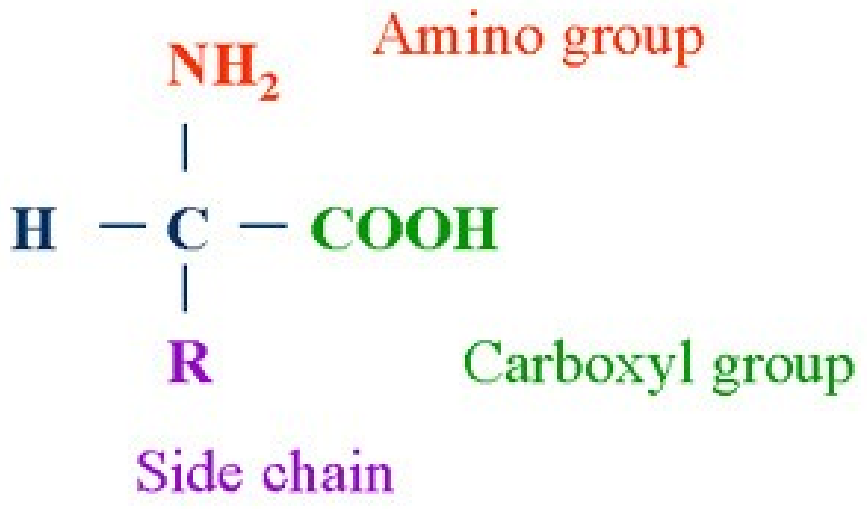
Prédiction des structure 3D des protéines

Karim Mezhoud
Ir. agronome
PhD. Toxicologie, Protéomique, Bioinformatique

Centre national des Sciences et Technologies Nucléaires
Sidi Thabet – Tunis

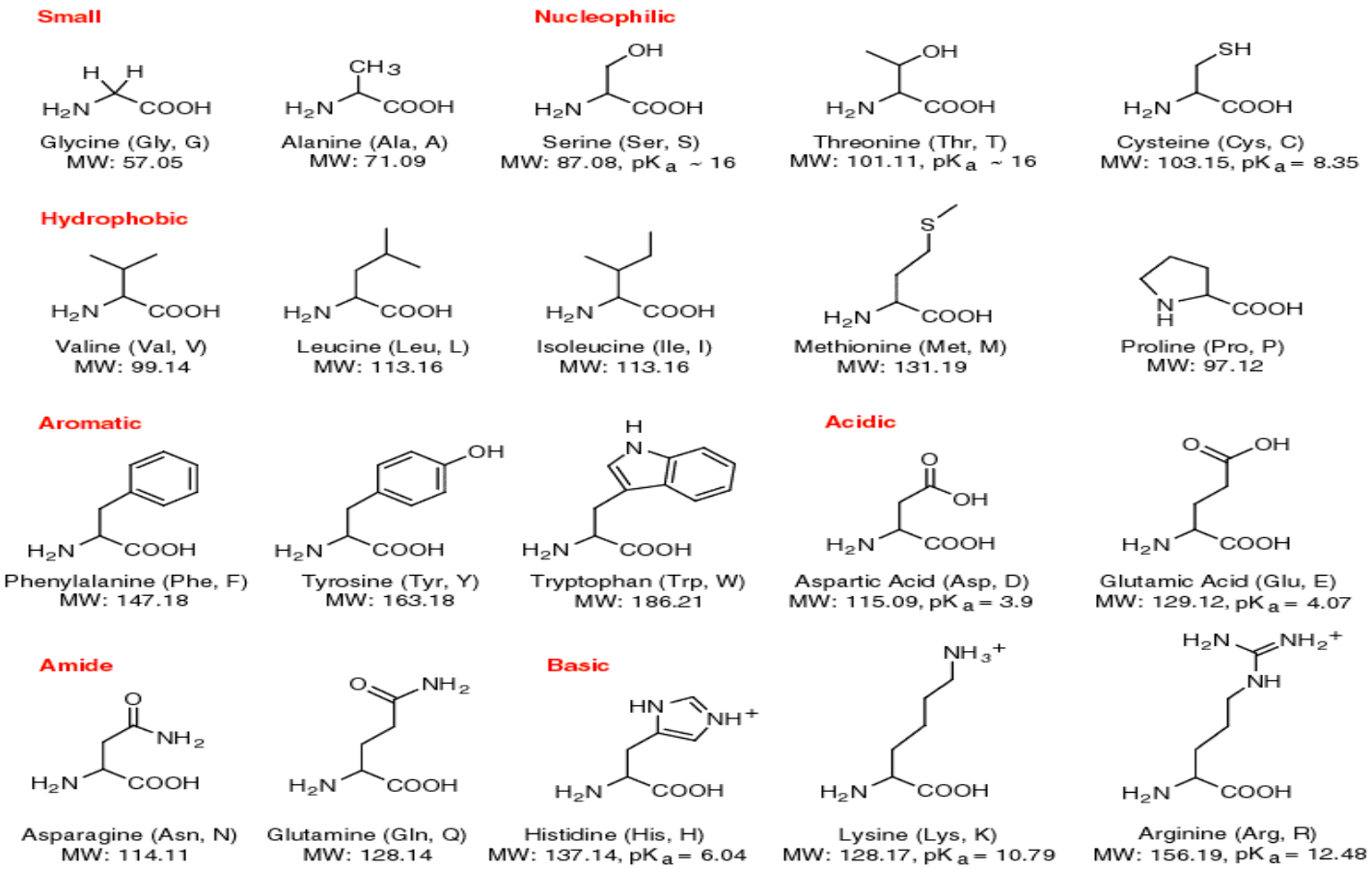
Les Protéines – Quelques notions de bases

- Qu'est ce qu'une protéine?
  - C'est une séquence linaires d'acide aminées...
- Qu'est ce qu'un acide aminée?

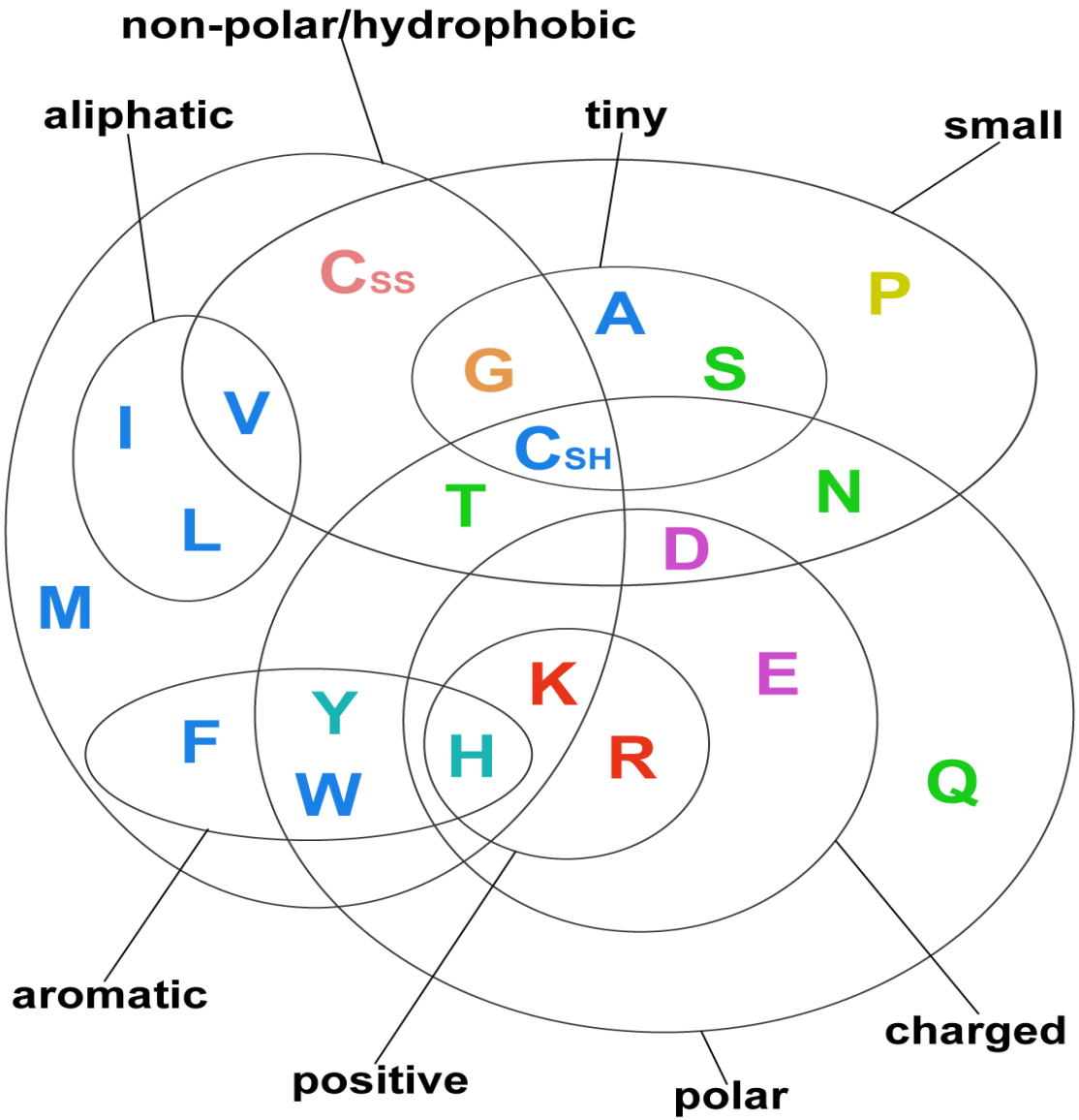


Les Protéines – Quelques notions de bases

- Combien y a t il d'acide aminée?

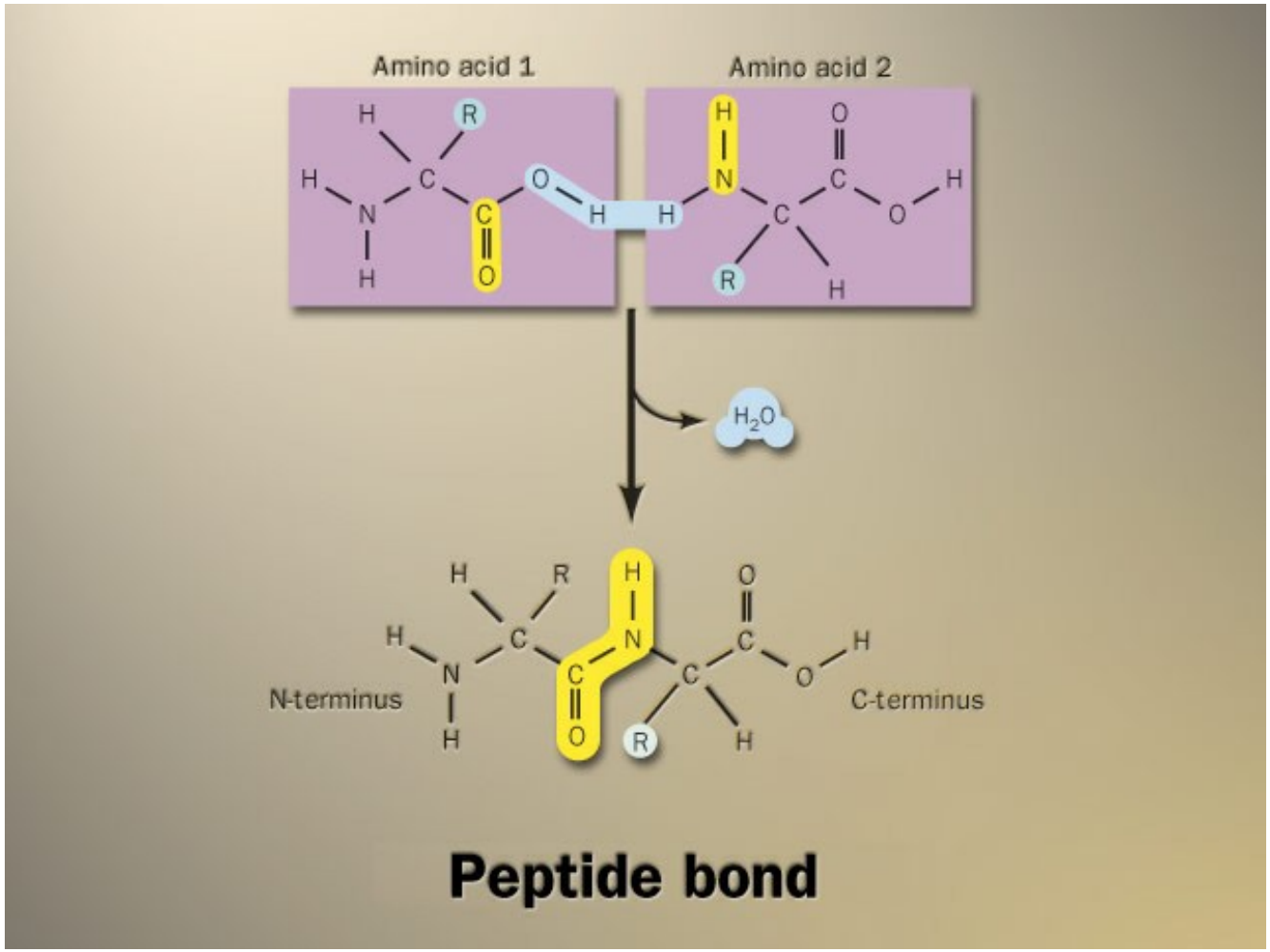


Les Protéines – Quelques notions de bases



Les Protéines – Quelques notions de base

- Les acides aminées se lient via Les Liaisons peptidiques

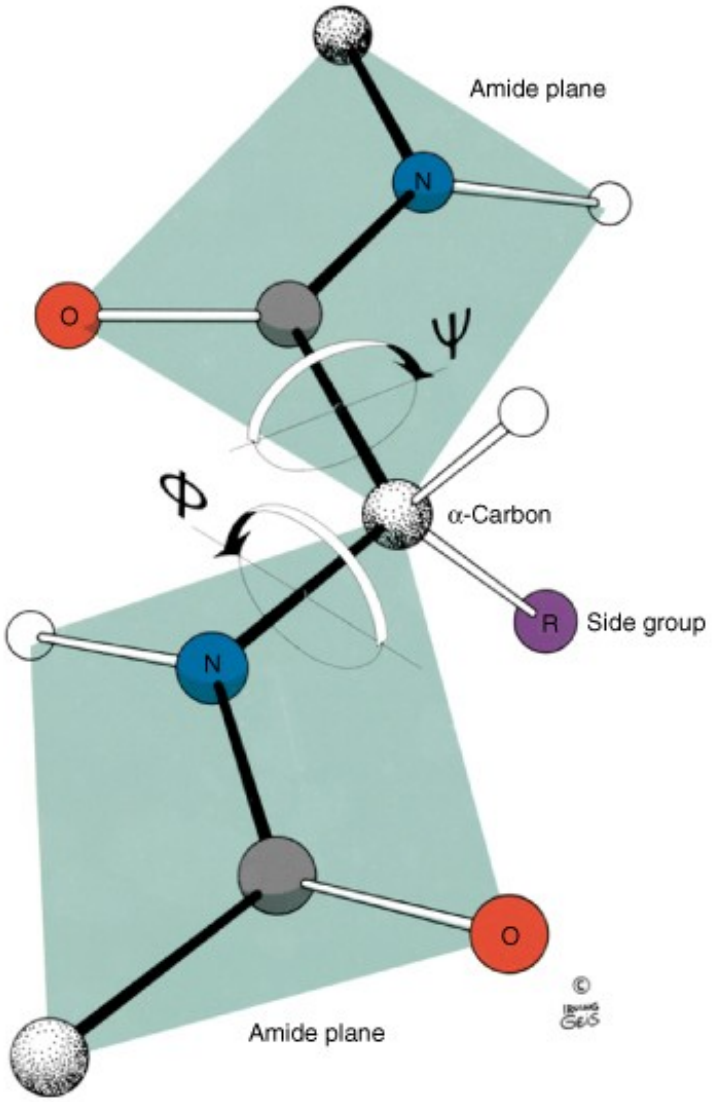


Les Protéines – Quelques notions de bases

- La séquence (Backbone) peut pivoter:

  Angle diédral (angle entre deux plans)

- Une protéine de 100 AA peut avoir 99 torsions!
  - Grande liberté de mouvement



Les Protéines – Quelques notions de bases

- Qu'est ce qu'elle fait une protéine?
  - Enzymes
  - Cellular Signaling
  - Antibodies
  - Structural protein

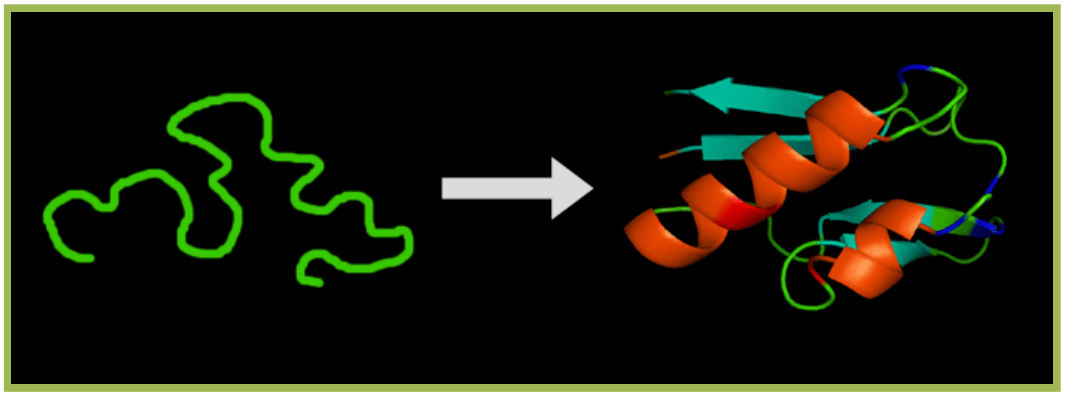
Qu'est ce qu'elle ne fait pas?

