Innovations structurales et fonctionnelles chez les dinoflagellés : l'alliance des masses de données omiques et de l'apprentissage profond nous permettra-t-elle de les identifier ?

Proposé par Mathilde Carpentier (porteur) et Lucie Bittner

Contexte

Deux révolutions méthodologiques importantes ont eu lieu ces dernières années en biologie. D'un côté, les NGS nous permettent d'obtenir des milliards de séquences provenant d'organismes cultivables mais aussi non cultivables pour lesquels il est possible maintenant de reconstruire des génomes à partir des données de l'environnement (MAGs). Il est néanmoins difficile d'exploiter cette nouvelle masse d'information car beaucoup des séquences produites sont trop différentes de celles que nous connaissons pour que nous puissions transférer nos annotations avec les méthodes classiques : 40% à 90% de séquences des métagénomés marins procaryotes de TARA n'ont pas d'annotation et 50% des séquences des métagénomés eucaryotes n'ont pas d'annotation taxonomique et/ou fonctionnelle. De l'autre côté, l'IA et en particulier le *deep learning* nous permettent d'analyser de très grandes quantités de données et d'en inférer des informations difficilement accessibles jusqu'à présent. Il y a néanmoins encore beaucoup à faire pour conjuguer ces deux avancées.

Les objectifs de cette thèse sont doubles : méthodologiques et analytiques. En effet, les récentes avancées des méthodes d'apprentissage profond en particulier pour la prédiction des structures protéiques ouvrent d'importantes perspectives pour l'exploration des quantités massives de données omiques générées grâce aux nouvelles techniques de séquençages. Néanmoins, il n'existe pas encore à l'heure actuelle de méthode utilisant la structure pour annoter fonctionnellement et taxonomiquement et c'est ce que nous proposons de développer ici. Nous voulons utiliser ensuite ces méthodes pour explorer l'univers des dinoflagellés qui sont des eucaryotes unicellulaires aquatiques très diversifiés, en particulier par leurs comportements alimentaires. Au-delà des enjeux de santé publique et de la compréhension voire de la préservation du fonctionnement des écosystèmes et des cycles biogéochimiques associés, ces jeux de données offrent un éclairage inédit pour étudier les mécanismes adaptatifs et évolutifs sous-jacents.

Verrous : La structure va-t-elle nous permettre d'annoter fonctionnellement et taxonomiquement les séquences ?

Actuellement, l'annotation taxonomique et fonctionnelle des séquences est classiquement faite par transfert des annotations des séquences suffisamment proches, mais nous avons atteint les limites de ces méthodes. Les récents progrès en IA permettent de prédire la structure 3D des séquences avec une très grande fiabilité, ce qui va nous permettre d'étendre considérablement nos connaissances. En quoi connaître la structure nous permettra de connaître la fonction et la taxonomie ? La structure est connue pour être bien plus conservée que la séquence. La structure étant prédite, il sera donc possible de la comparer aux structures connues et de retrouver des homologies devenues invisibles en séquence. Se posent néanmoins deux questions. La première question est de savoir si notre connaissance de l'univers des structures est suffisante pour retrouver des similitudes ou si nous les structures prédites seront toutes de nouvelles structures pour lesquelles il sera difficile d'inférer fonctions et des taxonomies. Le nombre de repliements structuraux est étonnamment petit, de l'ordre du millier, nous n'en trouvons actuellement pratiquement plus de nouveaux et la distribution des repliements dans les génomes suit une loi de puissance : quelques repliements sont extrêmement fréquents tandis que beaucoup le sont peu. A partir de ces éléments, il semble raisonnable de penser qu'une grande majorité de ces nouvelles séquences non annotées font partie des repliements déjà connus. Néanmoins, ce raisonnement repose sur des données extraites d'un ensemble plutôt biaisé d'organismes modèles. Il sera donc intéressant de répondre à cette question avec des données environnementales et cela pourra être

Mis en forme : Gauche : 2 cm, Droite : 2 cm, Haut : 2 cm, Bas : 2 cm, Distance de l'en-tête par rapport au bord : 1,25 cm, Distance du bas de page par rapport au bord : 1,25 cm

Supprimé: des séquences

Supprimé: la

Supprimé: la structure 3D des séquences

Supprimé: très

Supprimé: des connaissances

Supprimé: Le faible nombre de de repliements structuraux différents et la distribution en loi de puissance des repliements dans les génomes nous laisse à penser que c'est bien le cas et qu'une grande majorité de ces nouvelles séquences non annotées font partie des repliements déjà connus et très fréquents.

