Projet Bioinformatique EIDD

Introduction: Protein Units (Unités Protéiques)

Le concept d'Unité Protéique permet une nouvelle vision de l'architecture des protéines. Les protéines sont habituellement représentées selon deux niveaux de complexité : un niveau local (les structures secondaires qui sont une représentation très locale et élémentaire) et un niveau plus global (les domaines, souvent très complexes) (fig. 1). Les différences de taille et de complexité entre ces descriptions ne permettent cependant pas une vision continue de l'organisation structurale.

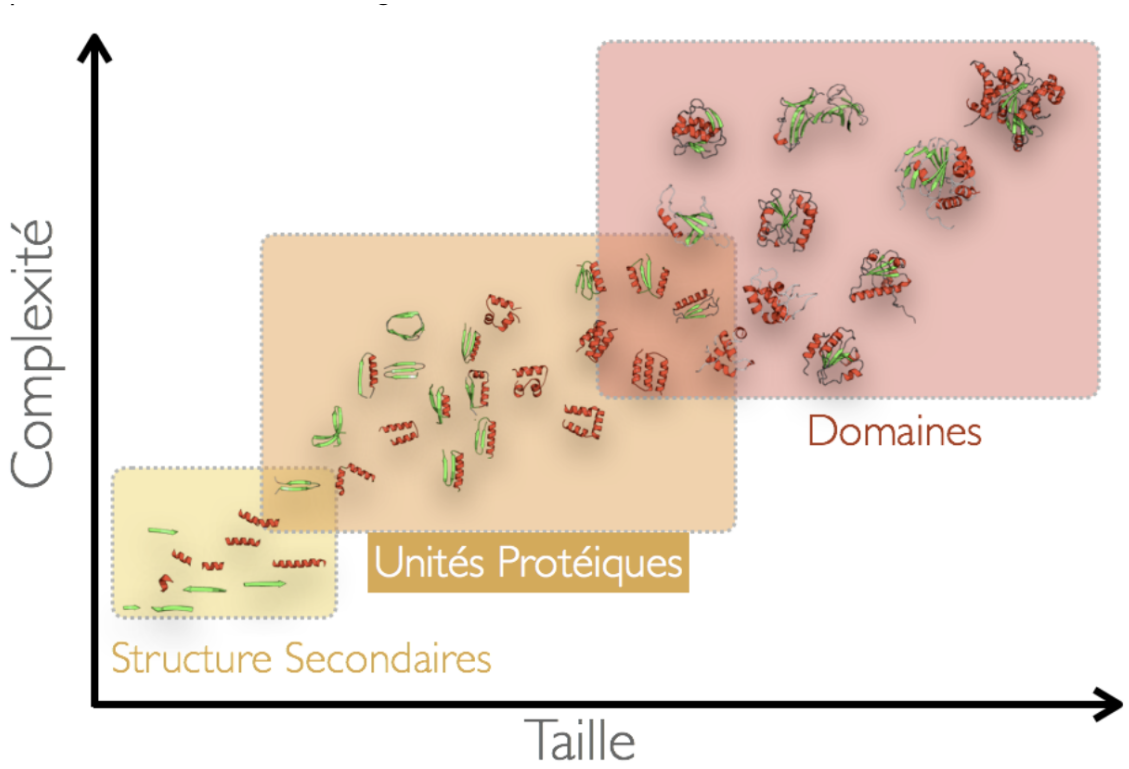

Figure 1 : Les UPs sont un échelon complémentaire entre structures secondaires et domaines protéiques.

Les UPs sont donc un échelon structural intermédiaire compact entre structures secondaires et domaines, pertinent pour décrire, analyser et comprendre l'anatomie et la structure des protéines.
Nous avons également réalisé une analyse systématique sur un ensemble évolutionnairement non redondant de structure des protéines. Nous avons segmenté en UPs ces structures puis les avons regroupées selon leurs similarités structurales. correspondant à 3000 superfamilles d'UPs ont été obtenues.

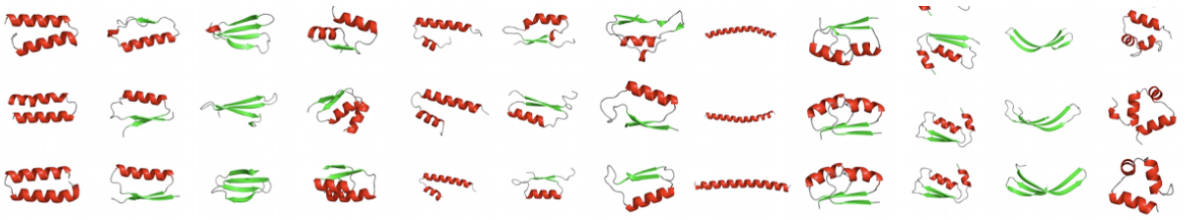

Figure 2: Quelques exemples de superfamilles d'unités protéiques

Ces Unités Protéiques, regroupées en superfamilles, ont été appelées Unités Protéiques Cores (UPC). Elles sont topologiquement simples, très compactes, récurrentes et énergétiquement favorables. Cette classification en superfamille permet de disposer d'une vision générale des UPs et de leur répartition au sein des protéines. Les travaux effectués jusqu'à aujourd'hui ont permis une caractérisation générale des unités protéiques et ont démontré la pertinence de l'approche, et ouvrent de nouvelles perspectives d'études théoriques et d'applications pratiques.