  - Réserve d'énergie (glucide, lipide) ou d'information (ADN, ARN)

- Vient du mot grec *Proteios* – PRIMAIRE
- Fondamental à la quasi-totalité des processus cellulaires

Séquence → Protéine

- Comment une protéine fait autant?
  - Les protéines se replient spontanément
  - Prennent une forme 3D spécifique
  - La forme dépend notamment des séquences en AA
  - La forme confère une fonction spécifique



La fonction d'une Protéine

- STRUCTURE 3D → FONCTION : Ce n'est pas automatique
- La détermination de la structure 3D d'une protéine ne résout pas sa fonction
- Deux techniques principales permettent de déterminer la structure:
  - Cristallographie aux rayons X
  - Résonance magnétique nucléaire

    → méthodes lourdes et beaucoup de protéines ne cèdent pas à ces méthodes

Niveau d'organisation des protéines

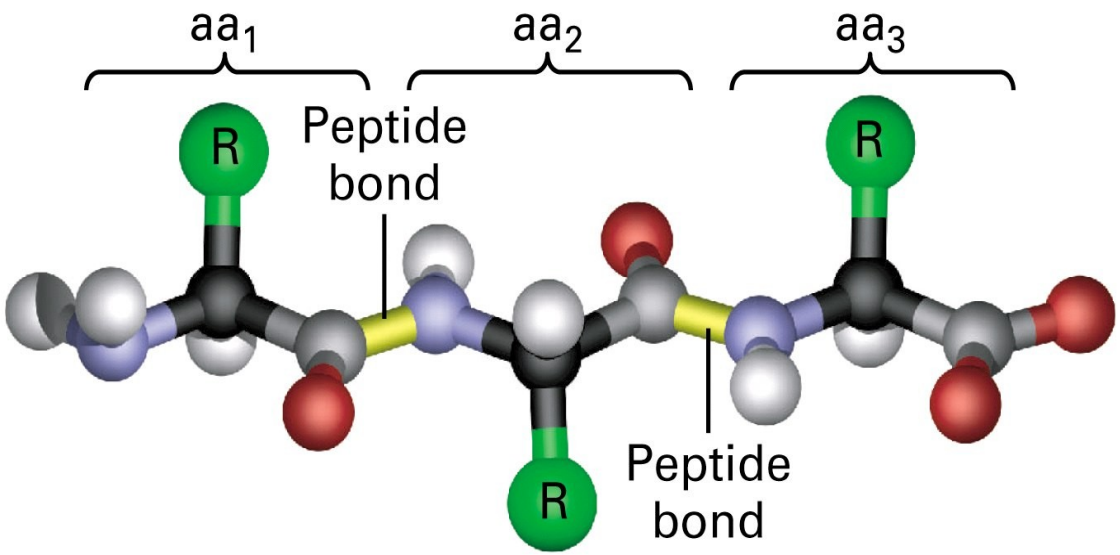
Les structures de protéines sont souvent décrite sur 4 niveaux différents:

- Structure primaire
- Structure secondaire
- Structure tertiaire
- Structure Quaternaire

Organisation Hydrophobique/Hydrophilique

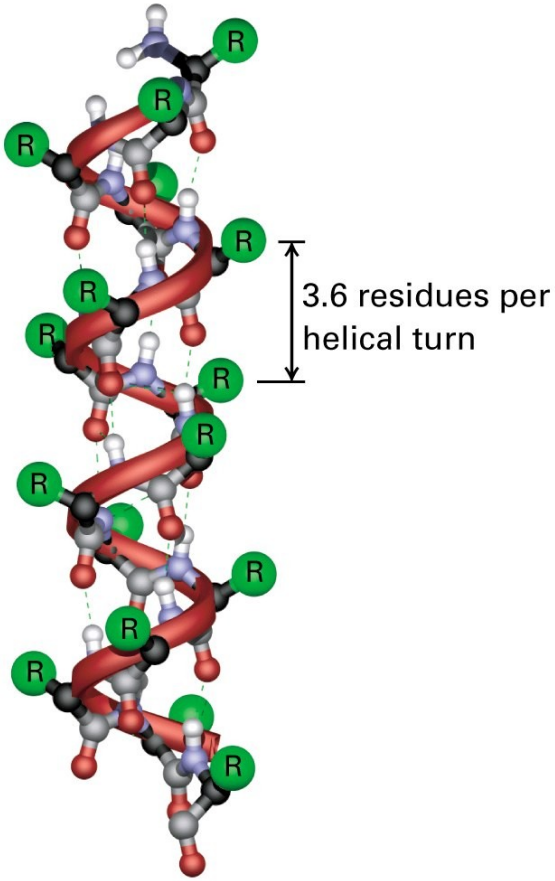
- La partie Hydrophobe se trouve à l'intérieur de la protéine

Niveau de description structural

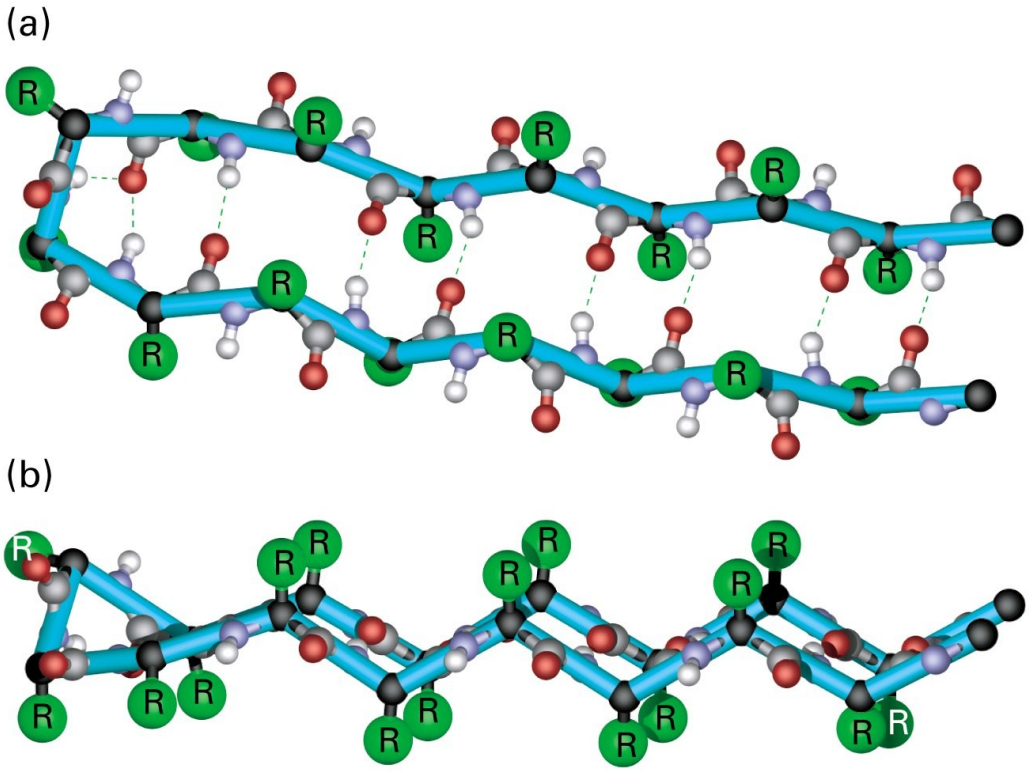


Structure primaire
Séquence d'acide aminée

Niveau de description structural



Hélice α



Feuillet β

3éme catégorie de str 2sd :
Coil (rouleau) ou loop (boucle)

Structure secondaire

Niveau de description

Structure tertiaire
Folded individual peptide

Structure quaternaire
Agrégation de deux ou plusieurs peptides

Le Repliement des protéines

Ce que nous savons:

- Repliement des protéines est rapide!
  - Généralement, un couple de secondes
- Le pliage est cohérent!!
- Appel à des forces faibles - Non-covalent
    - Liaison hydrogène, van der Waals, ponts salins
- Très peu d'intermédiaires
- Il est donc difficile d'étudier

Le Repliement des protéines

Quelles sont les propritétés physiques qui stabilisent les repliments :

1) La rigidité des backbones (Hélice α ; Feuillet β)

2)Les interaction entre acide aminées

- Interactions electrostatiques
- Forces de Van der Waals
- Ponts Hydrogène et disulfure

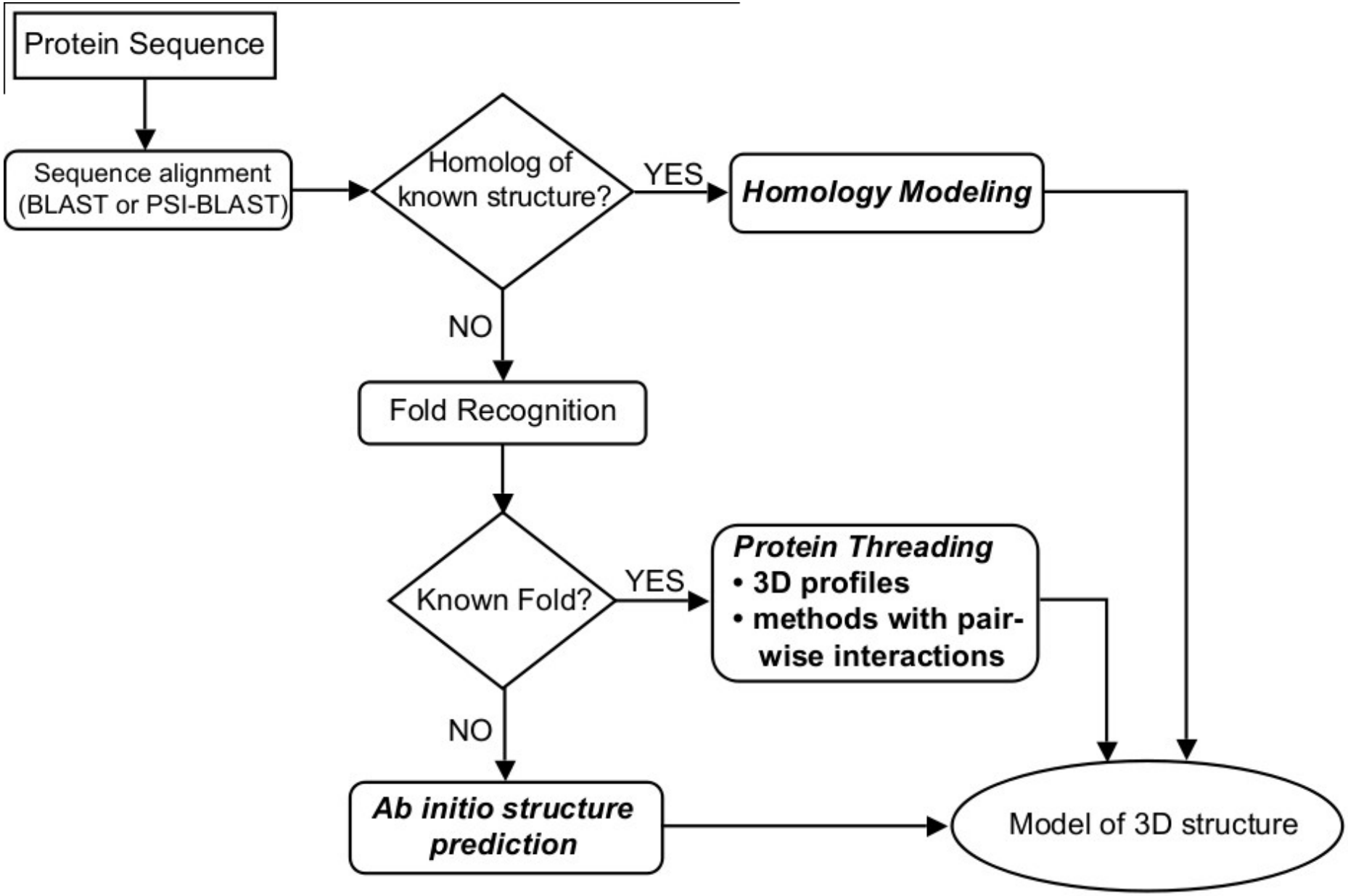
3) Les interaction des acides aminées avec l'Eau

Résumé de l'introduction

- Les protéines sont au cœur de tous les êtres vivants
- Critique pour toutes les études biologiques
- Le processus du repliement est largement inconnu
- Détermination de la structure protéique est expérimentalement difficile
- La fonction des protéines reste inconnue....

Prédiction des structures 3D en se basant sur des données connues.

Prédiction des structures 3D des protéines
Diagramme général



Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Prédiction des structures 3D des protéines

Techniques de prédiction de structure 3D:

A	Template modelling	En se basant sur un modèle connu dans les base de données
A1	Homology Modeling	Modélisation par Homologie des séquences
A2	Threading	En se basant sur la reconnaissance des coudes (Fold)

B	Template-Free Modeling	Modélisation sans un modéle connu dans les base de données
	Ab initio Methods •Physic-Based	Méthode ab initio •Basé sur les proriétés des AA •Thermodynamique •Hydrophobicité •electrostatique...

Modélisation par Homologie

Définition : Prédiction la structure 3D d'une protéine inconnue en utilisant comme modèle, une structure 3D connue d'une autre protéine homologue.

La méthode se base sur le fait que la structure de la protéine est plus conservé que sa séquence. Càd si les séquences sont similaires alors leurs structures le sont aussi.

Protein **D**ata **B**ank (http://www.rcsb.org/pdb/home/home.do) est la base de données qui regroupe toutes les structures des protéines connues (X-ray, crystallography, NMR).

