

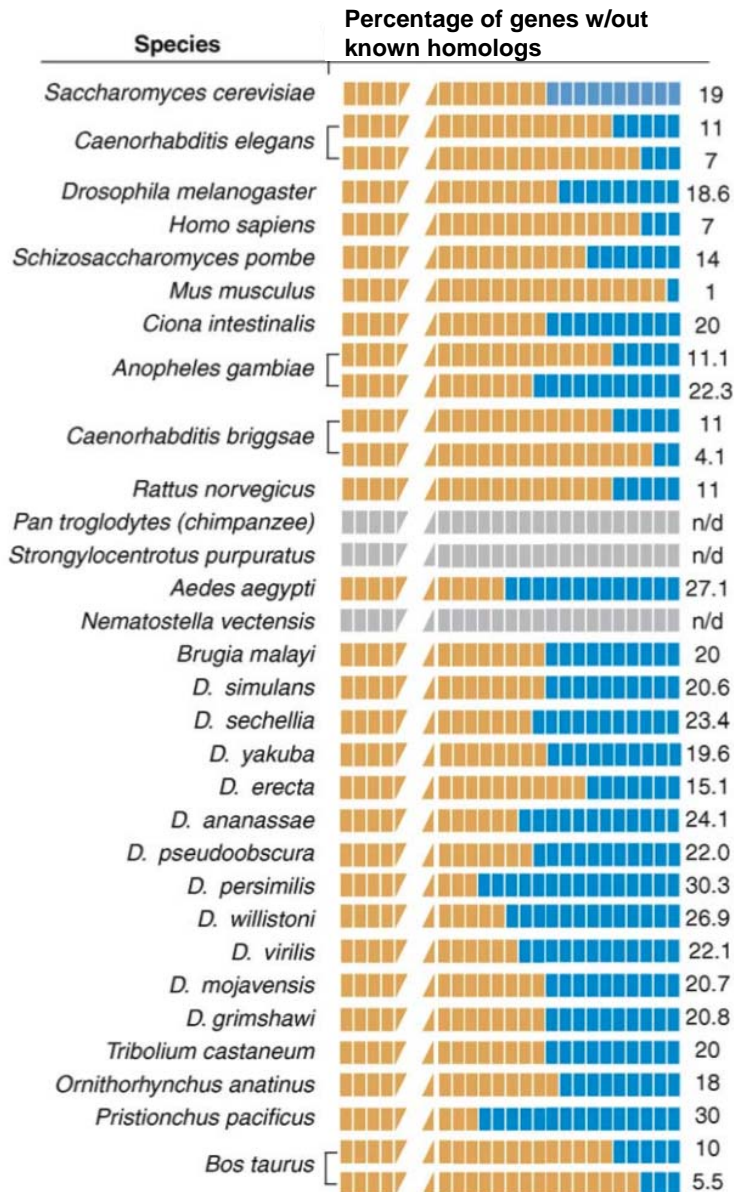
QuickTime™ et un
décompresseur
sont requis pour visionner cette image.

FROSTO

Un outil pour la détection de protéines homologues distantes

G. Launay, G. Collet, N. Maillet, O. Rousselet,
A. Cornu, A. Marin, R. Andonov, J-F. Gibrat

Génomique comparative



Nouvel organisme:

10 à 20 % de gènes orphelins

Spécifiques de(s) taxon(s)
« TRG »

Sensibilité des méthodes
de détection d'homologues

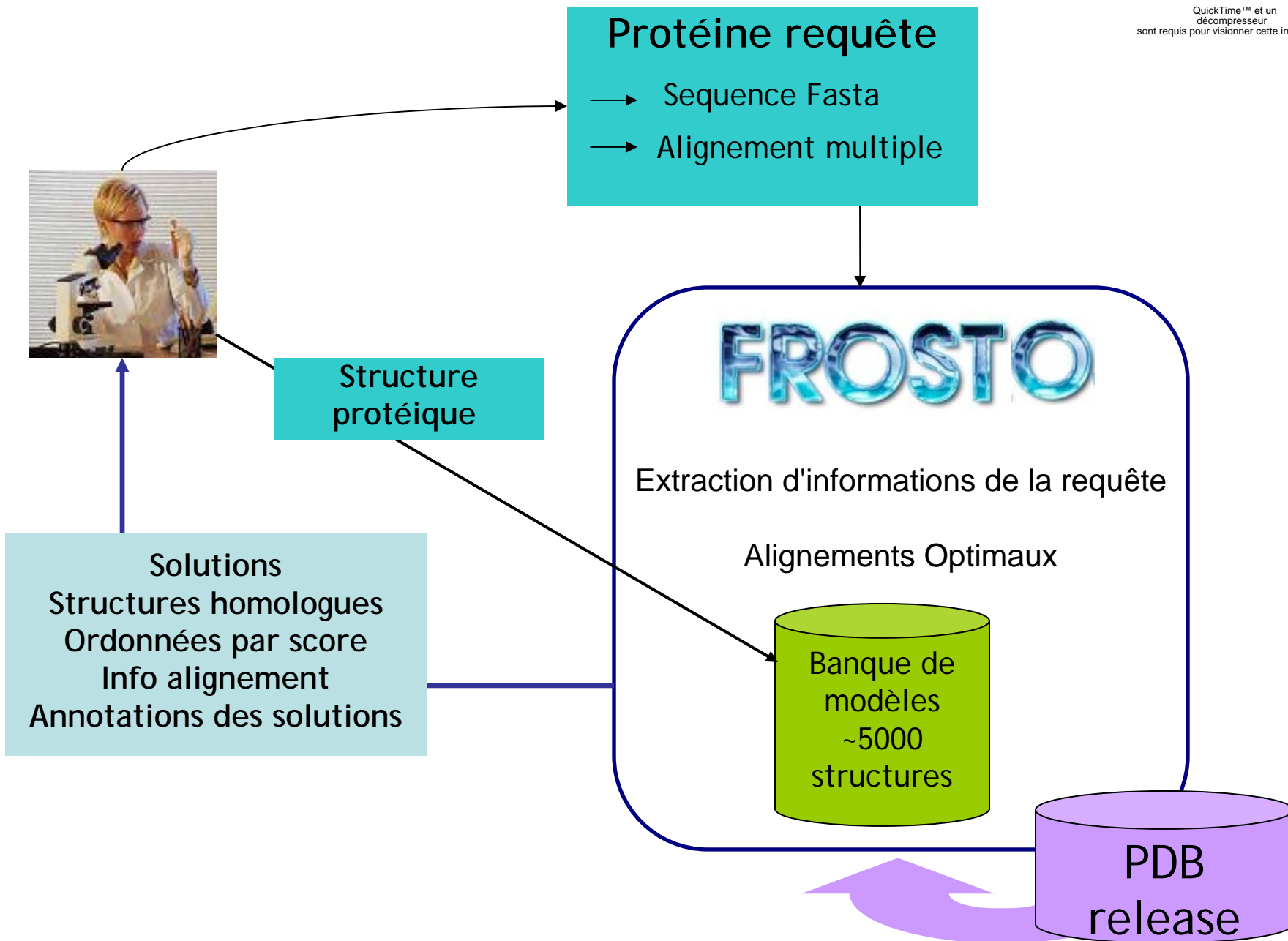
La reconnaissance de repliement

QuickTime™ et un
décompresseur
sont requis pour visionner cette image.

- La structure porte la fonction
- La structure est plus conservée que la séquence

QuickTime™ et un
décompresseur BMP
sont requis pour visionner cette image.

