut de biologie Paris-Seine (IBPS – Sorbonne Université/CNRS) dirigée par Martin Weigt, enseignant-chercheur à Sorbonne Université a proposé, en collaboration avec des chercheurs du Laboratoire de physique de l'ENS (LPENS, École normale supérieure/CNRS/SU/Université de Paris) et de l'École Polytechnique de Turin, une nouvelle méthode plus performante pour l'apprentissage de modèles génératifs. Cette approche dite « autorégressive », parant de familles connues de protéines et de leurs séquences pour ajuster un modèle statistique, permet à la fois de proposer de nouvelles séquences protéiques et de donner des informations sur la structure et la fonction des protéines associées.

Grâce à son efficacité, cette méthode peut être utilisée sur des milliers de familles de protéines, y compris celles ayant de très longues séquences. Elle permet de générer et d'évaluer de nouvelles séquences, qui n'ont jamais été trouvées dans la nature auparavant. Selon l'équipe de chercheurs, ces séquences artificielles seront importantes pour l'optimisation et la conception de grandes protéines de fonctionnalité donnée (par exemple, des enzymes efficaces et thermostables), question où la recherche fondamentale rejoint des enjeux technologiques et biomédicaux.

[1] Laboratoire biologie computationnelle et quantitative (LCQB)

À propos de l'initiative i-Bio :

L'initiative i-Bio (*initiative for interdisciplinary research in biology*) de l'Alliance Sorbonne Université s'inscrit dans une dynamique de recherche interdisciplinaire en biologie pour comprendre le vivant dans son intégralité et dans son histoire, à toutes les échelles de temps et d'espace. Cette initiative est portée par l'Institut de biologie Paris-Seine (IBPS – Sorbonne Université/CNRS) et l'Institut du Fer à Moulin (Sorbonne Université/Inserm). Son programme inclut des contrats doctoraux, un soutien à des projets de recherche transdisciplinaires, le recrutement de nouvelles équipes et une animation scientifique active.

Pour en savoir plus :

- L'initiative i-Bio
- Le programme instituts et initiatives de l'Alliance Sorbonne Université

Référence :

Efficient generative modeling of protein sequencesusing simple autoregressive models
Jeanne Trinquier, Guido Uguzzoni, Andrea Pagnani, Francesco Zamponi & Martin Weigt, *Nature Communications*, October 4, 2021

DOI : https://doi.org/10.1038/s41467-021-25756-4

À propos de Sorbonne Université :
Sorbonne Université est une université pluridisciplinaire de recherche intensive de rang mondial. Structurée en trois facultés, elle couvre les champs des lettres, de la médecine et des sciences. Ancrée au cœur de Paris et présente en région, Sorbonne Université est impliquée dans la réussite de sa communauté étudiante. Elle s'engage à répondre aux grands enjeux sociétaux et à transmettre les connaissances issues de ses laboratoires et de ses équipes de recherche. Grâce à ses 52 000 étudiantes et étudiants, 6 400 personnels d'enseignement et de recherche et 3 900 personnels administratifs et techniques, Sorbonne Université se veut diverse, créatrice, innovante et ouverte sur le monde. Avec le Muséum national d'Histoire naturelle, l'Université de Technologie de Compiègne, l'INSEAD, le Pôle Supérieur Paris Boulogne-Billancourt et France Education International, elle forme l'Alliance Sorbonne Université favorisant une approche globale de l'enseignement et de la recherche, promouvant l'accès au savoir, et développant des programmes et projets de formation. Sorbonne Université est également membre de l'Alliance 4EU+, un modèle novateur d'université européenne.
https://www.sorbonne-universite.fr *@ServicePresseSU*

Contacts presse

Marion Valzy 01 44 27 37 13 - 06 14 02 20 51
marion.valzy@sorbonne-universite.fr

Claire de Thoisy-Méchin 01 44 27 23 34 - 06 74 03 40 19
claire.de_thoisy-mechin@sorbonne-universite.fr

Contact chercheur

Martin Weigt, professeur à Sorbonne Université, chercheur à l'Institut de biologie Paris-Seine (IBPS)
martin.weigt@sorbonne-universite.fr