La modélisation par homologie est basée sur des données réelles :

- Les résultats ont typiquement une bonne résolution
- La qualité des résultats s'amériore avec la quantité des protéines résolues dans PBD

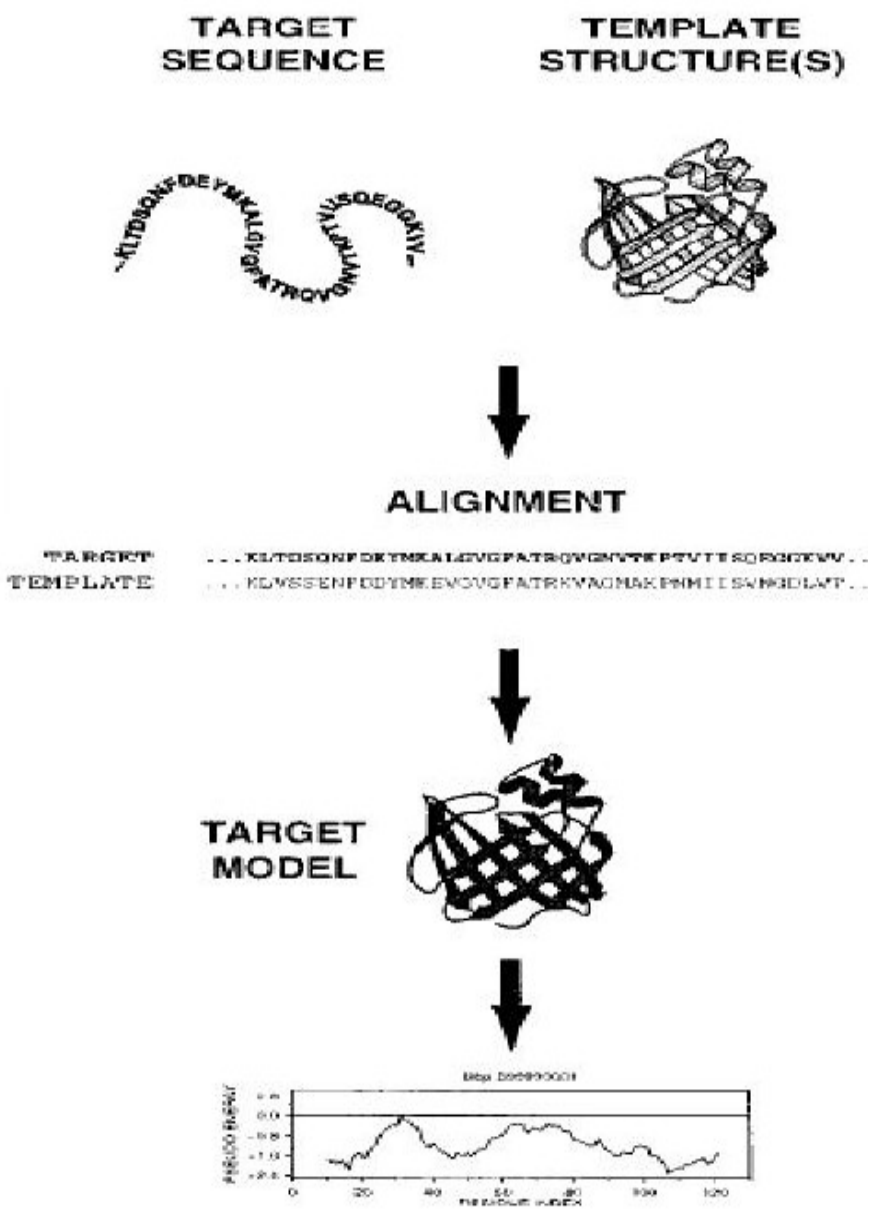
Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

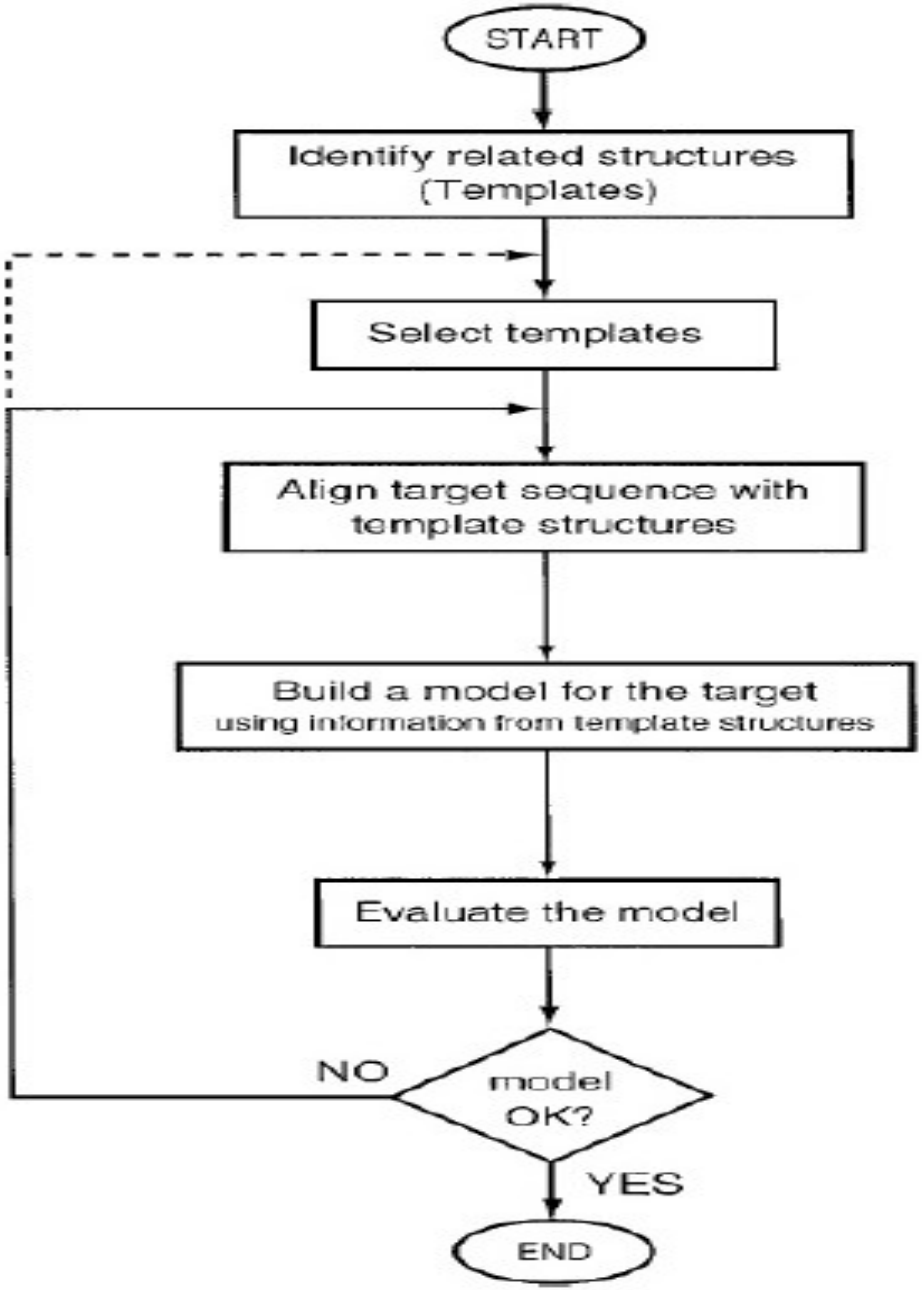
Les étapes pour la modélisation des structures en se basant sur l'homologie

- Utiliser la séquence inconnue en entrée pour la recherche de structures connues de protéines
- Produire le meilleur alignement global possible de la séquence inconnue et d'une (des) séquences de références
- Positionner les carbones Cα, en prenant les coordonnées de ceux de la protéine de référence comme modèle.
- Dans les régions comportant des brèches, soit dans la cible, soit dans la référence, utiliser une procédure pour modéliser les boucles, pour substituer les segments de longueur appropriée.
- Ajouter les chaînes latérales au modèle du squelette
- Optimiser la position des chaînes latérales (stoechiométrie)
- Optimiser la structure avec une minimisation d'énergie

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Prédiction basée sur l'homologie des séquences





Étape 1A : Recherche de modèle homologue

1) Recherche de similarité de séquence (BLAST) dans les bases de données de structure de protéines.

- Protein Data Bank (PDB) : http://www.pdb.org/
- Structural Classification Of Protein (SCOP) : http://scop.mrc-lmb.cam.ac.uk/scop/
- Protein StructureClassification (CATH) : http://www.cathdb.info/

2) Raffiner l'alignement de séquence avec PSI-BLAST pour détecter les faibles homologie de séquences.

3) S'il y a moins de 25% d'identité, la modélisation des reconnaissance des coude est favorisée.

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Étape 1B : Comment sélectionner le modèle

1) Sélectionner le modèle avec le meilleur pourcentage d'identité

2) Sélectionner plusieurs fragments de modèles

3) Tenir compte de la résolution de la structure du modèle

4) Tenir compte de la fonction de la protéine

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Étape 2A : alignement de le séquence avec le modèle

Alignement en paire (Smith-Waterman, CLUSTAL)

Étape 2B : Vérification du bon choix

Typiquement les méthodes d'alignement tend à maximiser la similarité des séquences (score), alors que la modélisation par homologie tend à maximiser la similarité des structures.
L'aginement peut être approuvé par d'autres informations comme la prédiction des structures secondaires.

§ Example: alignment of chymotrypsin (Cht) and trypsin (Trp) based on:

Sequence homology (Dayhoff, 1978)

```
Cht:   NTNCKK--YWGTKIKDAM
Trp:   NSSCKSA-YPG-QITSNM
```

Structural homology (Greer, 1981)

```
Cht:   NTNCKK--YWGTKIKDAM
Trp:   NSSCKS--AYPGQITSNM
```

Correct structural alignment even though it may not maximize sequence identity

Example from Greer (1981) *J. Mol. Biol.* 153:1027-1042

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Étape 3 : alignement de la séquence avec le modèle

Construire la structure de la protéine inconnu avec les données de la protéine modèle/

Exemples de méthodes :

- Modeling by rigid body assembly : COMPSER, SWISS-MODEL
- Modeling by segment matching or coordinate reconstruction : SEGMOD
- Modeling by satisfaction of spatial constraints : MODELLER

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Étape 4 : Évaluation et raffinage de la structure

Evaluer la stochiometry et d'autres propriétés structurales :

- Longueur des liaisons, les angles dihedral,les positions des chaînes latérales.

Exemples de méthodes :

- PROCHECK (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/)
- WHATCHECK (http://swift.cmbi.ru.nl/servers/html/index.html)
- PSVS (http://psvs-1_4-dev.nesg.org/)
- Eval123D (http://bioserv.cbs.cnrs.fr/HTML_BIO/frame_valid.html)
- JCSG Structure Validation Central (http://www.jcsg.org/prod/scripts/validation/sv2.cgi)

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

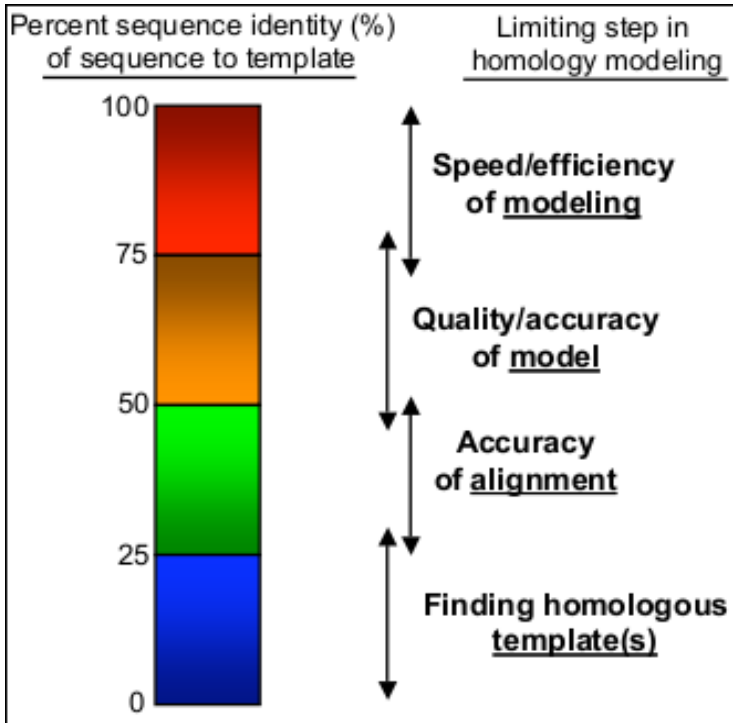
Limites de la modélisation par Homologie

1. Structure unique de protéine ne peut pas être prédite
2. Limitation avec performances des algorithmes d'alignement utilisées
3. Mauvaise gestion de positionnement des chaînes latérales
4. Difficulté de modéliser les régions des noeuds
5. Si le choix du modèle n'est pas le meilleur

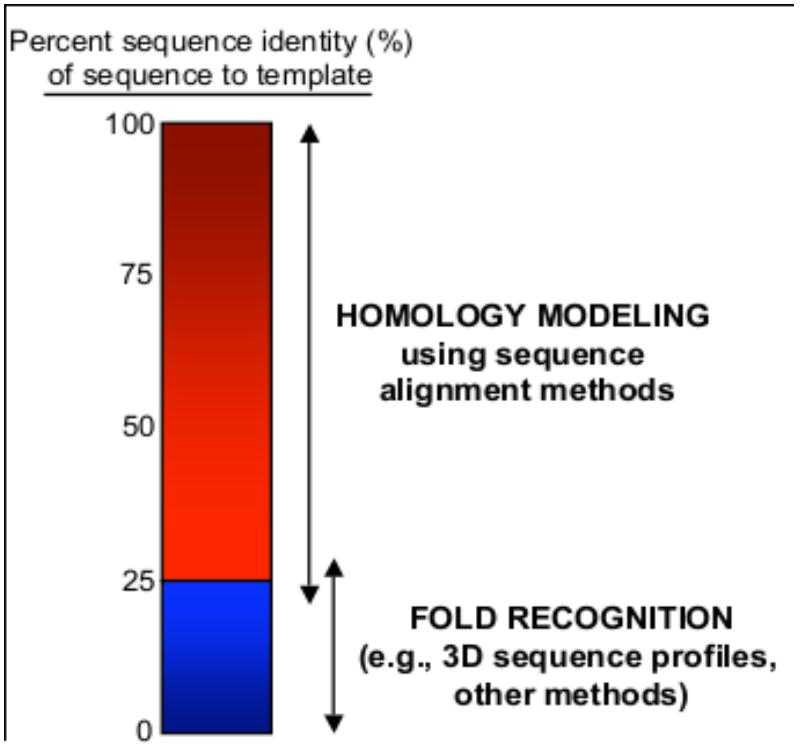
Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Reconnaissance des coudes (Fold Recognition)

Homology Modeling



Fold Recognition



Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Reconnaissance des coudes (Fold Recognition)

C'est une méthode qui identifie les structure 3D similaire mais qui n'ont pas de séquence homologue.

C'est une méthode qui se base sur le principe qu'il existe des protéine qui ont des structure qui se ressemblent avec un faible score d'alignement.

Il s'agit d'un alignement d'une séquence à un morceau de structure

La méthode derepliement (enfilage) intègre à la fois des informations de séquence et de structure (2D, 3D).
Exemple de programme : 3D-PSSM et PHYRE

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Reconnaissance des coudes (Fold Recognition) Environnement de la chaîne latérale

Elle se base sur la modélisation d'un profil 3D basé sur une matrice de score 1D-3D.

La matrice 1D-3D est similaire à la matrice BLOSUM mais elle tient compte en plus des propriétés environnantes des chaînes latérales des AA.

La structure environnante peut contenir des informations sur la structure secondaire, l'exposition au solvent, polarité, charge électrostatique.