➤ Alignement séquence / structure
Sensibilité accrue

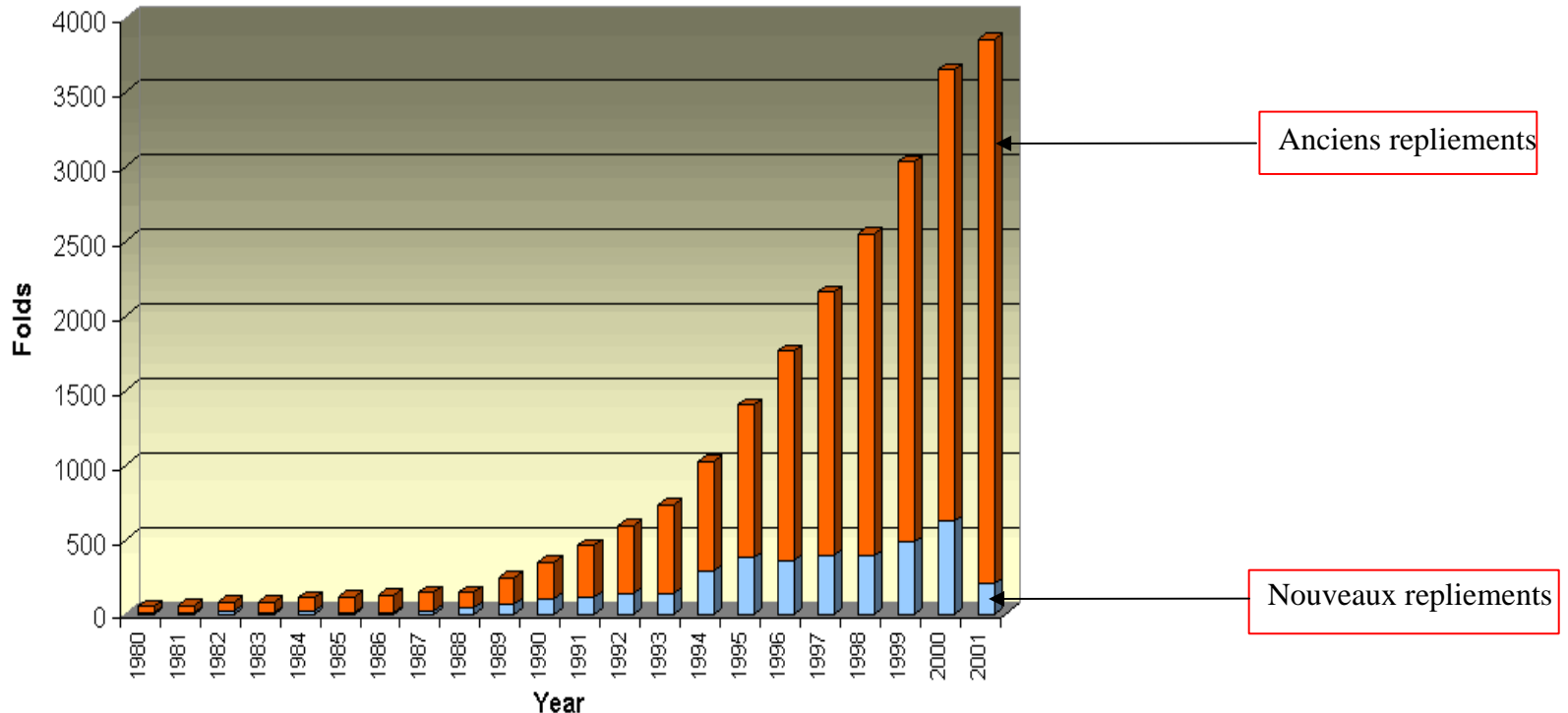


Reconnaissance de repliements ingrédients

- Une base de données de ~~séquences~~ structures
- Fonction(s) de scores quantifiant la compatibilité entre séquence(s) requête(s) et structures de la base de données
- Optimiseur la(es) fonction(s)
- Affinement du modèle

Banque de modèles

→ Croissance de la PDB

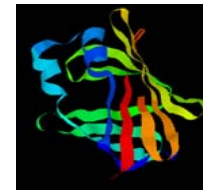
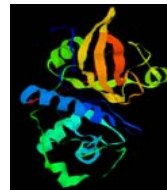
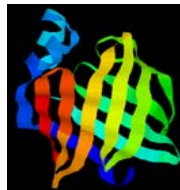
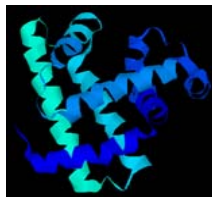


Nombre de repliements uniques relativement limité

Sur 3 ans, 90% des structures déposées dans la PDB sont proches de repliements déjà connus

