

Modélisation de la structure des protéines pour la reconnaissance de repliement

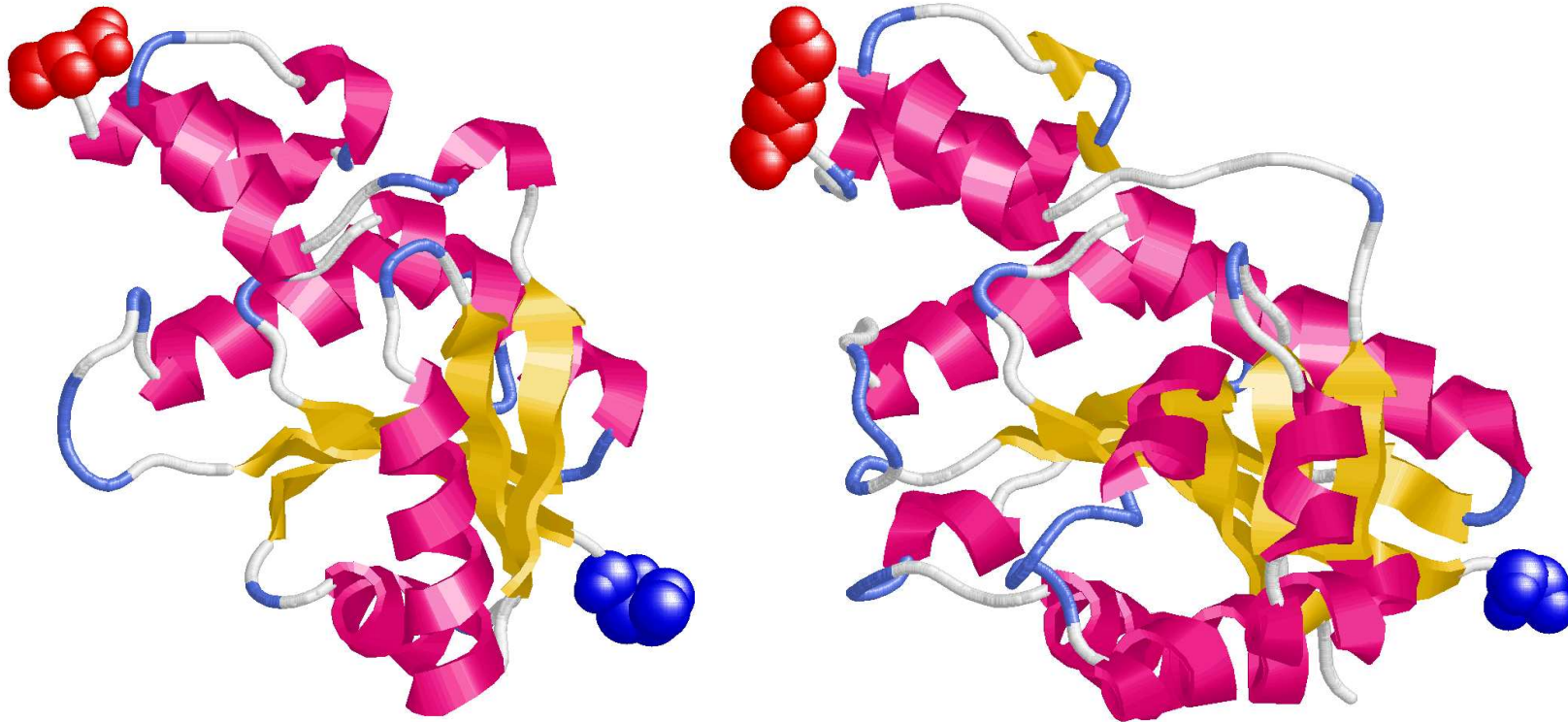
Antoine MARIN Jean-François GIBRAT
Jean-François TALY Juliette MARTIN



Pourquoi définir des « cœurs » ? Qu'est ce que c'est ?

1. définir des régions structurellement conservées (comparer ce qui est comparable) → approche algorithmique également ;
2. définir des états structurels pertinents → fonctions de score.

Qu'est-ce qu'une même structure ?



Adenylyltransferase *Mycobacterium tuberculosis* / *Homo sapiens*.

L'alignement structurel (17% id.)

```
1tfuA  ---..EEEEEE.....HHHHHHHHHHHHH-----..EEEEEE..--.....H
1nuuA  ..EEEEEEEEE.....HHHHHHHHHHHHHHHHH..EEEEEEEEE.....-..H
1tfuA  HHHHHHHHHH.....EEEEEE.-----..HHHHHHH...-----..EEEEEE.....
1nuuA  HHHHHHHHHHHHHH...EEE..HHHH.....HHHHHHHHHHHHH...EEEEEEH--HH
1tfuA  HHHHHHHH--HHHHHHH--..EEEEEE..-----HHH
1nuuA  HHH.....HHHHHHHHHH..EEEE.....HHHHHHH.HHHHH.HHHEEEE.....
1tfuA  ...HHHHHHHHH.....HHH..HHHHHHHHHHH.----.
1nuuA  .-..HHHHHHHHH.....HHHHHHHHHH.....-
```

PSI-BLAST = 62% correct.

Vous avez dits états structurels ?

- structure secondaire ;
- surface accessible au solvant ;
- enfouissement des atomes polaires ;
- zones d'angles ϕ et ψ ;
- etc.

État actuel de FROST

Les régions conservées sont définies par les éléments de structure secondaire.

Les états sont codés en « dur » : Hb, He, Cb, Ce, Eb, Ee.

Ce qui est en train de changer

Les régions conservées sont définies par des nombres réels entre 0 et 1.

→ pour le 3D nécessité de définir des blocs (pour le moment).

Les états sont des mots → matrices de substitution (1D et 3D).

À chaque ensemble d'états (N), un type de filtre est défini, qui est implanté automatiquement en 1D et 3D.

Exemple d'application possible

Modélisation des protéines membranaires.

Régions structurellement conservées = hélices transmembranaires.

États « structurels » = intérieur, extérieur, en contact avec des lipides, en contact avec d'autres hélices.

→ il faut avoir suffisamment d'homologues pour les comptages.

Construction d'un « cœur »

Tout est « automatique ».

- extraction de la séquence ;
- extraction de la fonction ;
- test de propriétés diverses (faible complexité, *coiled coils* et régions transmembranaires) → filtrés, états ;
- définition de la structure secondaire ;
- définition de la surface accessible au solvant et de la « polarité » ;
- définition de la conservation structurelle ;
- recherches d'homologues (construction de profils) ;
- prédiction de la structure secondaire ;
- recherche des homologues dans la PDB (profils de structure secondaire) ;
- définition des états structurels ;
- recherche de séquences non homologues (distributions statistiques) ;
- définition de la régions précise du cœur (sous structure).

Post-traitements

- superposition des structures → évaluation des résultats ;
- recherche des homologues → ne pas biaiser le calcul des paramètres et les tests.

Ce qui n'est pas fait « maison »

Découpage en domaines.

→ utilisation de CATH et SCOP.

Actuellement : 1175 cœurs.

Nouvelle estimation :

PDB40 \simeq 5000 chaînes.

CATH \simeq 2500 familles homologues.

SCOP \simeq 1300 « superfamilles ».

→ 10h pour 500 cœurs sur le cluster.

Et après ?

Modifier FROST pour utiliser ces modifications.

Faire des (jeux de) tests.