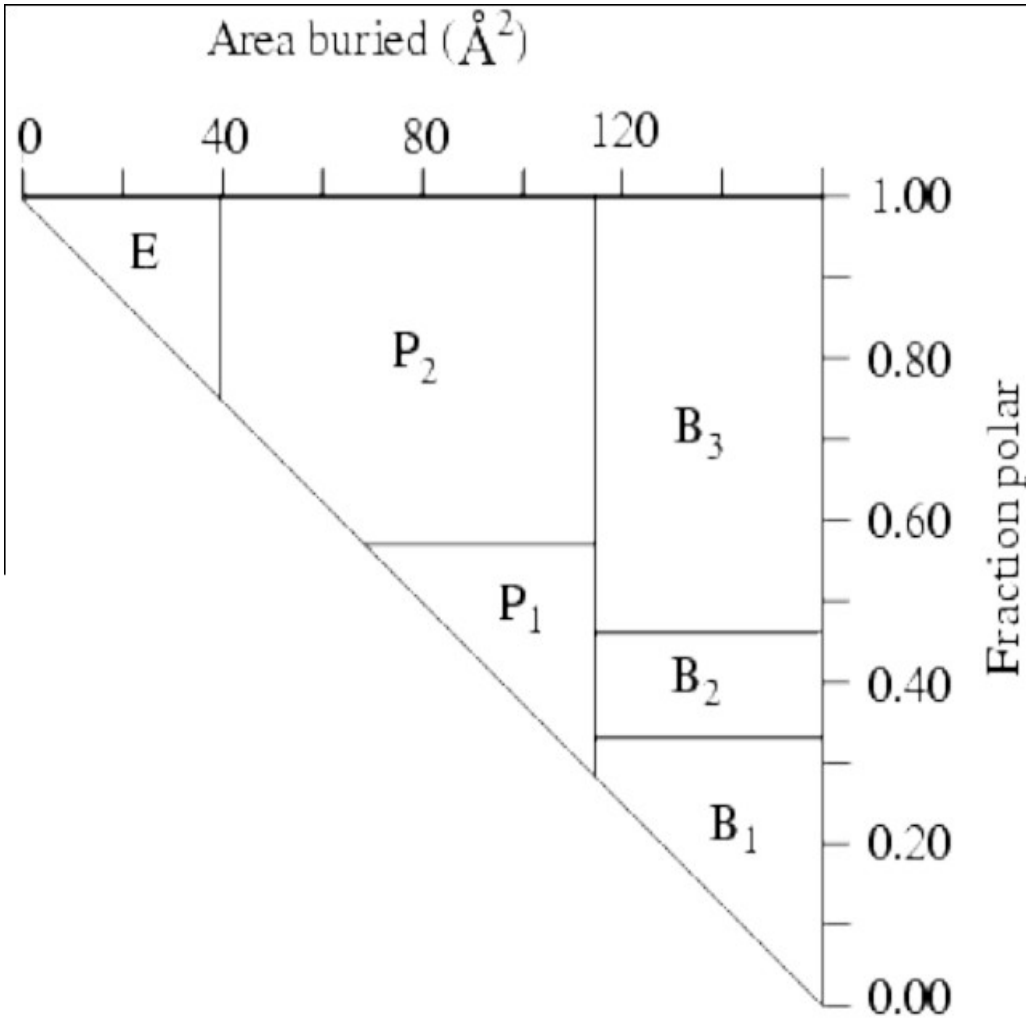
Bowie et al, 1991 Science 253:164-169 a défini 6 classes différentes d'environnement de la chaîne latérale qui dépendent de 3 états: Enfoncement – Hydrophobicité – Polarité.

Sachant que qu'il existe 3 structure secondaire différentes, la méthode de modélisation de profile structurale défini 3 x 6 = 18 environnement différents.

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Reconnaissance des coudes (Fold Recognition)
Environnement de la chaîne latérale

Environment:
B1 buried; hydrophobic environment
B2 buried; moderately polar environment
B3 buried; polar environment
P1 partially buried; moderately polar environment
P2 partially buried; polar environment
E exposed to solvent



§ Each environment is divided into three subclasses depending upon the secondary structure of the amino acid (alpha-helix, beta-strand, or other)

Bowie et al., (1991) Science 253:164-169

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

1D-3D Scoring Matrix

Bowie et al., (1991) Science 253:164-169

Environment class	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R
$B_1\ \alpha$	1.00	1.32	0.18	1.27	1.17	0.66	1.26	-0.66	-2.53	-1.16	-0.73	-1.29	-2.73	-1.08	-1.93	-1.74	-1.97	-0.34	-1.82	-1.67
$B_1\ \beta$	1.17	0.85	0.07	1.13	1.47	1.09	0.55	-0.79	-2.02	-0.94	-0.22	-1.12	-2.91	-1.67	-1.42	-1.93	-2.56	-1.91	-2.69	-1.16
B_1	1.05	1.45	0.17	1.10	1.11	1.02	0.98	-0.91	-1.92	0.26	-1.22	-1.53	-2.81	-1.17	-2.42	-2.52	-1.76	-1.12	-2.59	-2.16
$B_2\ \alpha$	0.50	0.90	0.85	1.01	0.63	0.68	1.12	-0.69	-1.49	-2.21	-0.10	-1.50	-1.47	-0.23	-0.61	-0.71	-1.62	0.23	-0.78	0.06
$B_2\ \beta$	0.01	1.18	1.06	0.76	1.31	1.06	0.64	-1.55	-2.26	-0.49	-0.87	-2.27	-1.77	-1.22	-2.07	-1.07	-1.41	-0.77	-1.14	-0.20
B_2	1.02	1.05	1.12	0.84	0.81	0.60	0.90	-0.68	-1.66	0.19	-0.05	-0.76	-1.17	-0.78	-0.65	-1.35	-1.28	0.46	-2.34	-0.80
$B_3\ \alpha$	0.92	-0.03	0.58	0.15	0.04	-0.02	0.69	-0.57	-1.86	-0.68	-1.56	-0.57	-0.96	0.22	-0.06	0.08	-0.50	0.73	0.43	0.96
$B_3\ \beta$	0.75	0.81	1.30	0.18	0.54	0.56	-0.57	-0.93	-1.93	-0.34	-0.54	-0.44	-0.74	0.21	-0.24	-0.14	-0.86	0.82	-0.53	0.13
B_3	1.07	0.70	1.13	0.35	-0.17	-0.03	0.23	-0.96	-0.98	-0.13	-1.20	-0.53	-0.54	0.05	0.04	-0.36	-1.05	1.01	0.10	0.66
$P_1\ \alpha$	-1.35	-0.82	-0.59	-0.52	-0.24	0.10	-0.03	0.73	-0.49	-0.25	0.95	0.31	0.34	-0.14	-0.54	-0.17	-0.25	-0.52	-0.21	-0.28
$P_1\ \beta$	0.36	-0.49	0.17	-1.03	0.20	0.46	-0.27	0.64	-0.82	-0.55	1.49	0.93	0.33	-2.27	-1.32	-0.73	-1.07	-0.42	-1.21	-0.77
P_1	-1.28	-1.20	-1.31	-0.82	-0.23	-0.01	-1.19	0.46	-0.24	0.66	1.35	0.56	0.49	-0.63	-0.13	-0.61	0.38	-1.12	-0.74	-1.29
$P_2\ \alpha$	-1.14	-1.43	-0.79	-0.35	-0.54	-0.48	-0.45	0.06	-0.50	-0.26	-0.93	-0.05	-0.18	0.55	-0.05	0.56	0.28	0.06	0.61	0.50
$P_2\ \beta$	-0.79	-0.54	-0.84	-1.30	-0.33	0.13	-0.72	-0.55	-0.98	-1.29	-0.57	0.84	0.59	-0.08	-0.16	0.32	0.19	-0.87	0.59	0.10
P_2	-0.82	-0.86	-0.51	-0.70	-1.09	-0.88	-0.89	-0.15	-0.40	0.44	-0.60	0.06	0.26	0.27	0.50	0.27	0.49	0.13	0.44	0.30
$E\ \alpha$	-1.35	-2.20	-2.10	-1.58	-2.76	-1.10	-0.72	0.46	0.68	0.04	-0.44	-0.17	0.15	0.36	0.28	0.59	0.44	-0.19	0.13	-0.34
$E\ \beta$	0.84	-0.90	0.30	-1.66	-1.47	-1.74	-0.88	0.06	1.46	-0.96	-0.24	0.14	0.65	-0.19	-0.06	-0.16	-0.78	-0.83	-0.52	-0.49
E	-2.14	-1.90	-0.94	-1.19	-1.61	-0.91	-1.67	0.12	1.13	0.20	-0.46	0.12	0.32	-0.03	0.41	0.03	0.22	-0.25	-0.14	-0.32

Amino acid preferences for solvent accessibility

Amino acid preferences for particular secondary structure

Interactions between neighboring amino acids (‘contact’ or ‘pair’ potentials)

Construction du profile 3D à partir de la structure 3D

Amino acid Position: N = 1 2 3 4 ...
Structural Environment = E B_2 Eα $P_2\alpha$ ...

§ Use 3D-1D scoring matrix columns to construct 3D profile for each position (amino acid) in structure:

3D profile

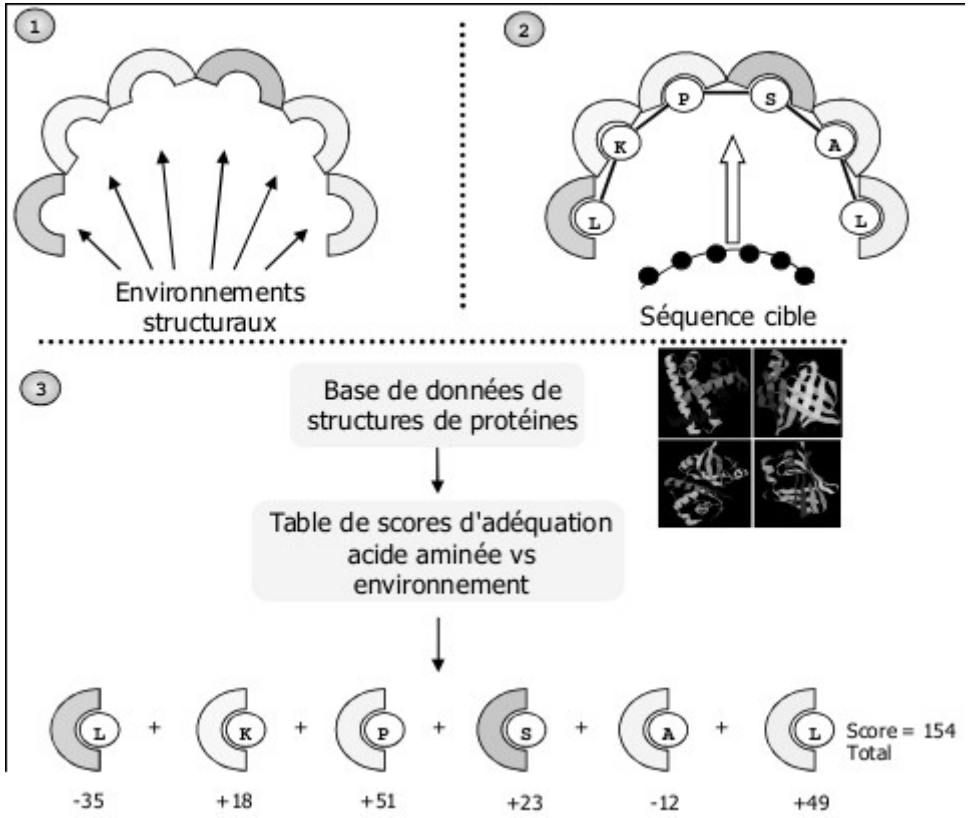
		Amino acid								
Position	3D Environment	A	C	D	E	.....	T	V	W	Y
1	E	12	-46	22	3	.....	12	-91	-214	-94
2	B_2	-66	-5	-128	-135	.....	-76	60	102	112
3	E α	46	-44	44	59	.....	-17	-110	-135	-210
4	$P_2\alpha$	6	-93	28	56	.....	-5	-48	-114	-79
.	.	.	.	.	.		.	.	.	.
.	.	.	.	.	.		.	.	.	.

(Note: 3D-1D scores were multiplied by 100 in profile)

Bowie et al., (1991) Science 253:164-169

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

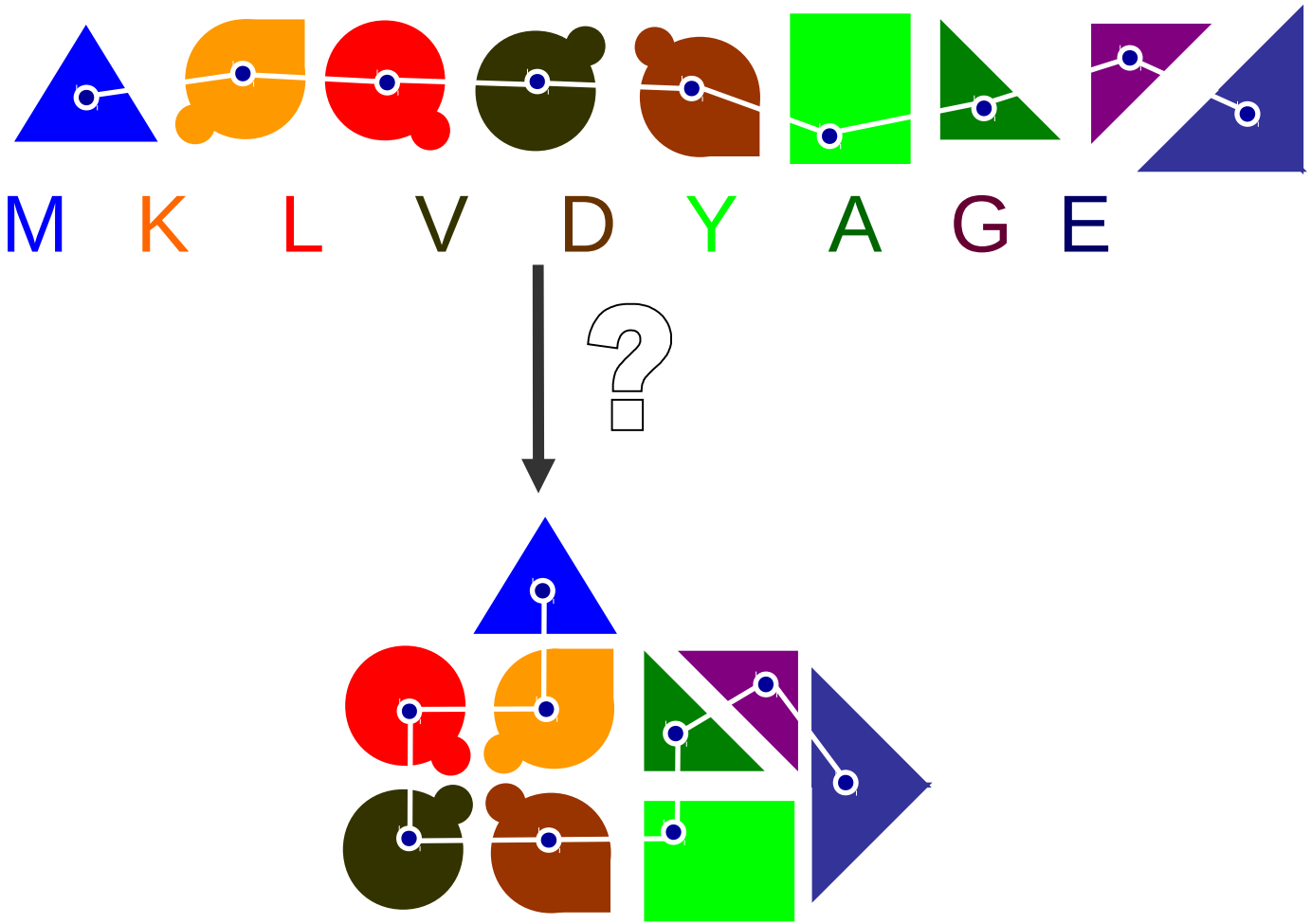
Threading modelling
(modélisation par enfilage ou repliement)