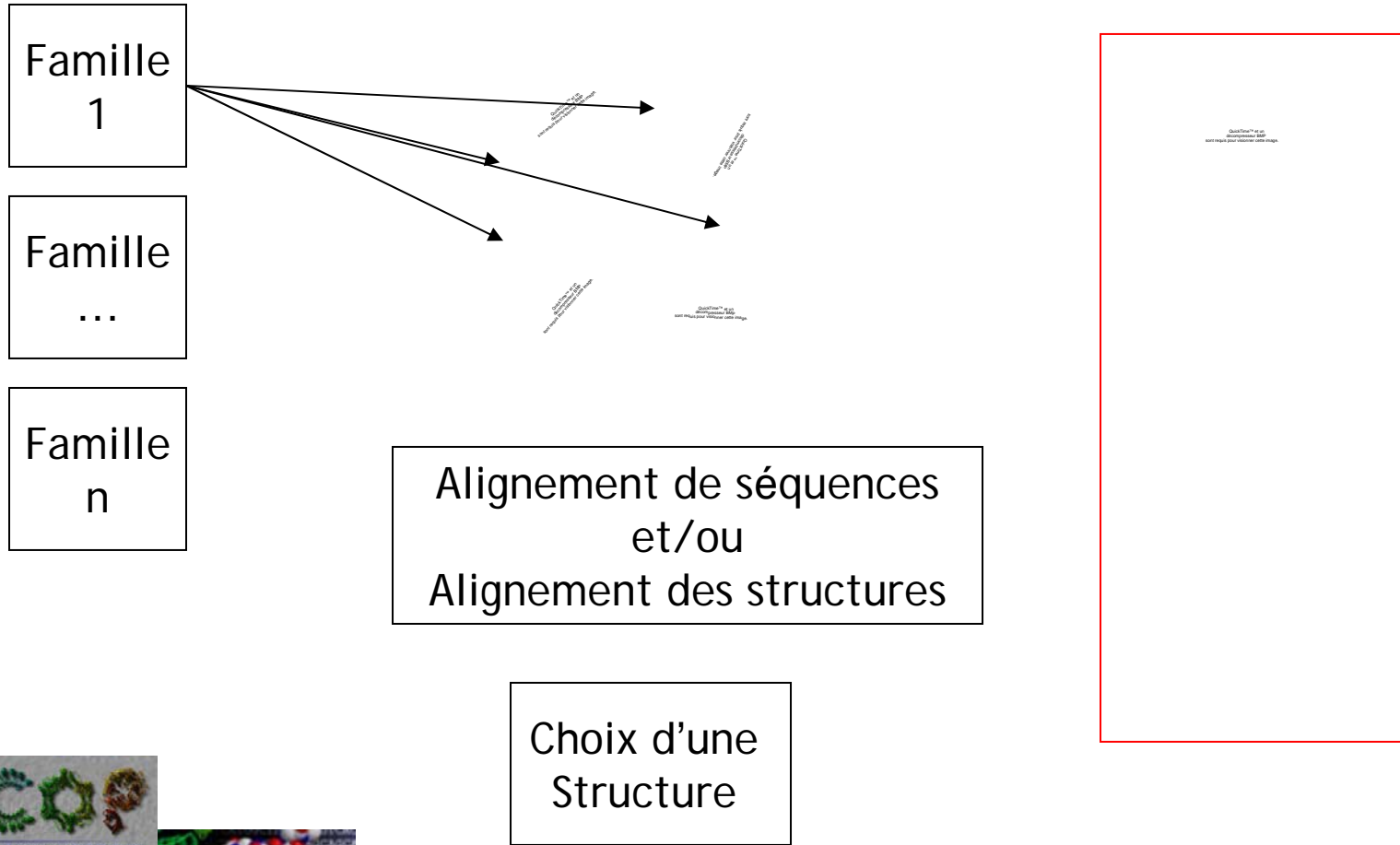
Banque de modèles

- Couverture d'un large spectre de protéines
- Résolution de chaque modèle de bonne qualité ($< 3.0\text{\AA}$)
- On préférera une structure cristallographique à une déterminée par RMN
- Identité de séquence entre les modèles $< 30\%$
(minimiser la redondance structurale)
- Régions les plus conservées



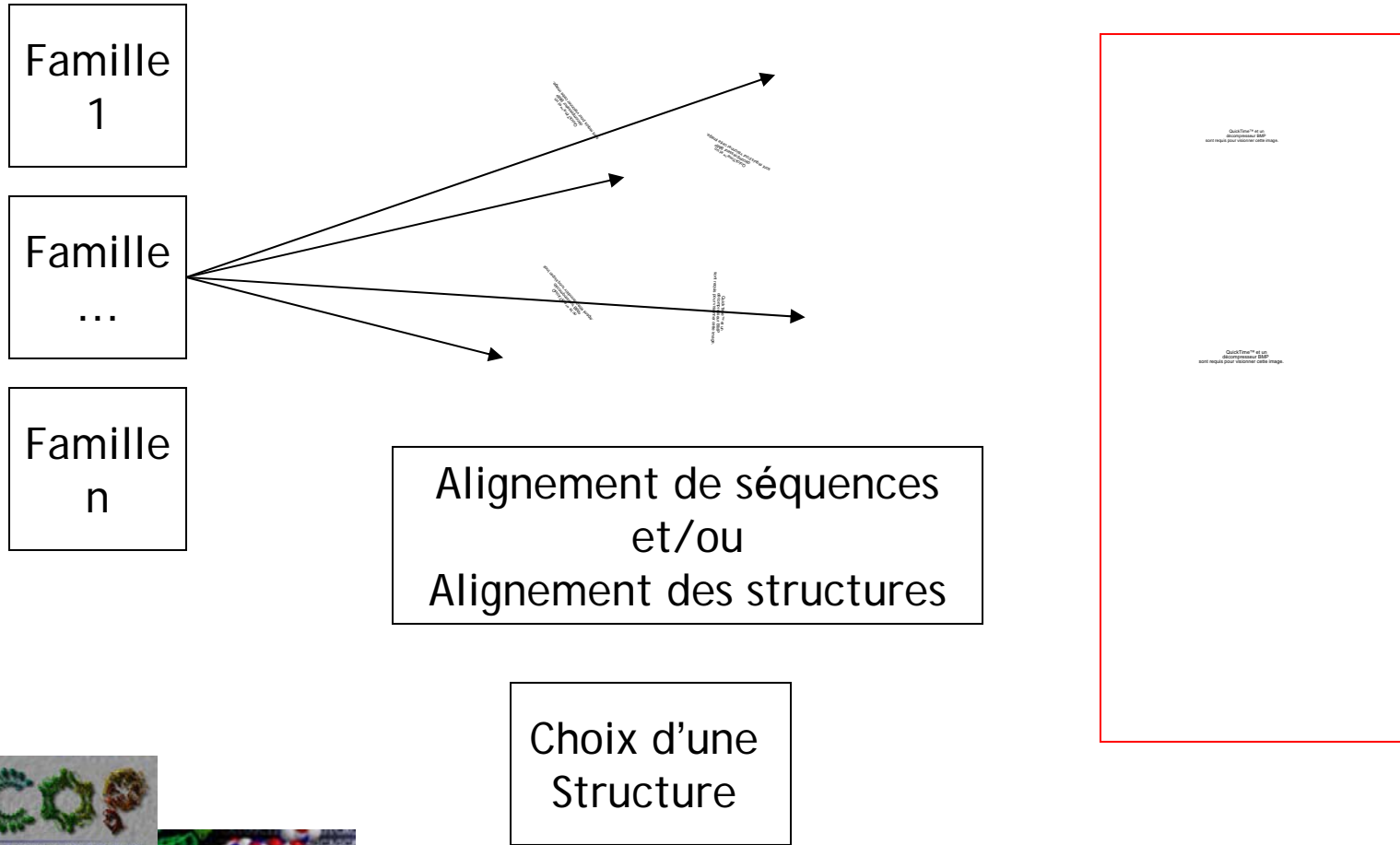
Construire la banque de modèles

→ Modèles représentatifs des structures connues



Construire la banque de modèles

→ Modèles représentatifs des structures connues



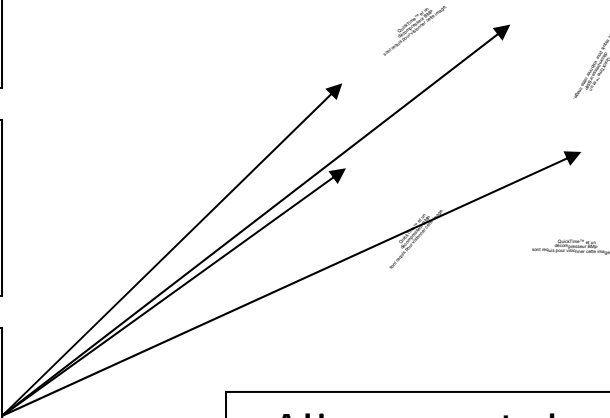
Construire la banque de modèles

→ Modèles représentatifs des structures connues

Famille
1

Famille
...

