

RÉSUMÉ

Les initiatives de génomique et de transcriptomique marines produisent aujourd'hui des volumes massifs de séquences protéiques issues d'organismes non modèles, révélant une diversité protéique encore largement inexplorée. Toutefois, une part importante de ces protéines reste difficile à interpréter fonctionnellement, les approches basées uniquement sur la similarité de séquence atteignant leurs limites lorsque la divergence évolutive est forte.

Le projet propose de dépasser ces limites en intégrant l'information structurale des protéines aux analyses de réseaux de similarité de séquences en utilisant les approches de graphes connaissances. L'hypothèse centrale est que la structure tridimensionnelle, plus conservée que la séquence, constitue un niveau pertinent pour relier évolution moléculaire, fonction et contexte écologique. L'objectif est ainsi de développer un cadre d'analyse permettant d'identifier des signatures structurales et évolutives associées à des fonctions biologiques et à des niches environnementales spécifiques.

Le travail reposera sur l'analyse de jeux de données issus de la génomique et de la transcriptomique environnementales marines, la construction de réseaux de similarité de séquences et leur enrichissement par des informations structurales issues de bases de données ou de méthodes de prédiction. Ces approches permettront de caractériser l'organisation de l'espace protéique, d'identifier des familles enzymatiques émergentes et de mieux comprendre les contraintes évolutives qui façonnent la diversification fonctionnelle.

Ce projet se situe à l'interface entre biologie évolutive, écologie microbienne et informatique. Il vise à produire des outils méthodologiques transférables pour l'analyse de données omiques à grande échelle, tout en apportant de nouvelles connaissances sur l'évolution des protéines marines et leur rôle dans le fonctionnement des écosystèmes.

ABSTRACT

Marine genomics and transcriptomics initiatives are now generating massive volumes of protein sequences from non-model organisms, revealing a vast and still largely unexplored protein diversity. However, a significant fraction of these proteins remains difficult to interpret functionally, as approaches based solely on sequence similarity reach their limits when evolutionary divergence is high.

This project aims to overcome these limitations by integrating protein structural information into sequence similarity network analyses using knowledge-graph approaches. The central hypothesis is that three-dimensional structure, which is more conserved than sequence, provides a relevant level of organization to link molecular evolution, biological function, and ecological context. The objective is therefore to develop an analytical framework capable of identifying structural and evolutionary signatures associated with biological functions and specific environmental niches.

The project will rely on the analysis of datasets from marine environmental genomics and transcriptomics, the construction of sequence similarity networks, and their enrichment with structural information derived from databases or prediction methods. These approaches will make it possible to characterize the organization of protein space, identify emerging enzymatic families, and better understand the evolutionary constraints shaping functional diversification.

This project lies at the interface between evolutionary biology, microbial ecology, and computer science. It aims to produce transferable methodological tools for large-scale omics data analysis, while providing new insights into the evolution of marine proteins and their role in ecosystem functioning.