Supprimé: r

Supprimé: , et quantifier la proportion de nouveau repliement,

publié dans article issu de ce travail. Si le nombre de repliements différents est si faible, on peut aussi raisonnablement se demander si les convergences structurales ne seraient pas fréquentes, ce qui rendrait impossible alors l'inférence d'homologie à partir de similitudes structurales, et l'annotation taxonomique. Nous avons récemment montré que ce n'est pas le cas et même que les repliements sont des caractères permettant de caractériser des clades même récents³. La seconde question est : connaître le repliement permet-il de connaître la fonction ? la réponse à cette question est loin d'être évidente. En effet, suivant les familles de repliements, les fonctions sont plus ou moins spécifiques : par exemple, la superfamille des α/β -Hydrolases regroupe des protéines ayant des rôles très diverses (lipases, **estérases** et hydrolases variées). Nous proposons ici d'apporter des solutions à ce problème en caractérisant fonctionnellement de manière plus précise les fonctions associées aux repliements, à la fois avec des motifs structuraux caractéristiques mais aussi des profils de coévolution (modèles de Pott, travail initié lors du Projet Emergence Pepper porté par M Carpentier en 2021-2022).

Supprimé: fera l'objet d'un

Supprimé: aussi

Supprimé: esterases

Supprimé: aux

Supprimé: avec

Résultats attendus

Nous avons choisi de travailler sur les dinoflagellés pour plusieurs raisons. C'est un groupe sur lequel nous travaillons déjà dans l'équipe et donc pour lequel nous avons une expertise⁴ et pour lequel de multiples questions se posent quant à ses modes trophiques son rôle dans l'environnement. Ce sont aussi des eucaryotes, super-règne dans lequel nous avons montré **qu'un très grand nombre de repliements sont apparus spécifiquement**^{3,5}. **Les dinoflagellés se développent dans des milieux très divers et ont des modes de vie très variés. On peut donc s'attendre à beaucoup d'innovation structurale** et donc aussi très probablement fonctionnelle dans ce groupe. Nous tenterons ainsi de relier les profils fonctionnels aux conditions environnementales pour identifier les principaux acteurs impliqués dans certaines niches et biomes de l'océan. A plus long terme, ce travail posera les jalons de la méthodologie qui pourra être appliquée dans le cadre du PEPR Atlasea dans lequel nous sommes impliqués et qui séquencera et analysera le génome de 4 500 espèces des côtes françaises.

Supprimé: qu'il y a une très grande innovation structurale

Commenté [hd3]: N'est-ce pas ce que vous cherchez à montrer ? si oui il faudrait peut-être moduler.

Supprimé:

Supprimé: lement une grande innovation

Les données sur lesquelles nous allons travailler sont déjà disponibles dans les bases MetDB et EukProt. Nous enrichirons notre base avec d'autres données provenant du PEPR Atlasea dans lequel l'équipe d'accueil est impliquée en particulier pour la réflexion autour de l'analyse des données. Nous attendons donc des résultats méthodologiques et analytiques de ce travail. Pour l'annotation structurale, fonctionnelle et taxonomique des données omiques il faut donc à la fois identifier les méthodes de deep learning les plus adaptées mais aussi développer des méthodes (comparaison de structures, caractérisation des fonctions) et c'est ce que nous ferons ici. Côté analytique, l'application de ces méthodes au jeu de données récoltées sur les dinoflagellés nous permettra de les tester mais aussi de répondre à certaines questions tant à propos de la proportion de repliements inconnus que des profils de répartition des repliements pour les différentes communautés étudiées. Nous tenterons de relier ces profils aux conditions environnementales pour identifier les principaux acteurs impliqués dans certaines niches et biomes de l'océan.

Nous proposons ce projet à l'Institut des Sciences des données et du Calcul car il repose sur l'utilisation d'algorithmes innovants et leur application sur de grandes masses de données, et le développement des méthodes permettant cette application

Bibliographie

- 1 Sillitoe, I. et al. *Nucleic Acids Res* 49, D266–D273 (2021)
- 2 Koonin, E. V. et al. *Nature* 420, 218–223 (2002)
- 3 Romei, M. et al. *Evolution* 76, 1706–1719 (2022)
- 4 Meng, A. et al. *Molecular Ecology* 27, 2365–2380 (2018)
- 5 Romei, M. et al. *Soumis* (2022)