Famille
n



Alignement de séquences
et/ou
Alignement des structures

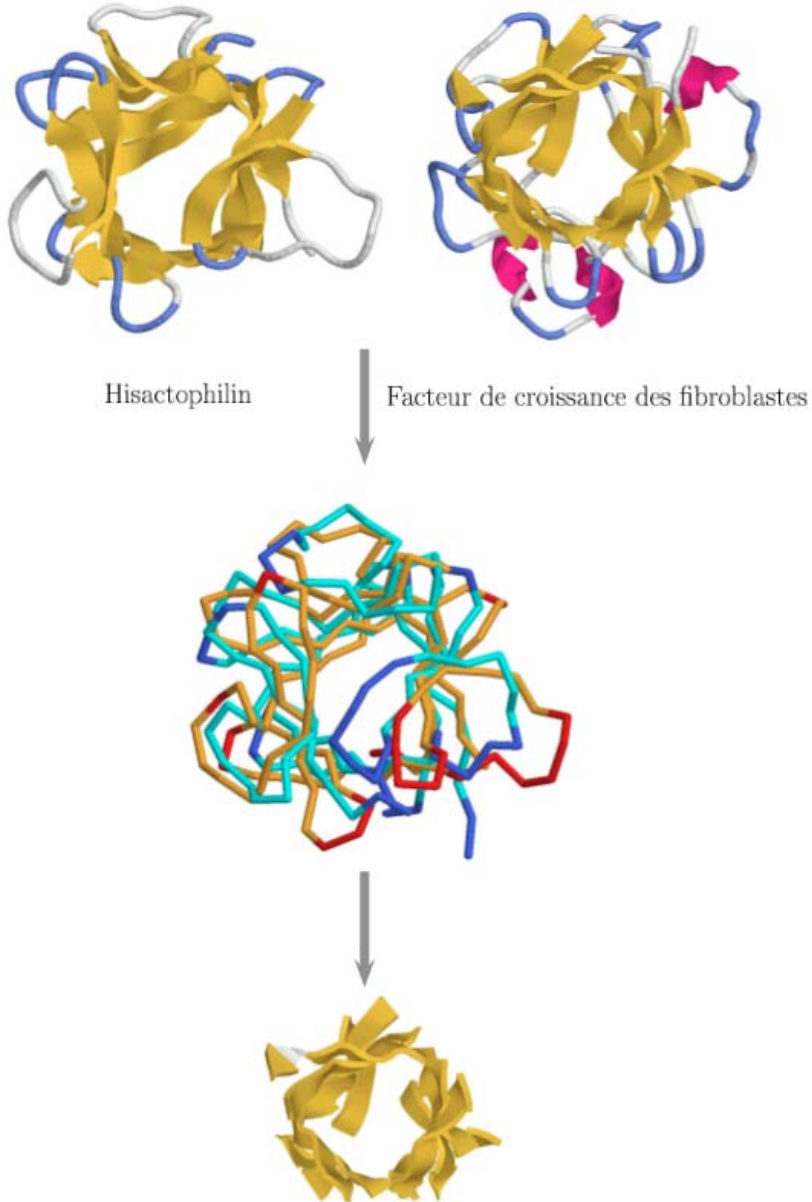
Choix d'une
Structure



Banque de
5000 modèles



Construire la banque de modèles



Régions des structures les plus conservées

Éléments de structures secondaires

→ Pas de gaps dans ces régions

Reconnaissance de repliements ingrédients

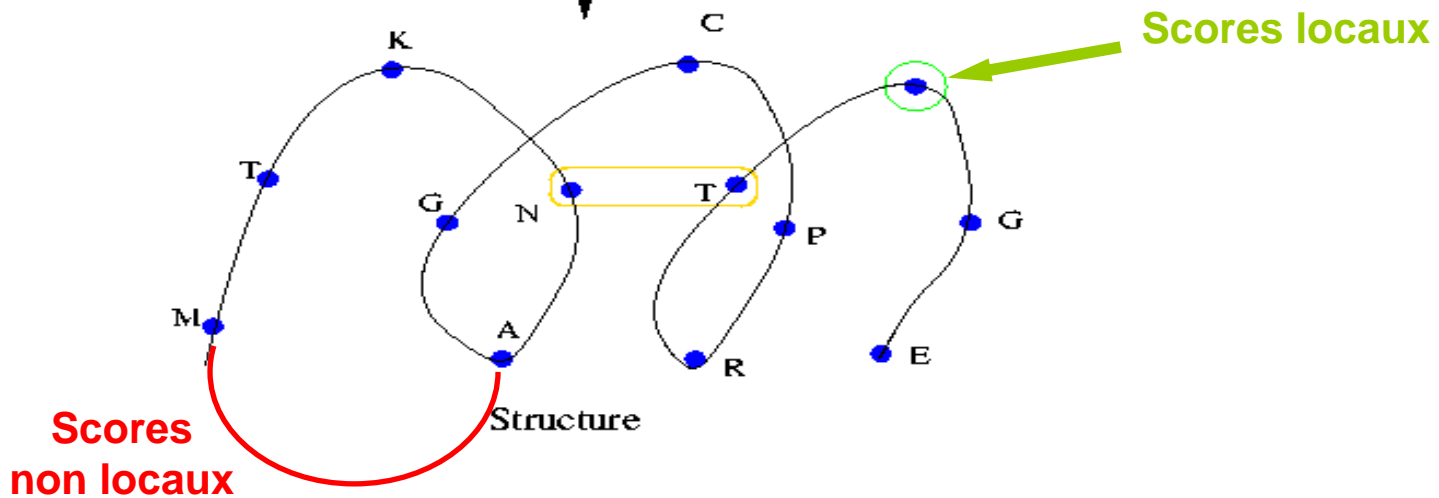
- Une base de données de structures
- **Fonction(s) de scores** quantifiant la compatibilité entre séquence(s) requête(s) et structures de la base de données
- Optimiseur la(es) fonction(s)
- Affinement du modèle

Fonction de score

$$f(\text{séquence, modèle})$$

Sequence: M T K L I L N A G C P R T G E W T Y T E

threading

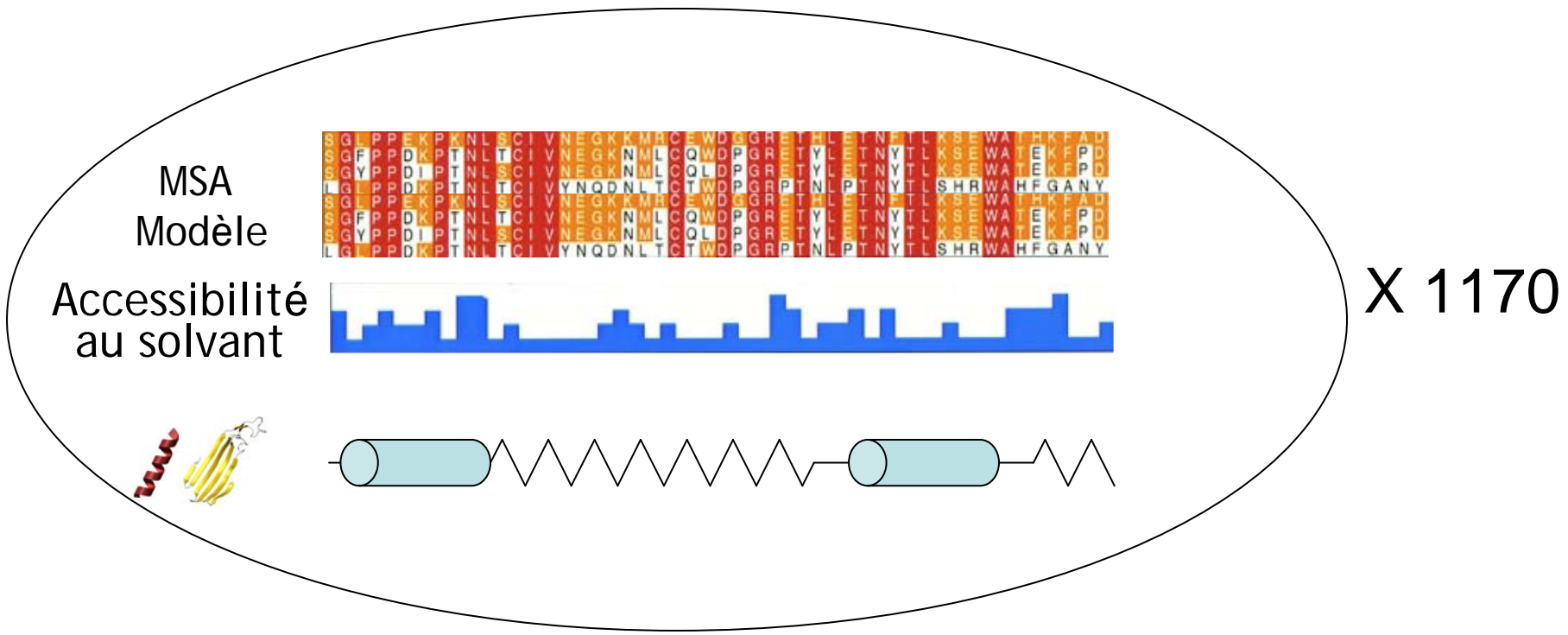


f est une somme de termes (scores)