Score = Énergie libre + Force électrostatique

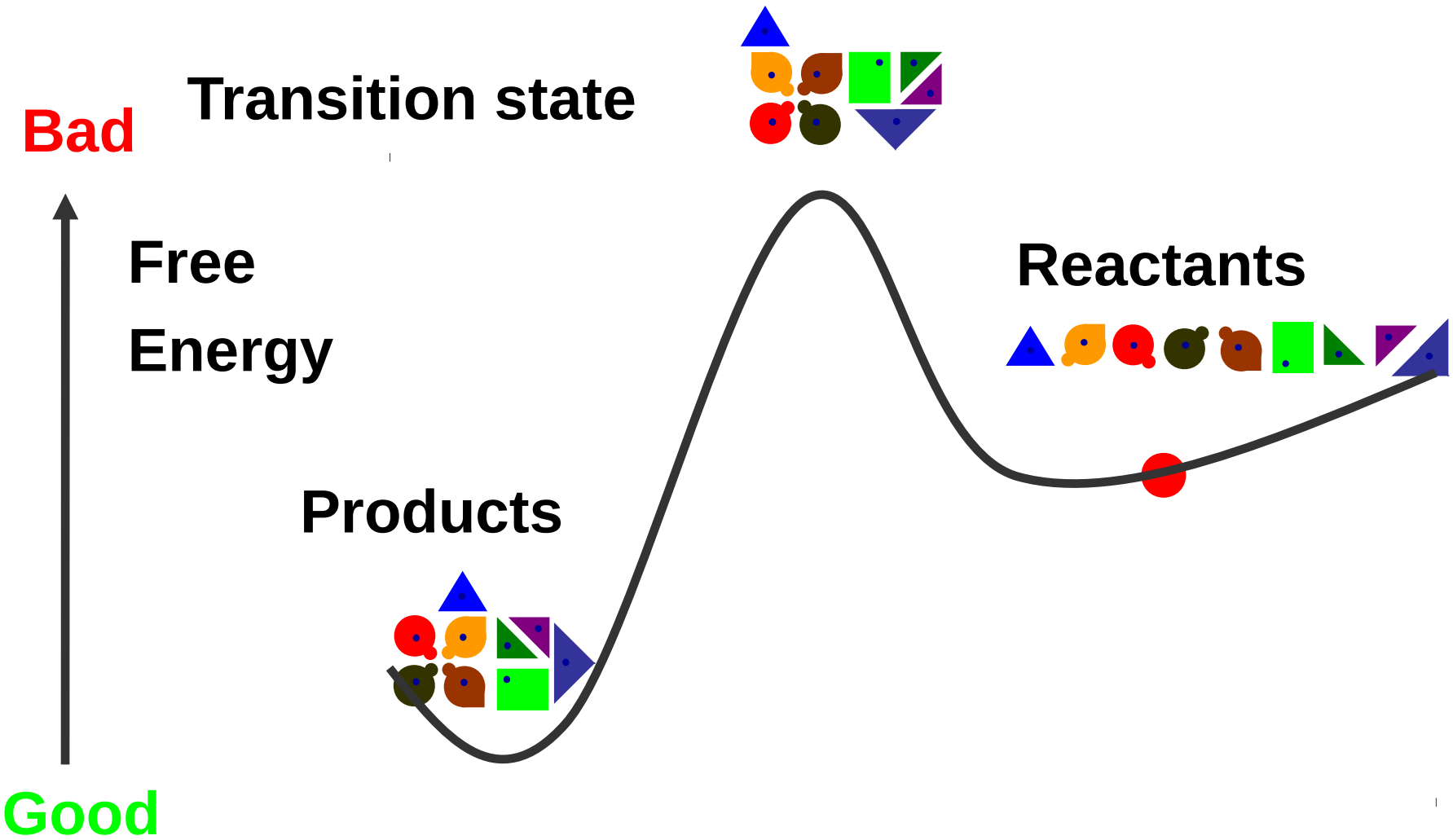
Optimisation de positionnement des chaines latérales on focntions de son niveau d'énergie

- **Proteins are hetero-polymers that adopt a unique fold**



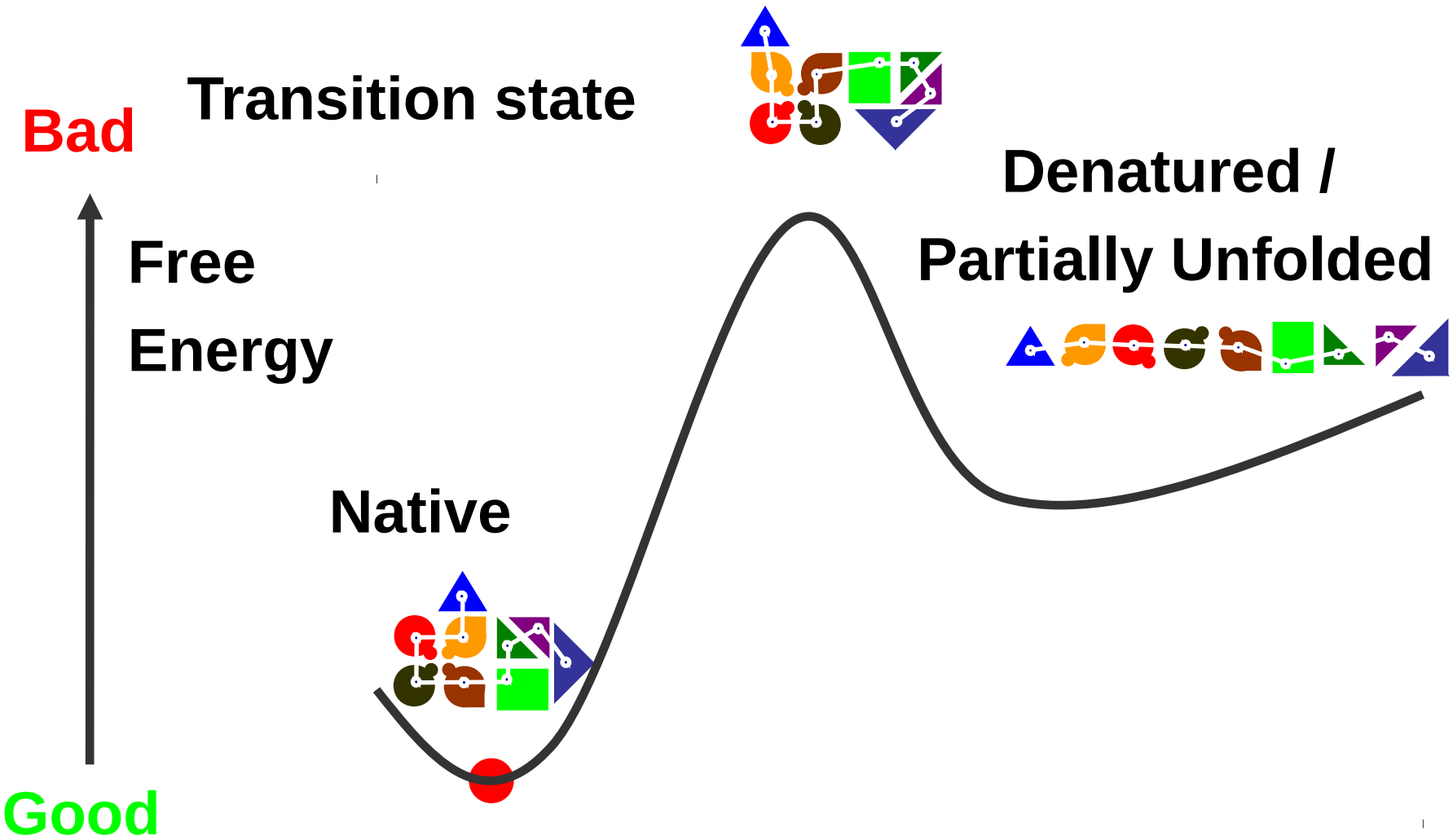
Proteins

- Protein folding as a reaction



Proteins

- Protein folding …



Usage des profils 3D

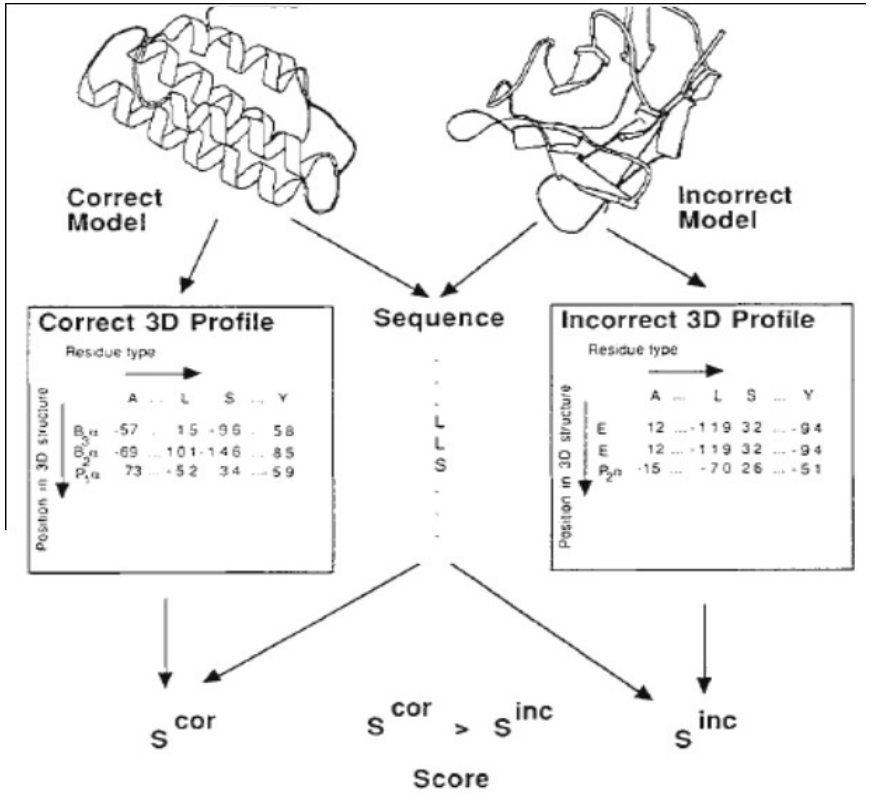
Les profils 3D sont utilisés dans la ercherche de similarité de structures/coude pour des protéines inconnues

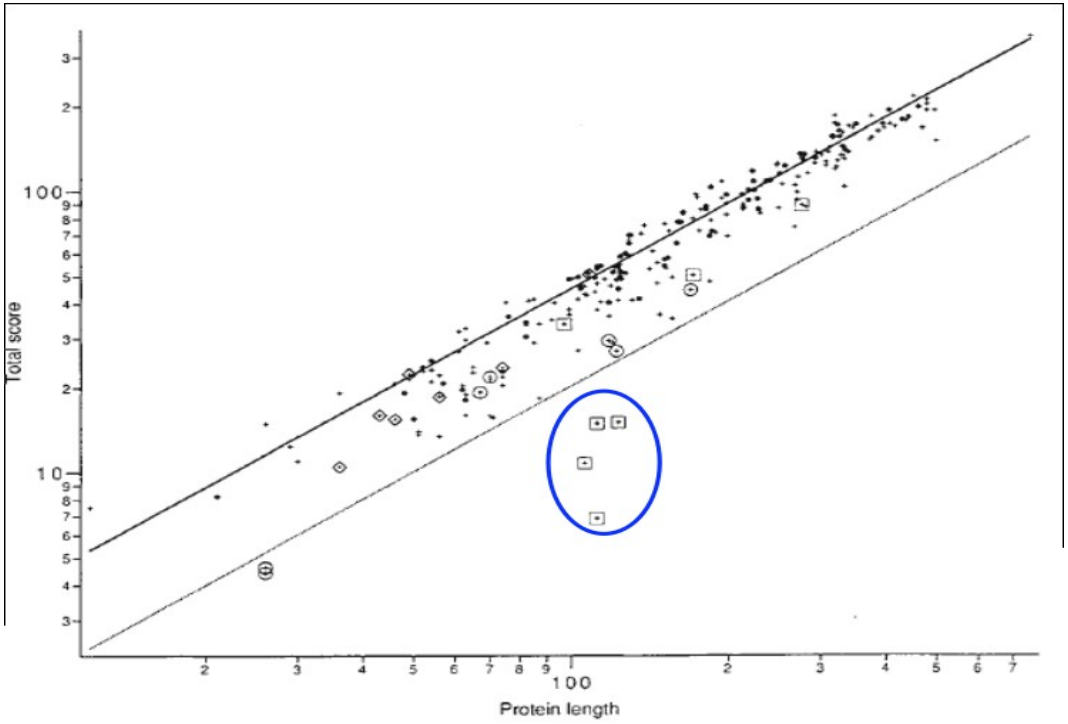
Les profil 3D peuvent être utilisés dans l'évaluation des modèles prédits

Les profils 3D peuvent être utilisé pour prédire des structures par l'alignement de de séquences avec une librairie de profils 3D (Threading structure prediction).

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Usage des profils 3D : Comparaison de modèles en utilisant les profils 3D





Images from Luthy et al., (1992) Nature 356:83-85

Voir cours de Dr. Justin Fisher : http://www.bioinformatics.wsu.edu/bioinfo_course/notes.htm

Threading Methodology

1. Construction d'une librairie de coudes modèle

2. Calcul de score pour évaluer l'emplacement de la séquences dans les coudes modèles

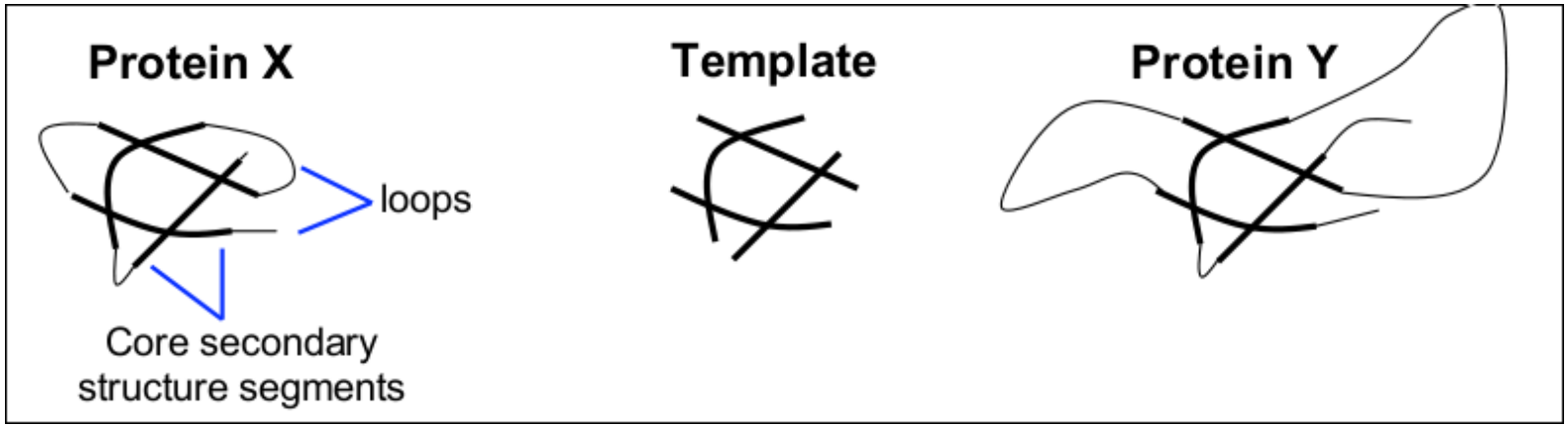
3. Recherche de l'alignement optimal entre la séquence et chaque coude modèle

4. Sélectionner les meilleurs coudes modèles avec la séquence de protéine

5. Le modèle 3D est la dérivé de l'alignement optimal (threading) de la séquence au meilleur score de la structure modèle

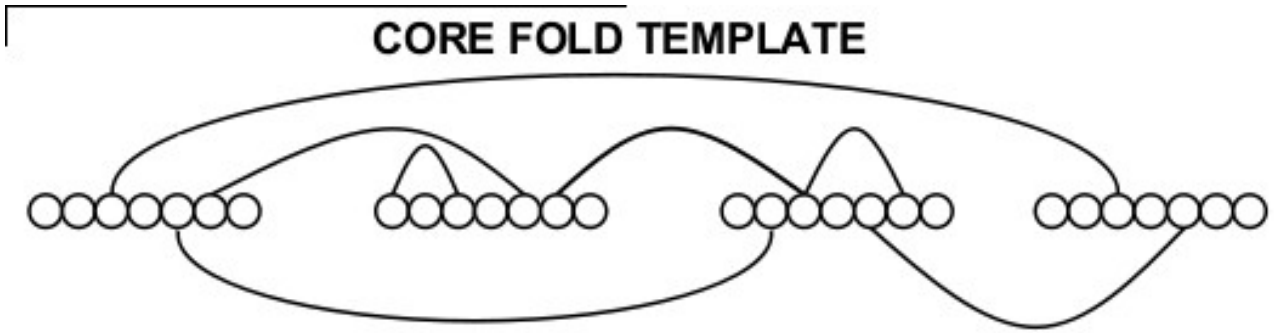
Threading :
étape 1- construire une librairie de modèle

Construire une librairie de Coeur (core)



Un modèle de coeur peut mimer les coudes en commun d'une famille une structure 3D d'une protéine.