Dans un premier temps l'idée du projet est de calculer la variance structurale (par le RMSD, TMscore et LDDT) d'Unités Protéiques contenu dans des protéine homologues dont les structures ont été résolus expérimentalement..Dans un deuxième temps le projet consiste à évaluer la "prédictibilité" *a priori* des régions contenant des Unités Protéiques.

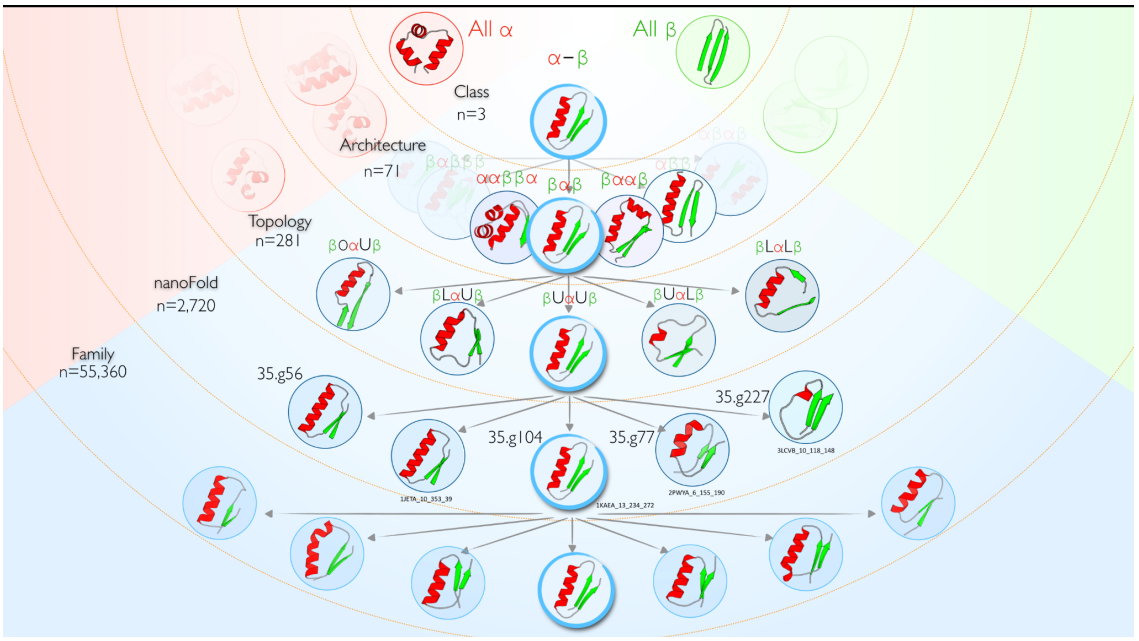

Figure 3: Classification hiérarchique des Unités Protéiques représentant les différents niveaux d'organisation

Introduction: Knottins

https://www.dsimb.inserm.fr/KNOTTIN/ : KNOTTIN Database
https://academic.oup.com/nar/article/46/D1/D454/4607803?login=false

Protein Units and/or knottin recognition from sequence using transformers & embedding

Le projet vise à détecter des Unités protéiques et/ou Knottins dans des séquences de protéines en utilisant du transfert learning à partir de Transformers déjà entraînés. La prédiction sera à plusieurs niveaux : détection de la présence ou de l'absence d'Unité Protéique, puis détection du type d'Unité Protéique selon les différents niveaux hiérarchiques (classe, architecture, Topologie, nano-repliement, famille : voir figure 3).

Exploring the protein structure universe using auto-encoders on knottin and/or Protein Units

Objectif : L'idée de ce projet est d'étudier l'espace de projection des structures et fonctions, et de déterminer les relations et bijection d'un espace à un autre.

Les méthodes se baseront sur les *embedding* de structures obtenues par un *Variational auto-encoder*.

PSI-Blast using *Protein Language Model embedding*

Objectif : Réalisez le programme BLAST et PSI-BLAST en reprenant la méthode décrite dans l'article ou sur la page wikipedia : https://fr.wikipedia.org/wiki/Basic_Local_Alignment_Search_Tool .

La matrice de score sera remplacée par un calcul de cosine similarity utilisant un embedding en remplacement de l'acide aminé. Afin d'obtenir l'embedding à partir de la séquence vous utiliserez EMS2 (exemple d'utilisation : https://shorturl.at/hmDM3)

Référence : Stephen Altschul, Warren Gish, Webb Miller; Eugene Myers; David J. Lipman (1990). "Basic local alignment search tool". Journal of Molecular Biology. 215 (3): 403–410. doi:10.1016/S0022-2836(05)80360-2.

Alignement multiple using *Protein Language Model embedding*

Objectif : Réalisez un programme reprenant la méthode décrite dans l'article ci-après. L'algorithme de construction de l'arbre pourra être remplacé par la construction d'un arbre par embranchement séquentiel. L'algorithme heuristique d'alignement séquentiel pourra être remplacé par un alignement basé sur la programmation dynamique.

La matrice de score sera remplacée par un calcul de cosine similarity utilisant un embedding en remplacement de l'acide aminé. Afin d'obtenir l'*embedding* à partir de la séquence vous utiliserez EMS2 (exemple d'utilisation : https://shorturl.at/hmDM3)

Référence : Desmond G. Higgins , Paul M. Sharp, Fast and sensitive multiple sequence alignments on a microcomputer, *Bioinformatics*, Volume 5, Issue 2, April 1989, Pages 151–153, https://doi.org/10.1093/bioinformatics/5.2.151