Maximiser f pour trouver l'alignement entre
ma séquence et chaque modèle

Formes des scores locaux

➤ Paramétrage sur : Banque de 1170 structures protéiques



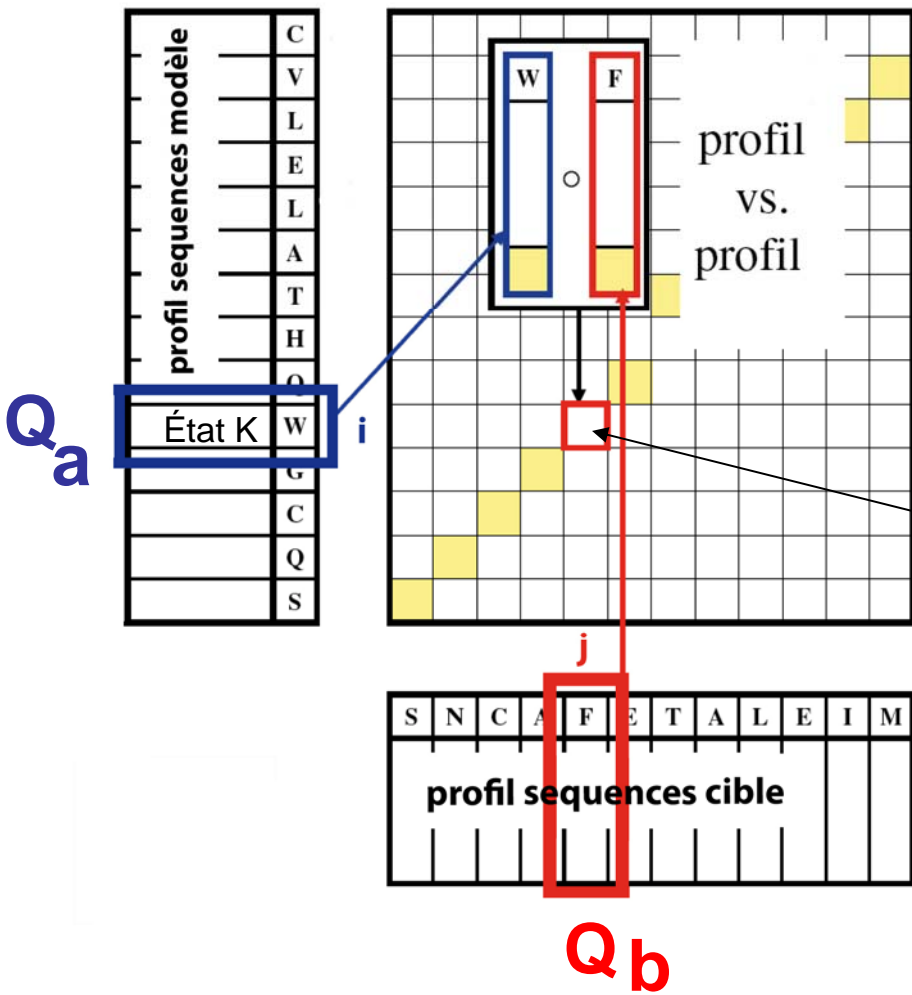
Définition de 9
Etats structuraux (S)

9 matrices
Blosum-like

$$\text{score}_{S_k}(r, R) = 2 \log_2 \frac{P_{S_k}(r, R)}{P_{S_k}(r) P_{S_k}(R)}$$

1D

Fonction de scores locaux Optimisation



Somme de scores indépendants

➤ Programmation dynamique

$$S_{i,j} = \sum_{a=1}^{20} \sum_{b=1}^{20} Q_a^i Q_b^j M_{ab}^K$$

M est la matrice de substitution déterminée par l'état de la position i

FROSTO

Alignement Séquence / Structure

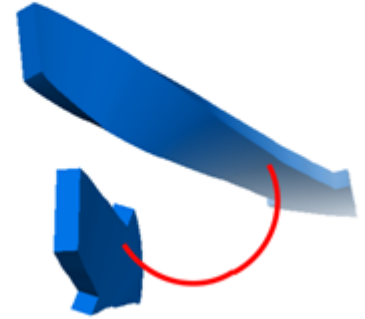
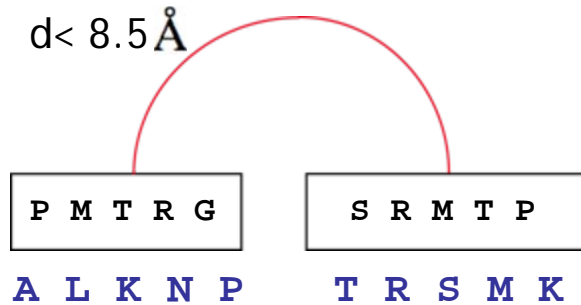
2 Classes de fonctions de scores

- Fonctions de scores locaux 1D
 - Programmation dynamique
- Fonctions de scores non locaux 3D

3D

Formes des scores non locaux

➤ Interactions entre acides aminés



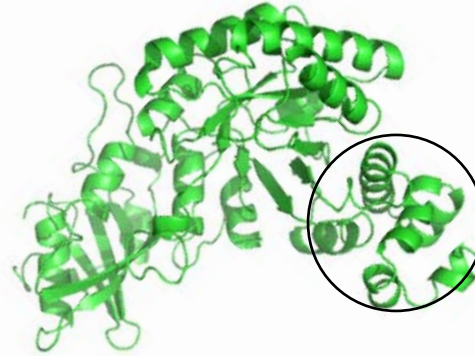
Potentiels distance dépendants

Terme de contacts

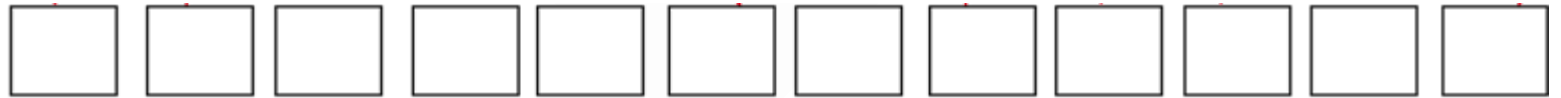
Coût de **substitution** d'une paire d'acides aminés en contact dans une paire d'états donnée par une autre paire d'acides aminés .
(Généralisation des scores 1D)