Un modèle de coeur peut inclure des informations des interaction intra-proteine



Threading :
étape 2 - Calcul de Score

Amino acid Position: N = 1 2 3 4 ...
Structural Environment = E B_2 Eα $P_2\alpha$ ...

§ Use 3D-1D scoring matrix columns to construct 3D profile for each position (amino acid) in structure:

3D profile

		Amino acid								
Position	3D Environment	A	C	D	E	.....	T	V	W	Y
1	E	12	-46	22	3	.....	12	-91	-214	-94
2	B_2	-66	-5	-128	-135	.....	-76	60	102	112
3	E α	46	-44	44	59	.....	-17	-110	-135	-210
4	$P_2\alpha$	6	-93	28	56	.....	-5	-48	-114	-79
.	.	.	.	.	.		.	.	.	.
.	.	.	.	.	.		.	.	.	.

(Note: 3D-1D scores were multiplied by 100 in profile)

Amino acid preferences for solvent accessibility

Amino acid preferences for particular secondary structure

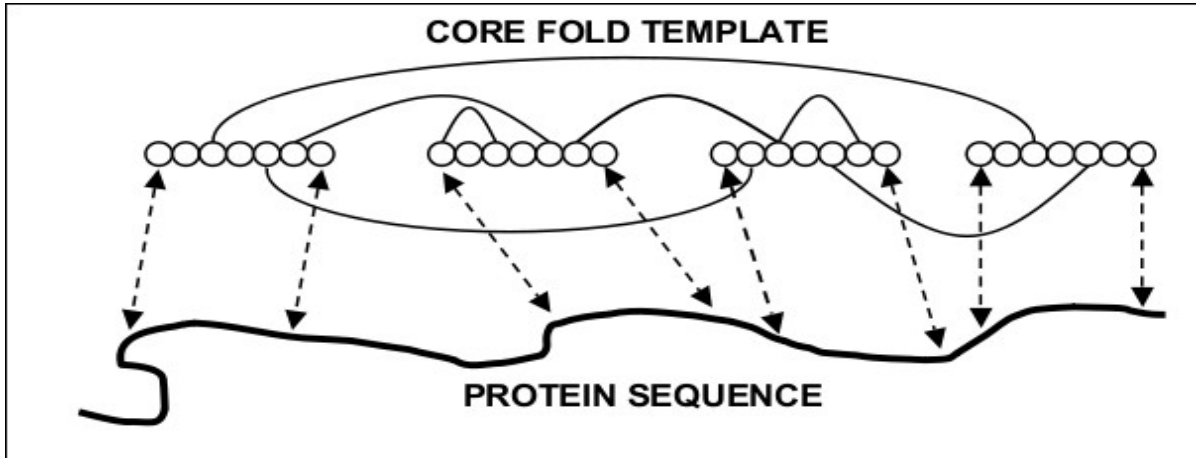
Interactions between neighboring amino acids (‘contact’ or ‘pair’ potentials)

Threading :
étape 3 - Alignement de la séquence avec le meilleur modèle de coeur

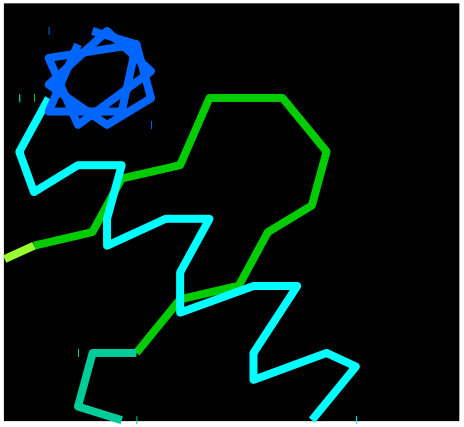
1. Si les interactions intra-protéine ne sont pas autorisées
   - La méthode de la programmation dynamique trouve l'alignement optimal entre la séquence et le modèle

2. Si les interactions intra-protéine sont autorisé :
   - Les méthodes Heuristiques peuvent ne pas trouver le meilleur alignement

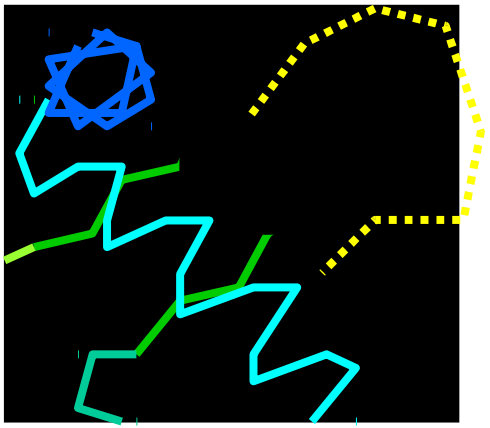
   - Les méthodes exact (Branch & bound) peuvent trouver un alignement optimal mais nécessite une durée exponentielle.



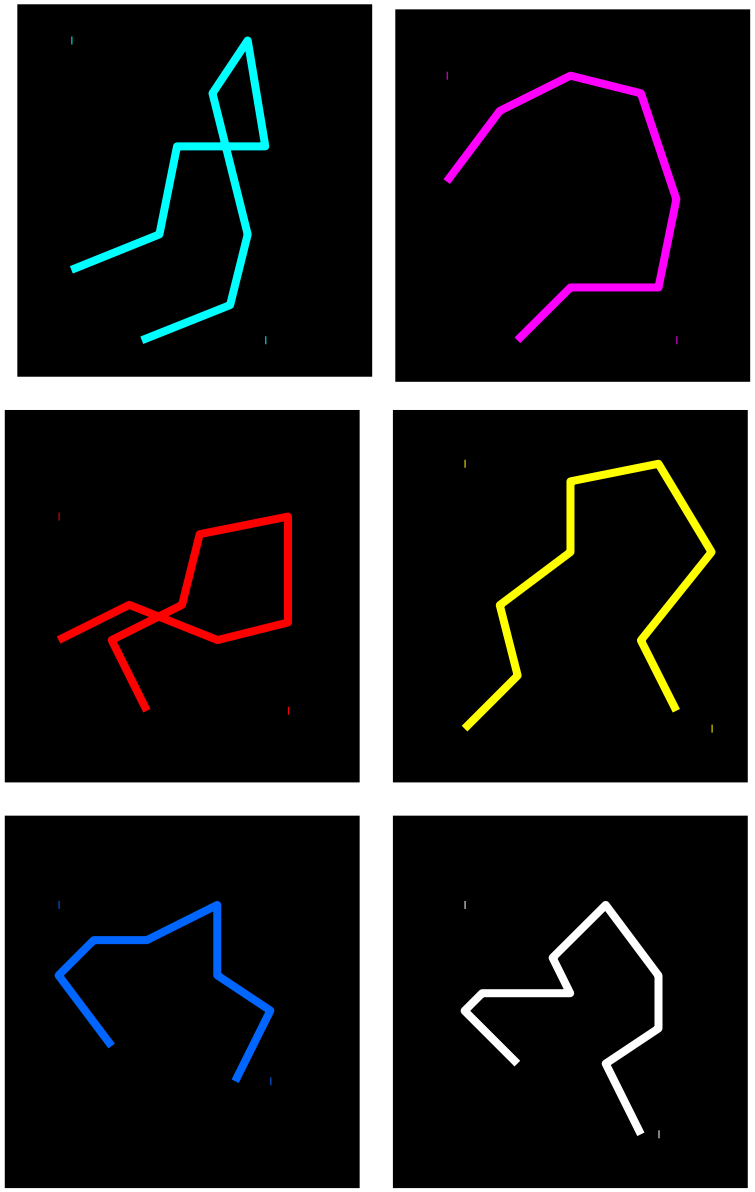
Threading :
Étape 4 – Sélection du modèle



Query FGHQWERT



Hit #1 YAYE--KS



Profil 3D et Threading

Les méthodes de profils 3D et threading alignent des séquences à des structures

La méthode des profils 3D peut être utilisée

- Reconnaissance des coudes
- Evaluation des modèles 3D
- Prédiction des structures de protéine

La méthode Threading reconnaît les coude pour prédire les structures des protéines.

La méthode threading a généralement des rendement moyen de résolution de structures modèles

Exemple de programmes qui utilisent Threading :

- RAPTOR (http://www.bioinfor.com/all-products/raptor)
- MUSTER (http://zhanglab.ccmb.med.umich.edu/MUSTER/)
- THREADER (http://bioinf.cs.ucl.ac.uk/software_downloads/threader/)

Prédiction des structures 3D des protéines

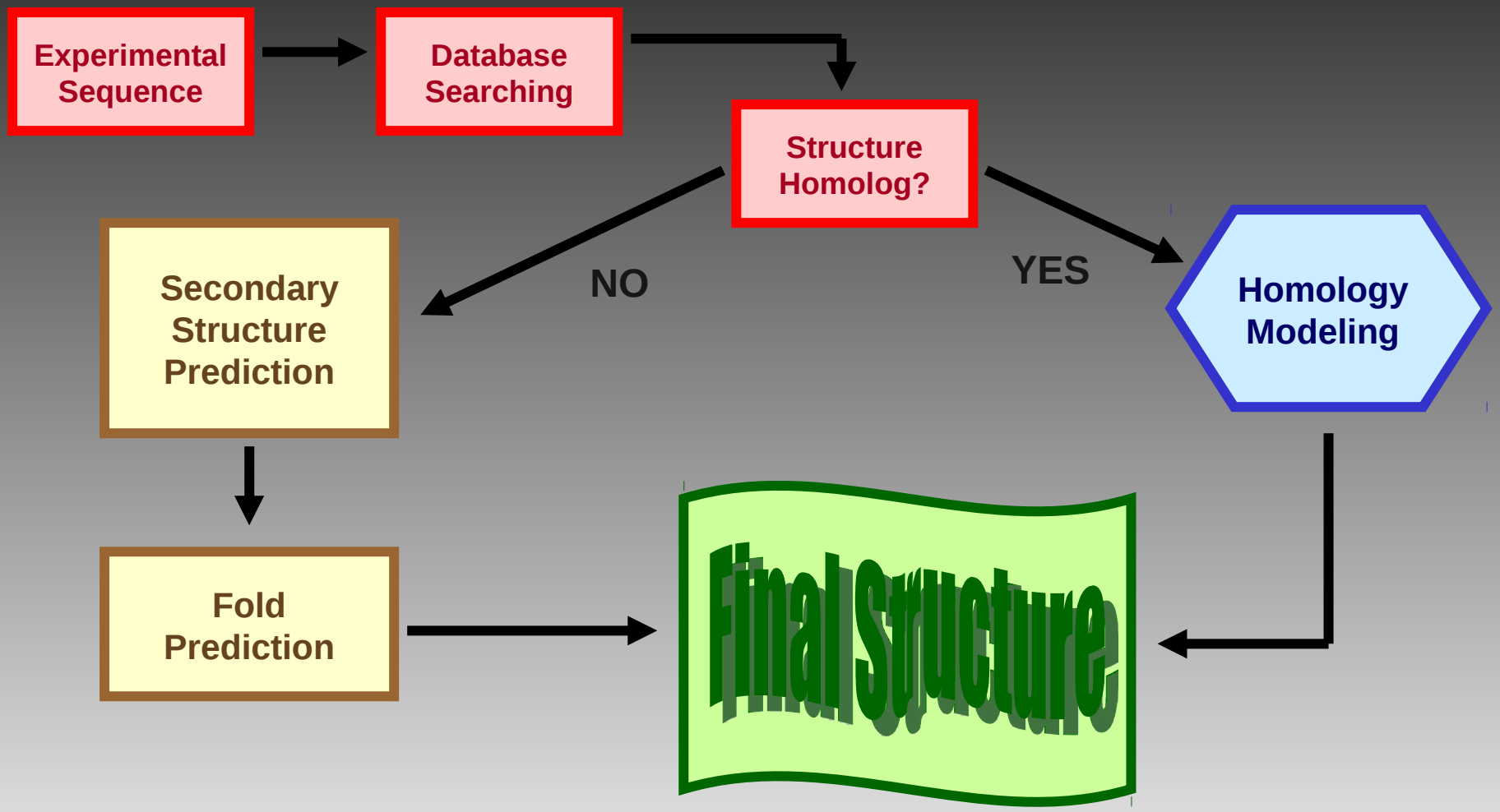
Comparaison des modélisations par Homologie et reconnaissance des coudes

modélisation par homologie	modelisation par reconnaissance des coudes (Fold)
recherche d'identité	Essayer toutes les possibilités
Determine l'alignement optimal (pairwise ali.)	essayer plusieurs alignements (MSA)
optimiser un model	evaluer plusieurs modèles

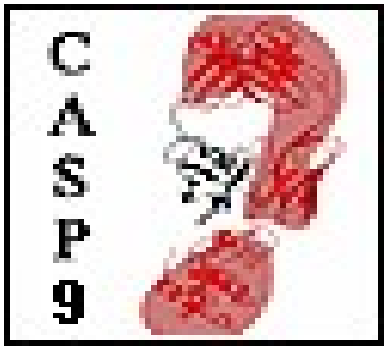
Résumé sur l'état des prédictions des structures 3D

Method	Sequence Similarity	Template Coverage	Accuracy	Difficulty	Comput. Expense
Homology	>30%	>90%	1-3 Å	Trivial	Seconds
FR/Homology	20-30%	>75%	2-5 Å	Easy	Minutes
Fold Recognition	<20%	>50%	3-10 Å	Moderate	Hours
Ab initio	<10%	0	5-20 Å	Hard	Days

Protein structure prediction flowchart

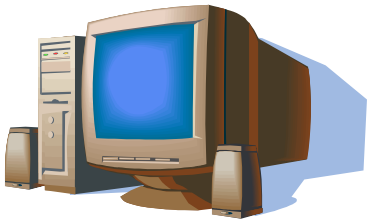


Protein Structure Prediction Center



- CASP: Critical Assessment of Structure Prediction

- CAFASP: Critical Assessment of Fully Automated Structure Prediction



Haut débit

CAFASP

Le but de CAFASP est d'évaluer la performance des serveurs de prédiction structurale disponible sur le réseau. Contrairement au CASP permet d'évaluer les serveurs sans intervention humaine

Classification des serveurs de prédiction 3D

CAFASP5 results using MaxSubDom (based on 89 targets out of 100)

#	Server	Total	%Diff	Time(hh:mm)	283	288	289	290	291	292	293	294	295	296	297	298	299	300
1	Zhang-Server	49.621	0	35:02	22	81	37	98	92	67	45	75	74	5	65	74	0	34
	*MQAPCons	47.361	4	-	33	74	43	91	90	59	41	77	74	5	64	70	15	33
	*Verify	47.008	5	-	33	83	41	97	91	61	45	64	73	9	64	46	15	37
2	HHpred2	45.343	8	0:15	0	73	37	96	72	62	45	74	79	0	61	76	12	0
	*Taruna	45.329	8	-	0	79	36	97	88	63	44	76	75	0	58	70	15	0
3	Pmodeller6	45.086	9	-	0	75	32	91	77	61	43	75	75	0	58	66	15	37
4	HHpred3	44.860	9	0:21	22	73	36	96	72	62	44	74	79	0	63	76	0	0
5	Pcons6	44.662	9	31:03	0	77	37	96	91	62	41	71	74	6	65	65	0	34
6	CIRCLE	44.589	10	37:00	20	78	42	97	90	63	44	79	75	6	59	70	0	30
7	ROBETTA	44.262	10	23:05	33	80	41	96	91	61	45	64	73	9	64	46	15	34
8	BayesHH	44.191	10	0:35	23	74	47	96	61	62	44	75	77	0	63	72	0	0
9	HHpred1	44.050	11	0:11	0	79	38	97	91	61	46	72	73	0	62	69	0	0
10	MetaTasser	43.258	12	46:02	0	78	34	88	54	51	38	53	75	0	68	44	0	0
11	UNI-EID_expm	43.134	13	10:02	0	80	36	96	67	63	44	60	76	0	68	69	0	0
12	FAMSD	42.962	13	20:02	0	78	42	96	91	63	47	74	82	6	61	70	0	0
13	FAMS	42.760	13	43:12	0	76	33	90	90	66	46	79	79	6	54	60	0	0
14	beautshot	42.723	13	2:02	0	78	37	97	88	64	44	77	75	0	62	64	0	0
15	RAPTOR-ACE	42.554	14	45:27	0	79	43	88	90	60	42	74	73	0	65	70	0	35
16	UNI-EID_bnmx	42.337	14	9:54	0	82	37	96	91	49	43	60	75	0	66	69	0	0
17	SP3	41.899	15	10:37	0	80	33	91	66	63	41	79	75	5	65	70	0	35