Fonctions de scores non locaux

3D Optimisation



Structure
modèle



F—V—N—H—K—S—R—E—A—L—F—D

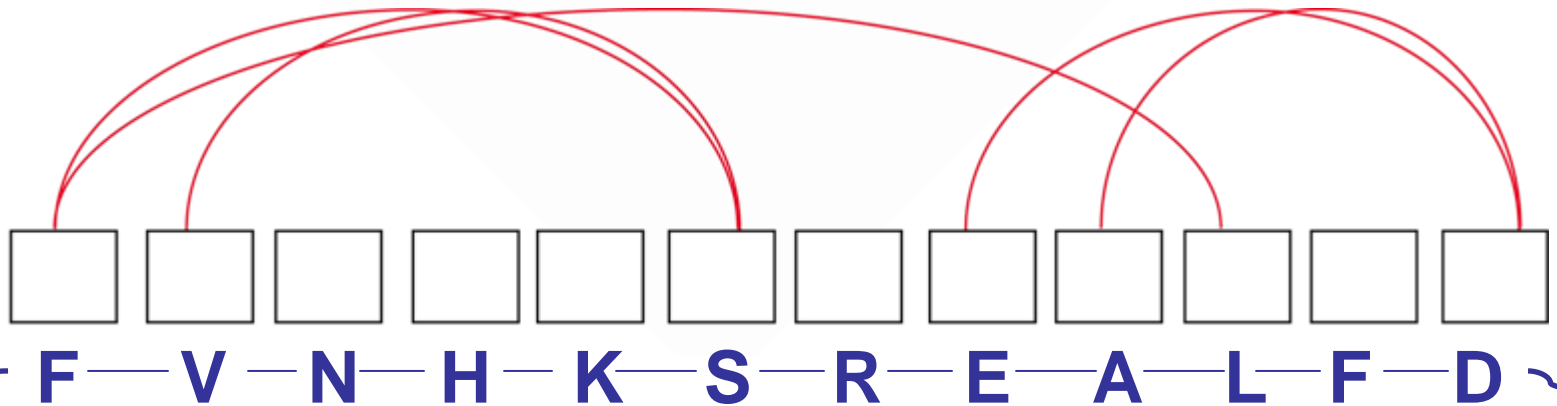
Séquence requête

Fonctions de scores non locaux

3D Optimisation



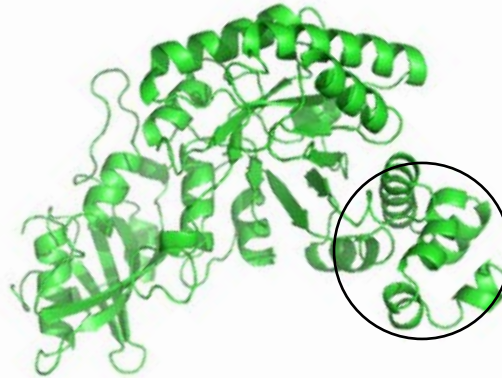
Structure
modèle



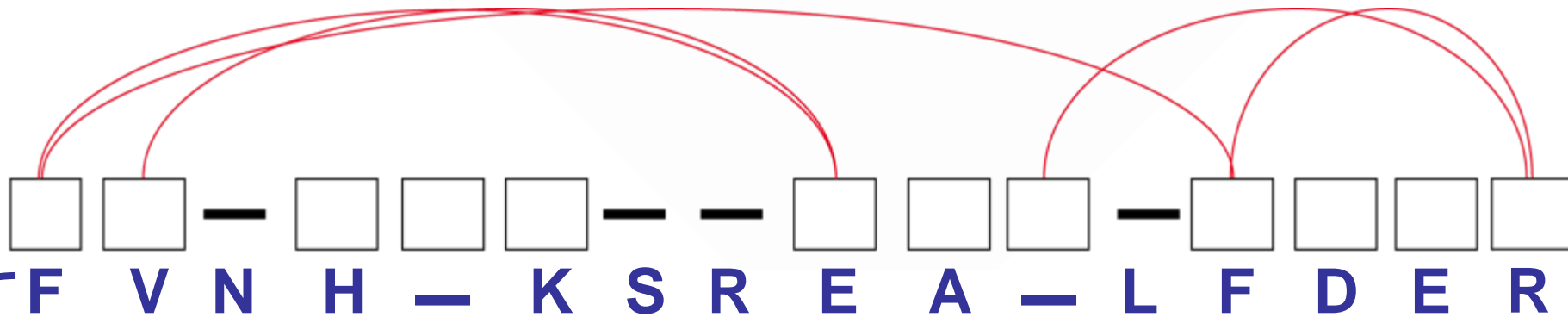
Séquence requête

Fonctions de scores non locaux

3D Optimisation



Structure modèle



Séquence requête

Problème NP
complet

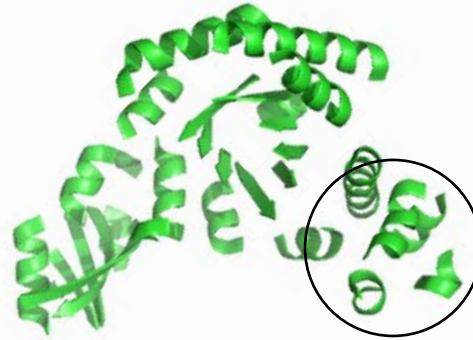
Lathrop 1998

Fonctions de scores non locaux

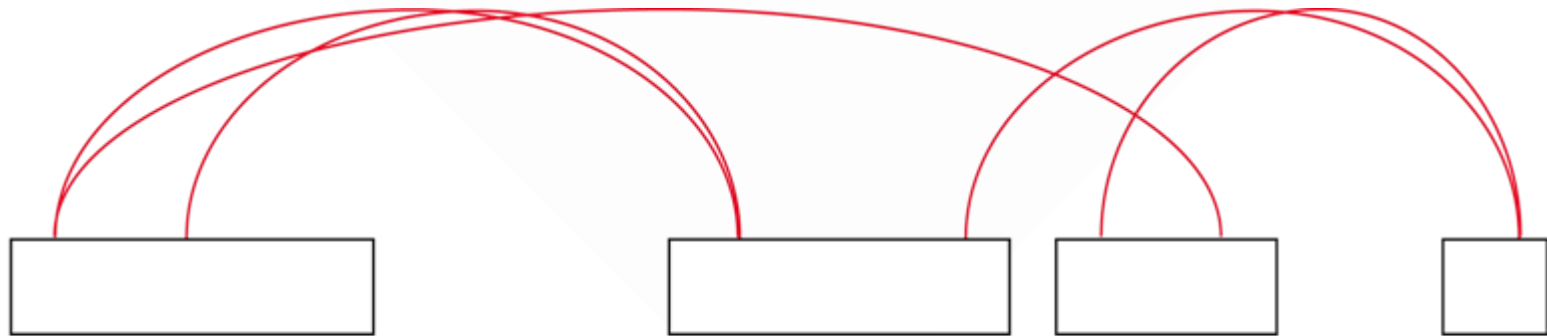
3D

Optimisation

Blocs structuraux reliés par arcs somme des interactions entre les acides aminés de chaque bloc



Structure modèle



F—V—N—H—K—S—R—E—A—L—F—D

Séquence requête

➤ Branch&Bound

Programmation linéaire en nombre entier

FROSTO

Alignement Séquence / Structure

2 Classes de fonctions de scores

- Fonctions de scores locaux

- Programmation dynamique

1D

- Fonctions de scores non locaux

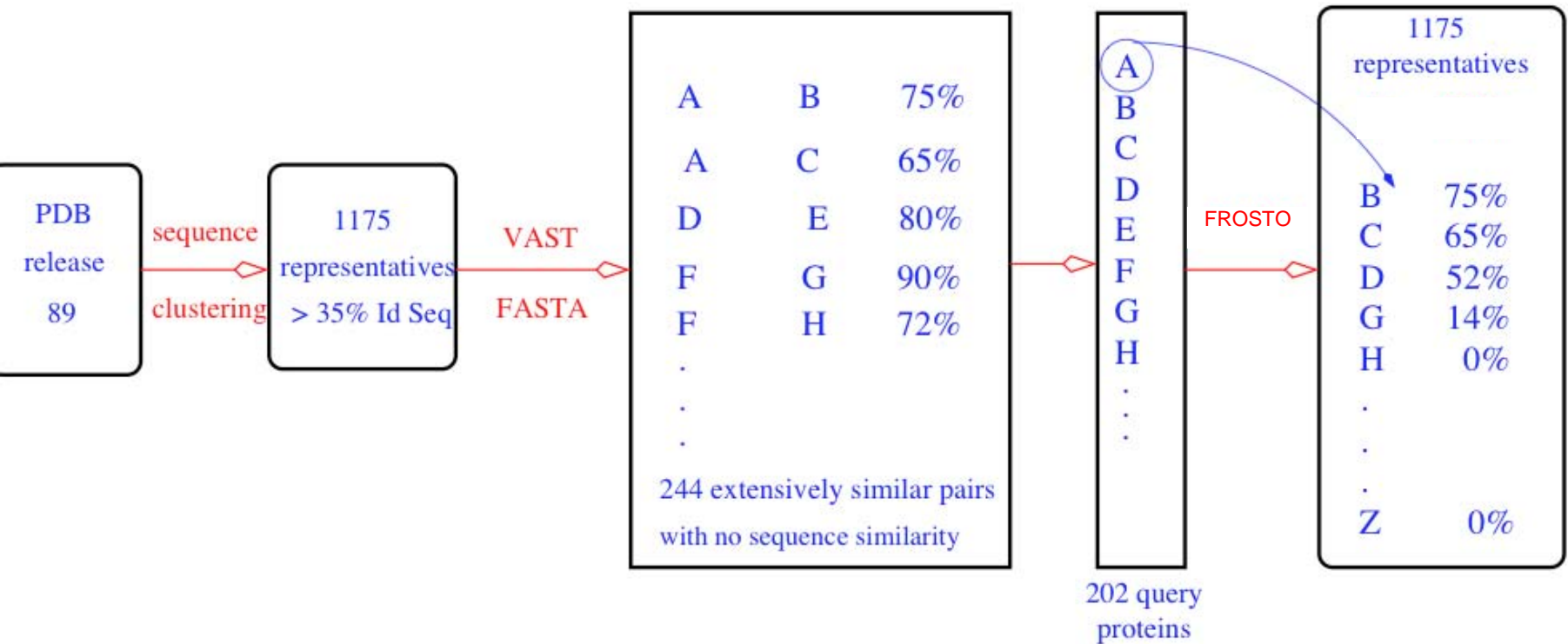
- Branch&Bound

- Programmation linéaire en nombres entiers

3D

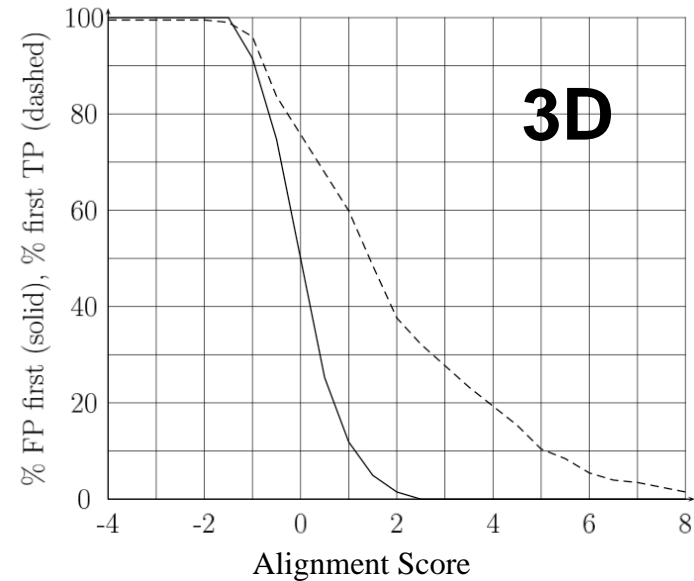
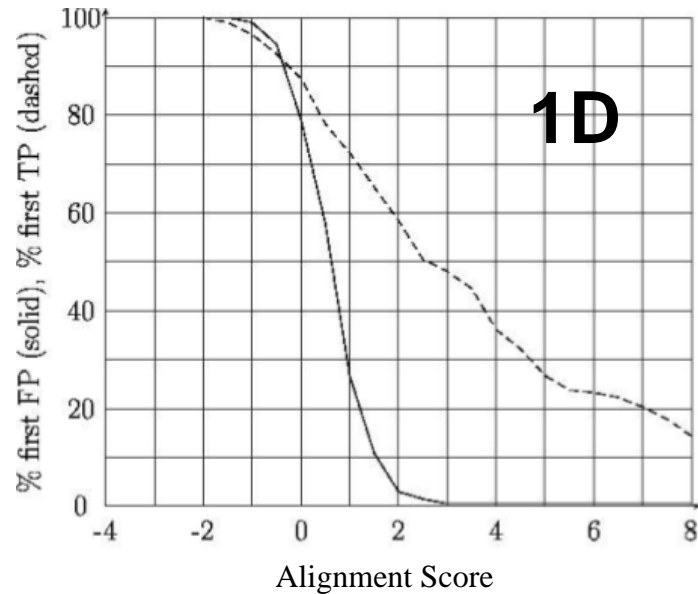
FROSTO

Evaluation des performances



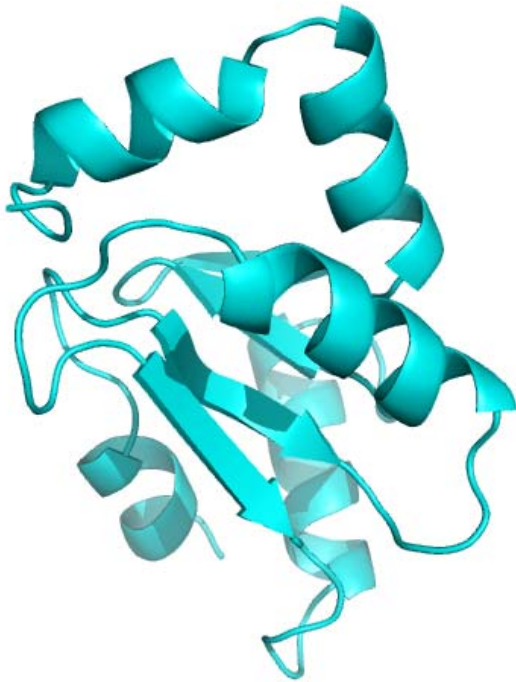
Jeux test: 202 requêtes vs 1174 modèles

Sensibilité des fonctions de scores

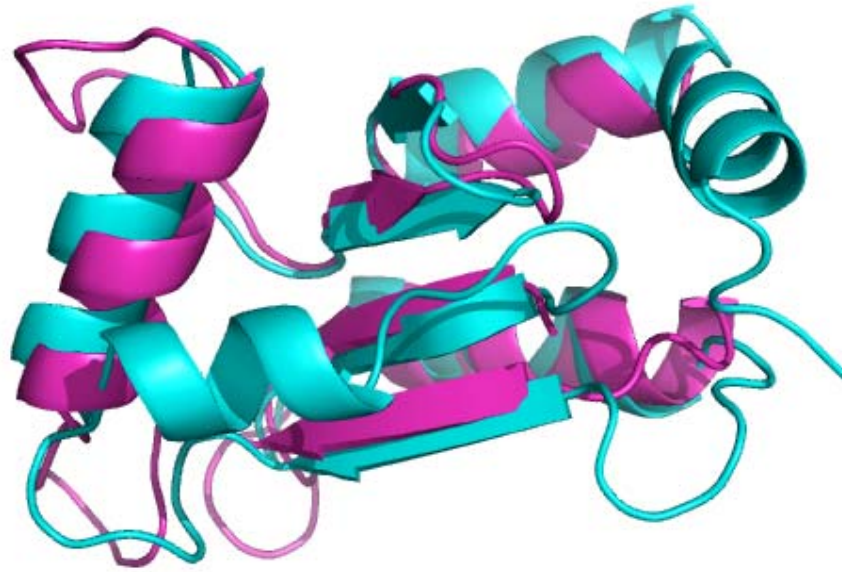


Rates of false positives	1%	5%	10%
Normalized distance	3.2	2.6	2.3
Coverage 1D filter	51%	60%	65%
Normalized distance	3.0	2.5	1.8
Coverage 3D filter	38%	49%	60%
<i>E</i> value	0.2	1.1	2.5
Coverage PSI-BLAST	33%	40%	42%
<i>E</i> value	0.09	0.7	1.4
Coverage 3D-PSSM	56%	74%	79%