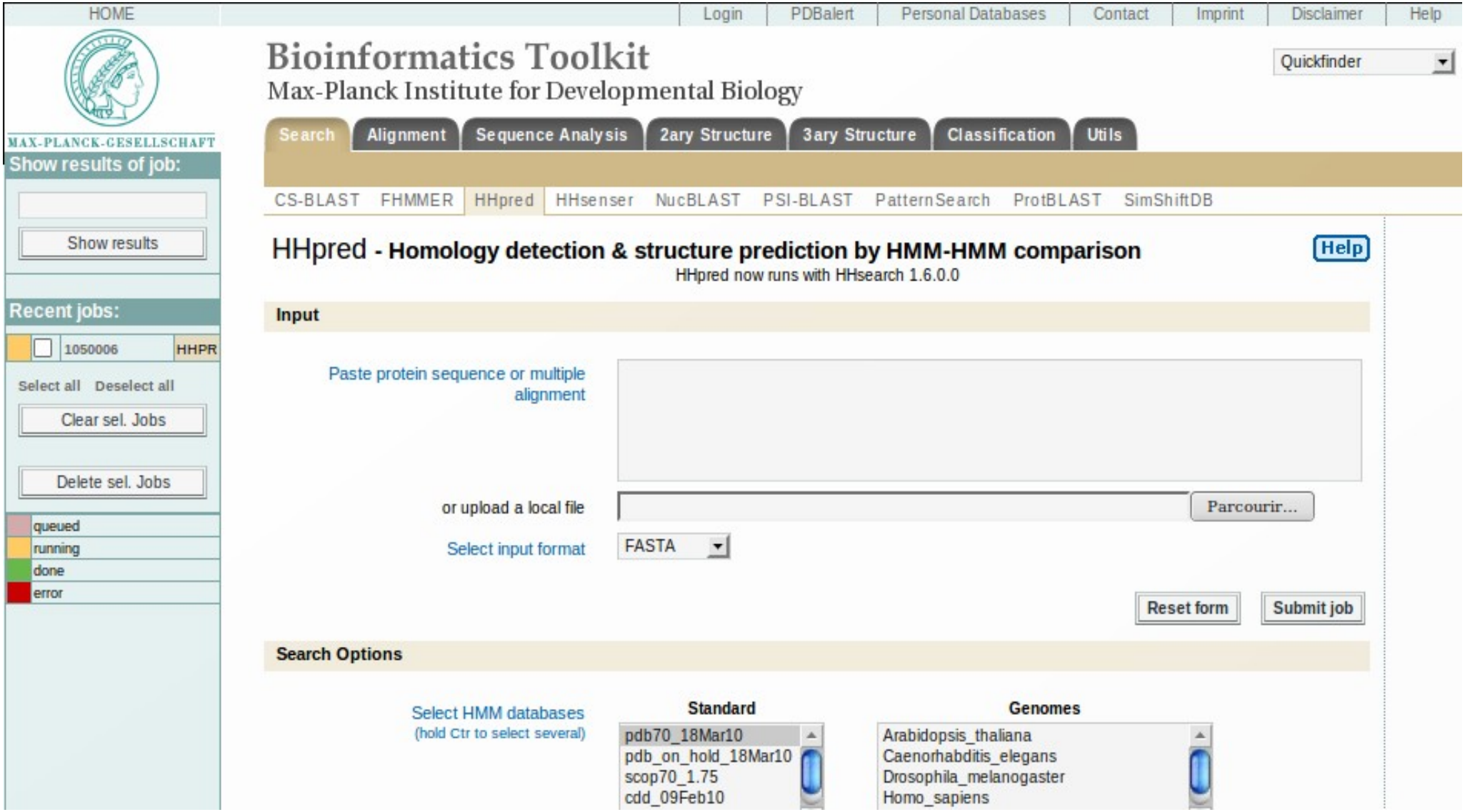
Différents scores utilisés par CASP

- Predictors---Name of groups.
- N----------Number of targets used to calculate the cumulative score.
- Rank---------Rank of the predictors on the basis of TM-score or GDT-score.
- TM_1---------TM-score of the first models. A TM-score < 0.17 implies that the prediction is close to random structures. TM-score=0 means either that the target was not submitted or that there is no overlap between the predicted model and the solved structure.
- Zscore-------Z-score derived from corresponding scores for each given target.
- MS_1---------MaxSub-score of the first model. (The MaxSub-score is calculated based on the TM-score searach engine).
- GDT_1--------GDT-score of the first model. (The GDT-score is taken from official CASP7 assessors).
- RM_1---------RMSD to native of the first model.
- cov----------The percentage of residues in the predicted model relative to native.
- DGyr---------Difference of radius of gyration between model and the native structure, i.e. DGyr=|Gyr_model-Gyr_native|, where Gyr_model and Gyr_native are the radius of gyration of the predicted model and the native structure, respectively. When the radius of gyration is calculated, the same set of residues in model and the native struture are used.
- NC----------Number of clashed models. A clashed model is defined according to the Valencia's rule (Proteins Suppl 7: 27-45, 2005), i.e. the model with more than 4 severe clashes (Ca-Ca pair-wise distance < 1.9 Angstroms) or with more than 50 bumps (Ca-Ca pair-wise distance < 3.6 Angstroms).
- TM_B---------TM-score of the best in top-five models.
- MS_B---------MaxSub-score of the best in top-five models
- GDT_B--------GDT-score of the best in top-five models.
- RM_B---------RMSD to native of the model with the best TM-score in top-five models.
- L_seq--------Length of the target sequences for modeling.
- L_native-----Number of the residues in the solved structures.

http://predictioncenter.org/casp9/CD/data/html/groups.2.html

#	GR #	GR name	Domains Count	SUM Z-score (GDT_TS)	AVG Z-score (GDT_TS)	AVG GDT_TS	SUM Z-score (AL0P)	AVG AL0P	AVG GDT_HA	AVG CA(i)-CA(i+1)	AVG Mammoth (Z-Score)	AVG Dali (Z-Score)
1.	380	QUARK	147	115.883	0.788	62.675	0.731	60.259	45.669	3.780	16.998	14.843
2.	428	Zhang-Server	147	113.344	0.771	62.765	0.756	60.999	45.772	3.781	17.127	14.650
3.	077	RaptorX-MSA	147	103.356	0.703	61.774	0.668	58.445	44.942	3.765	17.018	15.090
4.	286	RaptorX	147	103.103	0.701	61.731	0.700	58.852	44.671	3.771	17.029	14.814
5.	276	RaptorX-Boost	147	99.938	0.680	61.453	0.672	58.309	44.223	3.778	17.047	14.729
6.	453	HHpredA	147	93.199	0.634	59.528	0.683	56.984	44.013	3.787	15.907	14.163
7.	449	HHpredB	147	93.199	0.634	59.528	0.683	56.984	44.013	3.787	15.907	14.317
8.	346	HHpredC	147	91.916	0.625	59.361	0.684	56.648	43.899	3.787	15.867	14.276
9.	452	Seok-server	147	89.636	0.610	60.158	0.572	55.974	43.936	3.760	16.069	14.363
10.	002	MULTICOM-CLUSTER	145	87.836	0.606	60.358	0.590	55.776	43.750	3.761	16.353	14.436
11.	321	BAKER-ROSETTASERVER	145	87.346	0.602	58.768	0.624	55.271	42.552	3.798	16.139	13.914
12.	119	MULTICOM-REFINE	147	86.512	0.589	59.519	0.534	54.806	43.366	3.759	16.141	14.379
13.	215	MULTICOM-NOVEL	147	82.903	0.564	59.371	0.510	54.037	43.135	3.761	15.959	14.380
14.	236	gws	145	82.737	0.571	58.931	0.533	54.414	43.010	3.781	15.999	14.595
15.	457	chunk-TASSER	147	82.710	0.563	58.846	0.598	55.982	41.465	3.798	16.387	13.466
16.	174	Phyre2	147	78.887	0.537	58.823	0.557	55.149	42.506	3.798	16.137	13.790
17.	080	MULTICOM-CONSTRUCT	147	76.526	0.521	58.703	0.492	53.194	42.654	3.763	15.999	14.236
18.	253	pro-sp3-TASSER	147	75.440	0.513	58.117	0.531	54.707	40.747	3.798	16.131	13.250

HHpred - Homology detection & structure prediction by HMM-HMM comparison



I-TASSER (Zhang-Lab)

Zhang Lab

THE UNIVERSITY OF MICHIGAN

Home | Publications | Research | People | Funding | Job Opening | Services | Labonly

Online Services

- I-TASSER
- LOMETS
- MUSTER
- TM-score
- TM-align
- MM-align
- SVMSEQ
- ANGLOR
- REMO
- Potential
- Decoys
- SPICKER
- HAAD

I-TASSER ONLINE

Protein Structure & Function Predictions

(The server has completed structure and function predictions for 36305 proteins submitted by 8337 users from 86 countries)

(Number of visitors: 180407)

I-TASSER is an Internet service for protein structure and function predictions. 3D models are built based on multiple-threading alignments by LOMETS and iterative TASSER simulations; function inslights are then derived by matching the predicted models with protein function databases. I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent CASP7 and CASP8 experiments. The server is in active development with the goal to provide accurate structural and function predictions using state-of-the-art algorithms. The server is only for non-commercial use. Please send your questions and comments to yangzhanglab@umich.edu. (More about the server ...)

[**Queue**] [**Search**] [**About**] [**Statistics**] [**Remove**] [**Potential**] [**Decoys**] [**What's new**]

Copy and paste your sequence here (<1,500 residues):

RCSB PDB
PROTEIN DATA BANK

Historique:
La banque PDB a commencé à répertorier les coordonnées des structures cristallographiques de macromolécules des les années 1770.
Depuis développement de logiciels qui permettent de visualiser, localiser les sites actifs, et calculer des propriétés spatiales comme l'accessibilité aux solvants ou le potentiels électrostatiques.

http://www.pdb.org/pdb/home/home.do

Format de fichier PDB

1ère Partie: appelé en-tête
Contient des informations bibliographiques attachées à la structure, sur la résolution et les paramètres cristallographiques, la séquence et parfois la structure secondaire.

2ème partie: Elle contient les coordonnées atomiques
Dans cette partie les atomes désignés par ATOM se situent sur la chaine protéique, tandis que les atomes désignés par HETATM (HETeroAToM group) font partie des molécules cofacteurs, substrats, ions ou autres groupes qui sont liés par une liaison covalente à la chaîne protéique

Compound/Molecule

Keywords

Method used (also NMR, simulation)

Date Deposited

PDB Identifier

```
HEADER    TRANSCRIPTION/DNA                          07-FEB-00   1EFA
TITLE     CRYSTAL STRUCTURE OF THE LAC REPRESSOR DIMER BOUND TO
TITLE    2 OPERATOR AND THE ANTI-INDUCER ONPF
COMPND     LAC REPRESSOR/DNA COMPLEX
KEYWDS     PROTEIN-DNA COMPLEX, HELIX-TURN-HELIX, GENE REGULATION,
EXPDTA     X-RAY DIFFRACTION
AUTHOR     C.E.BELL, M.LEWIS
JRNL         AUTH   C.E.BELL, M.LEWIS
JRNL         TITL   A CLOSER VIEW OF THE CONFORMATION OF THE LAC
JRNL         TITL 2 REPRESSOR BOUND TO OPERATOR.
JRNL         REF    NAT.STRUCT.BIOL.                V.   7   209 2000
REMARK   1 REFERENCE 1
REMARK   1  AUTH   M.LEWIS, G.CHANG, N.C.HORTON, M.A.KERCHER, H.C.PACE
REMARK   1  TITL   CRYSTAL STRUCTURE OF THE LACTOSE OPERON REPRESSOR
REMARK   1  TITL 2 AND ITS COMPLEXES WITH DNA AND INDUCER
REMARK   1  REF    SCIENCE                         V. 271  1247 1996
REMARK   1  REFN   ASTM SCIEAS  US ISSN 0036-8075
SEQRES   1 D   21    G   A   A   T   T   G   T   G   A   G   C   G   C
SEQRES   2 D   21    T   C   A   C   A   A   T   T
...
SEQRES   1 A  333  MET LYS PRO VAL THR LEU TYR ASP VAL ALA GLU TYR ALA
SEQRES   2 A  333  GLY VAL SER TYR GLN THR VAL SER ARG VAL VAL ASN GLN
SEQRES   3 A  333  ALA SER HIS VAL SER ALA LYS THR ARG GLU LYS VAL GLU
...
SEQRES   1 B  333  MET LYS PRO VAL THR LEU TYR ASP VAL ALA GLU TYR ALA
SEQRES   2 B  333  GLY VAL SER TYR GLN THR VAL SER ARG VAL VAL ASN GLN
...
HETNAM     NPF ORTHONITROPHENYL-BETA-D-FUCOPYRANOSIDE
FORMUL   6  NPF    3(C12 H15 N1 O7)
HELIX    1   1 THR A    5  TYR A   12  1                                  8
HELIX    2   2 SER A   16  ASN A   25  1                                 10
HELIX    3   3 ALA A   32  ASN A   46  1                                 15
HELIX    4   4 ASN A   50  ALA A   57  1                                  8
HELIX    5   5 LEU A   73  LEU A   90  1                                 18
...
SHEET    1   A 4 SER A  93  MET A  98  0
SHEET    2   A 4 LEU A  63  THR A  68  1  N  ILE A  64   O  SER A  93
SHEET    3   A 4 GLY A 121  ASN A 125  1  O  GLY A 121   N  GLY A  65
SHEET    4   A 4 ALA A 145  PHE A 147  1  O  LEU A 146   N  ILE A 124
```

Title of this PDB submission

Authors of this PDB submission

Primary Literature Reference

Other Literature Reference(s)

Sequence of DNA chain D

Sequence of protein chain A

Sequence of protein chain B

Heteroatom (non-protein/DNA) declaration

Heteroatom definition

Define where the author thinks helices are

Define where the author thinks beta-sheets are

```
CISPEP   1 ILE C  283    PRO C  284          1        -0.51
...
CRYST1  251.438  251.438  204.791  90.00  90.00 120.00 H 3 2        54
ORIGX1      1.000000  0.000000  0.000000        0.00000
...
SCALE1      0.003977  0.002296  0.000000        0.00000
...
ATOM      1  O5*   T D   5      -6.441  91.288 -22.163  1.00136.32          O
ATOM      2  C5*   T D   5      -6.339  92.635 -21.685  1.00136.72          C
...
ATOM     12  N3    T D   5     -11.923  96.235 -21.510  1.00137.53          N
ATOM     13  C4    T D   5     -12.400  95.306 -20.603  1.00137.44          C
...
ATOM     18  P     G D   6      -4.758  96.798 -21.981  1.00137.50          P
ATOM     19  O1P   G D   6      -3.345  96.432 -21.701  1.00137.33          O
...
TER     696        A E  18
ATOM    699  C   LYS A   2       7.988 103.719  -3.783  1.00105.99          C
ATOM    700  O   LYS A   2       8.319 103.402  -4.923  1.00106.22          O
ATOM    701  N   PRO A   3       7.446 102.838  -2.929  1.00106.03          N
ATOM    702  CA  PRO A   3       7.183 101.430  -3.245  1.00106.19          C
ATOM    703  C   PRO A   3       6.592 101.207  -4.633  1.00106.39          C
...
TER    5575      ALA B 331
HETATM 5576  C1  NPF   901     -13.446  87.461  32.376  1.00 58.71          C
HETATM 5577  C2  NPF   901     -12.756  88.680  32.106  1.00 59.04          C
...
HETATM 5616  O   HOH     1       1.628  87.097  33.699  1.00 35.03          O
...
CONECT 5576 5577 5581 5585
CONECT 5577 5576 5578
...
MASTER       30    0    0   36   36    0    0    6 7838    5   40   82
END
```

This molecule has a cis-proline

Crystallographic rotation and scaling matrices

sugar atoms

base atoms

backbone atoms

Nucleic acid residue atom coordinates

Indicates end of a chain

Amino acid residue atom coordinates (parts of residues 2 and 3 of chain A are shown)

Indicates end of a chain

Heteroatom atom coordinates

Water atom coordinates

Instructions for connecting heteroatom atoms

Some wrap-up thing

Signals end of the entire file (subsequent lines ignored)

"ATOM"	Atom No	Atom Name	Residue Name	Chain	Residue Seq No	Insert Code	Atom x-coord (Å)	Atom y-coord (Å)	Atom z-coord (Å)	Occupancy	Temperature Factor	Element Symbol	Charge
1-4	7-11	13-16	18-20	22	23-26	27	31-38	39-46	47-54	55-60	61-66	77-78	79-80
ATOM	702	CA	PRO	A	3		7.183	101.430	-3.245	1.00	106.19	C	

```
12345678901234567890123456789012345678901234567890123456789012345678901234567890
         10        20        30        40        50        60        70        80
```

La structure chimique des protéines

Les motifs conservés au niveau des séquences 1D correspondent souvent à des caractéristiques structurales 3D et fonctionnelles de la famille étudiée. La conservation de ces motifs au cours de l'évolution à des emplacements spécifiques reflète leur fonction dans la stabilité de la structure protéique, la formation de site actif enzymatique, ou la formation d'un site de liaison à un ligand. Ainsi, l'étude de la structure 3D d'une protéine permet de proposer une hypothèse sur le rôle d'un motif conservé.