Problèmes liés à la définition de blocs structuraux



Problèmes liés à la définition de blocs structuraux



Problèmes liés à la définition de blocs structuraux

Norm score : 0.337084
Raw score : -8.69376
Query name : labaA
Template name: lkteA

Template SSE : HHHHHHHHHH--bbbbbb--HHHHHHHHHHHHHHH---bbbbbb---HHHHHHHHHHH--bbb---HHHHHHHHHH-HHHHHHHHHHH

Template Seq : AOAFVNSKI--KVVVFI--CPFCRKTQELLSQL---LLEFVD---NEIODYLOOLT--RVFI---GCTDLESMHK-GELLTRLQQVG

Query Seq : MFKVYGYDSNIHKCGPCDNAKRLTVKKQPFEFINIMPEKGVFDDEKIAELLTKLGRDTQIGLTMPOVFAPDGSHIGGFDOLREYFK

Query SSE* : .bbbbbbb.....HHHHHHHHHHHH.....bbbb.....HHHHHHH.HHHHHHHHHHH.....bbbb...bbb.HHHHHHHH.

(*: Predicted by PsiPred)

Problèmes liés à la définition de blocs structuraux

```
Norm score   : 0.337084  
Raw score    : -8.69376  
Query name   : labaA  
Template name: lkteA
```

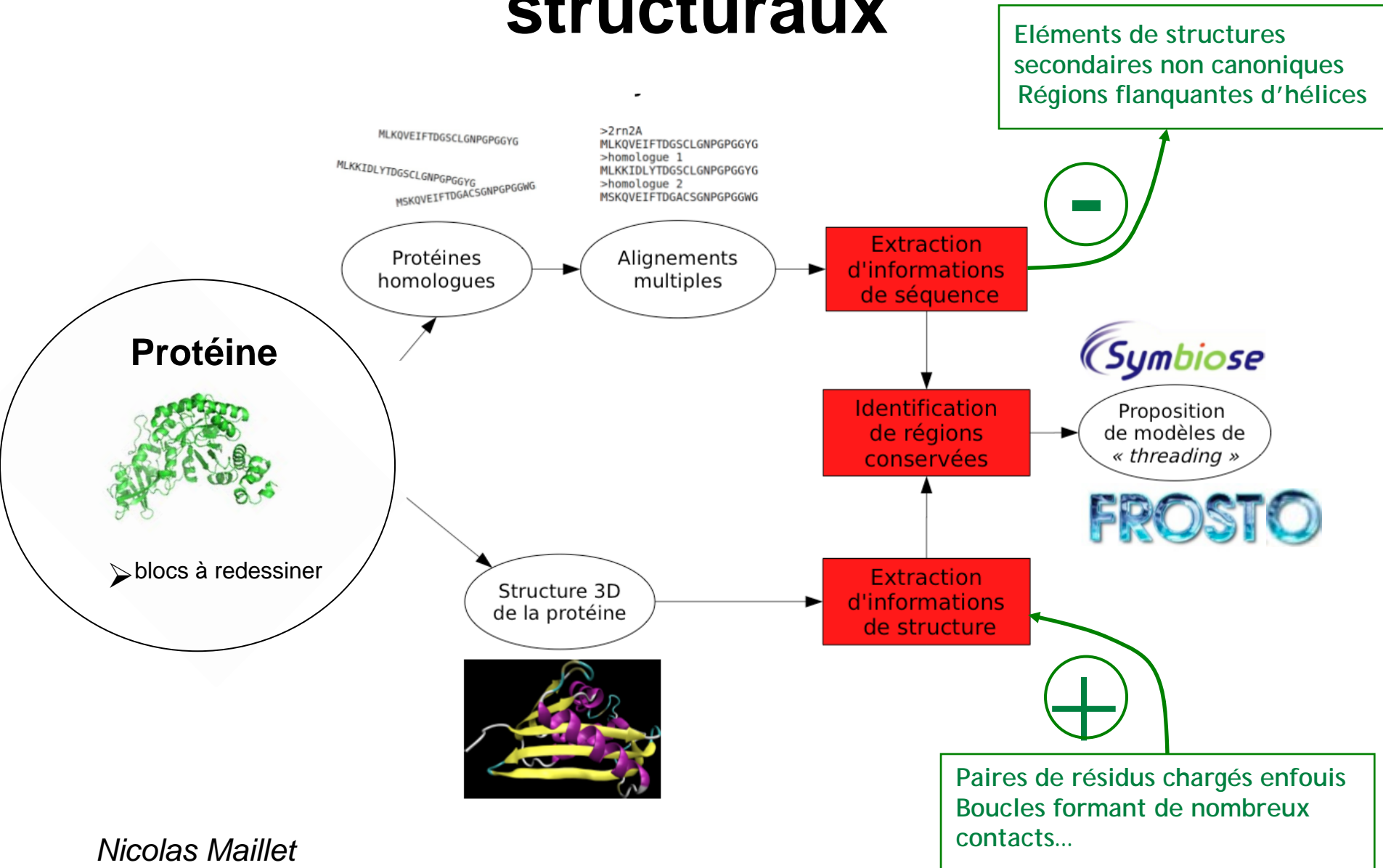
```
Template SSE : HHHHHHHHHH--bbbbbb--HHHHHHHHHHHHHHH---bbbbbb---HHHHHHHHHHHH--bbb---HHHHHHHHHH-HHHHHHHHHHH  
Template Seq : AOAFVNSKI--KVVVFI--CPFCRKTQELLSQL---LLEFVD---NEIODYLOOLT--RVFI---GCTDLESMHK-GELLTRLQQVG  
Query Seq    : MFKVYGYDSNIHKCGPCDNAKRLTVKKQPFEFINIMPEKGVFDDEKIAELLTKLGRDTQIGLTMPOVFAPDGS HIGGFDOLREYFK  
Query SSE*   : .bbbbbbb.....HHHHHHHHHHHH.....bbb.....HHHHHHH.HHHHHHHHHHH.....bbb...bbb.HHHHHHHH.
```

(*: Predicted by PsiPred)

La totalité des acides aminés des structures secondaires -> **faible sensibilité**

Nécessité de **redécouper** les protéines modèles en nouveaux blocs structuraux

Vers une autre définition des cœurs structuraux



Travaux en cours

« TOUT 3D »

- Combinaisons linéaires des scores 1D et 3D.
- Extrémités de structures secondaires accessoires
- Ajout de termes:
 - Profils de mutations distants
 - Termes de solvation

Participation à CASP

- Déploiement sur Genocluster
- Parrallélisation du code