Superposition des structures 3D

Objectif de la superposition des protéines:

Constat N°1:

•La fonction d'une protéine est définie par certains motifs (domaines) sous leurs formes natives (actives) en 3D.

- Une séquence linaires, même si elle a les motifs, elle ne sera pas fonctionnelle

Constat N°2:

•Deux séquences similaires à plus de 70% peuvent avoir des structures qui se ressembles et des fonctions proches ou identiques.

Constat N°3:

•L'identité structurale des protéines ne nécessite pas automatiquement une identité linaires des séquences:

- Càd il se peut que deux séquences sont identiques à 27% mais leurs structures sont identiques à 90%

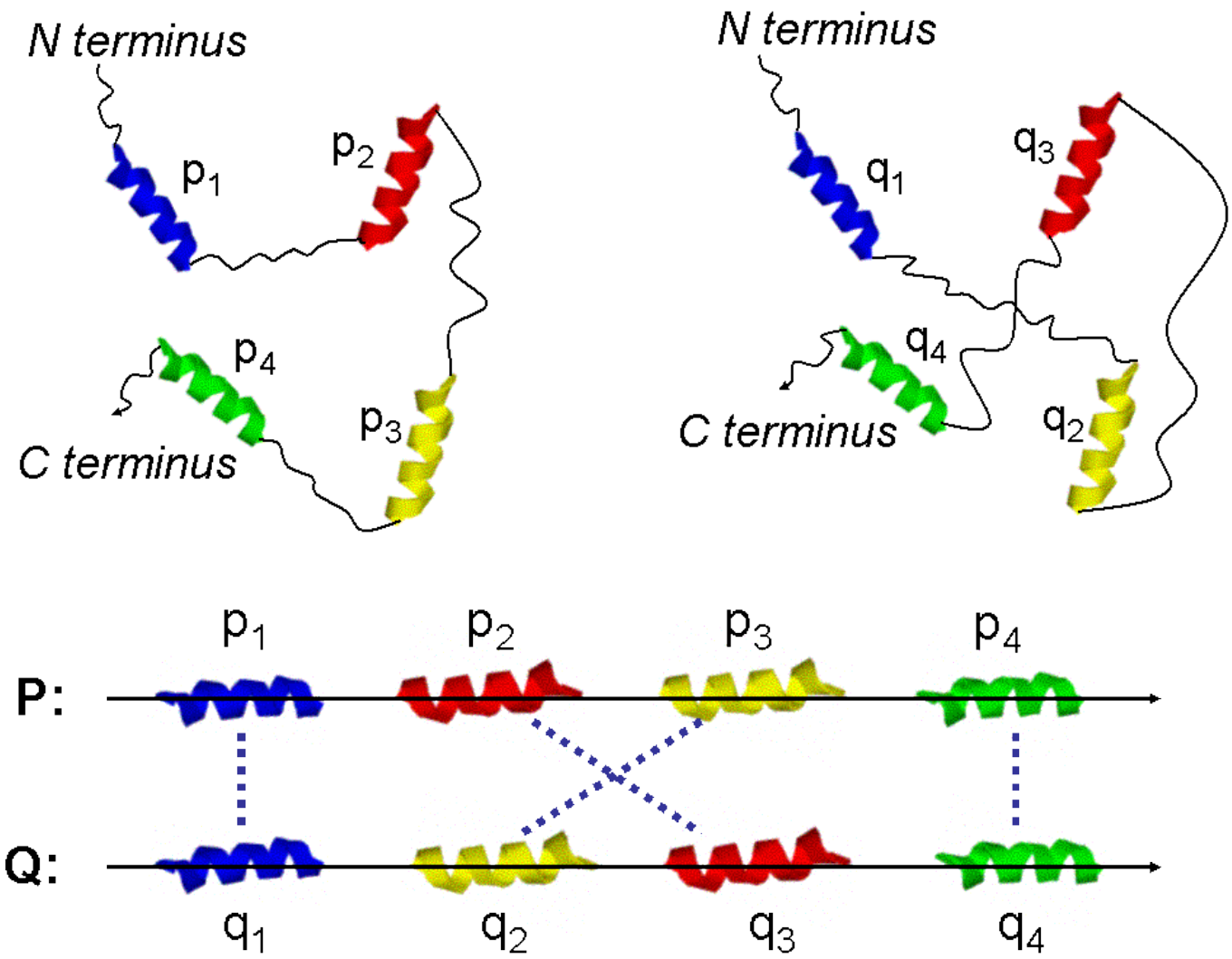
Sequence id: 27%

Structural id: 90%

Human Hemoglobin alpha-chain pdb:1jebA

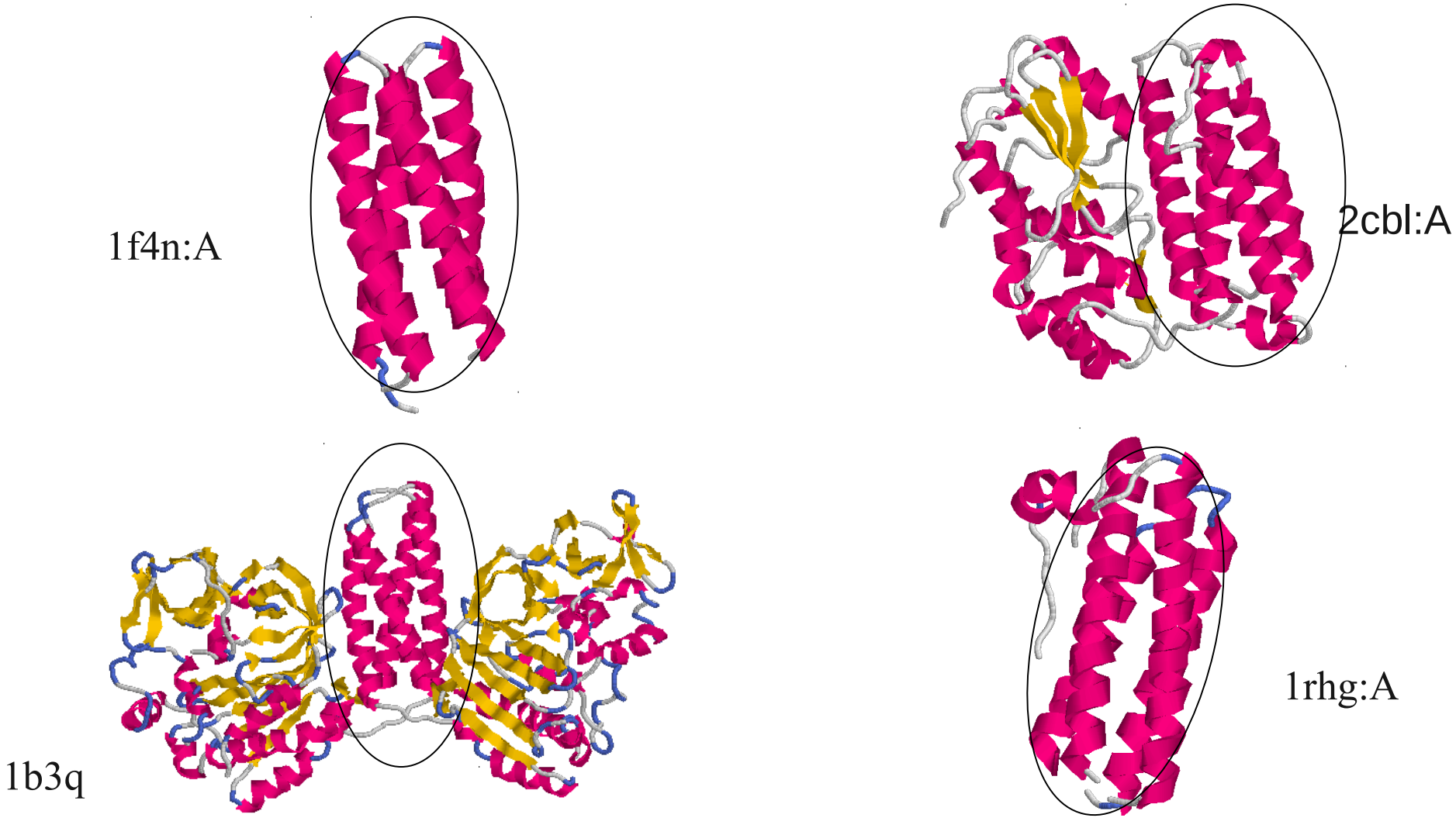
Human Myoglobin pdb:2mm1

Superposition des structures 3D



Superposition des structures 3D

Bouquet à 4 hélices



La superposition structurale est un outil qui permet de comparer des séquences avec un faible pourcentage de similarité

Superposition des structures 3D

Domaine d'application

1) La superposition des structures peut se faire sur deux ou plusieurs séquences

2)Elle utilise que les structures connues (RMN/X-ray, cristallo) ou prédites (PDB file)

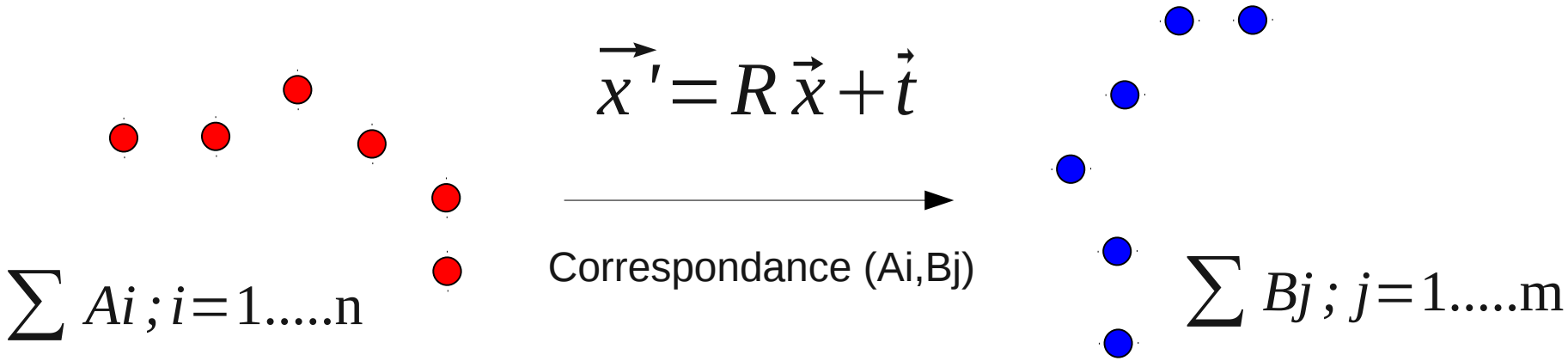
3)La superposition permet d'évaluer une structure prédite sans homologue connu.

4)Le output de la superposition est une double matrices de coordonnés d'atomes et leurs écart-types moyens (distance cartésienne: RMSD (Root Mean Squared Deviation)

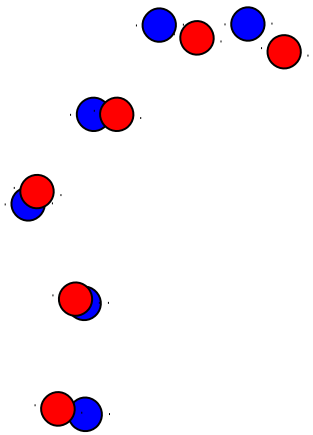
Procédure d'alignement de structure

Cas simple: **Deux protéines identiques avec le même nombre d'AA**

$$\vec{x}' = R\vec{x} + \vec{t}$$



Une fois que les relations atomes-atomes entre deux structures ont été établis, le travail d'un programme de superposition consiste à obtenir la meilleure superposition avec le plus petit RMSD possible



$$RMSD(Ai, Bj) = \sqrt{\sum |Ai - Bj|^2 / n}$$

Procédure d'alignement de structure

1. Alignement global des séquences
Relier les atomes similaires en paires

2.1 Translation : $\vec{x}' = \vec{x} + \vec{t}$

2. 2 Rotation: $\vec{x}' = R\vec{x} + \vec{t}$

2.3 Échelle: $\vec{x}' = s(R\vec{x} + \vec{t})$

$$RMSD(Ai, Bj) = \sqrt{\sum |Ai - Bj|^2 / n}$$

Root **M**ean **S**quared **D**eviation
écart-type à la valeur moyenne

Exemple de programme de superposition

Dali server

Institute of Biotechnology

SERVICES & TOOLS | GROUP MEMBERS | NEWS & VACANCIES | RESEARCH | PUBLICATIONS

Protein Structure Database Searching by DaliLite v. 3

The Dali server is a network service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the Protein Data Bank (PDB). You receive an email notification when the search has finished. In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences.

Requests can also be submitted by e-mail to *dali-server at helsinki dot fi*. The body of the e-mail message must contain atomic coordinates in PDB format.

If you want to know the structural neighbours of a protein already in the Protein Data Bank (PDB), you can find them in the Dali Database.

If you want to superimpose two particular structures, you can do it in the pairwise DaliLite server.

Upload a structure:

Parcourir...

Or enter PDB identifier: **chain:** (optional)

(Keyword search for PDB identifiers)

Job name:

(optional)

Enter email address for notification:

http://ekhidna.biocenter.helsinki.fi/dali_server/

Exemple de programme de superposition

Structural similarity between Acetylcholinesterase and Calmodulin found using CE (Tsigelny et al, *Prot Sci*, 2000, **9**:180)

Databases and Tools for 3-D Protein Structure Comparison and Alignment

Using the Combinatorial Extension (CE) Method

Select from the following options by clicking the links on the right

More Info

FIND

Find structural alignments by selecting from **ALL** (this mode is in the process of fixing, please use REPRESENTATIVES for now)

Updated 2001-10-1

or **REPRESENTATIVES** from the PDB.

Updated 2004-08-1

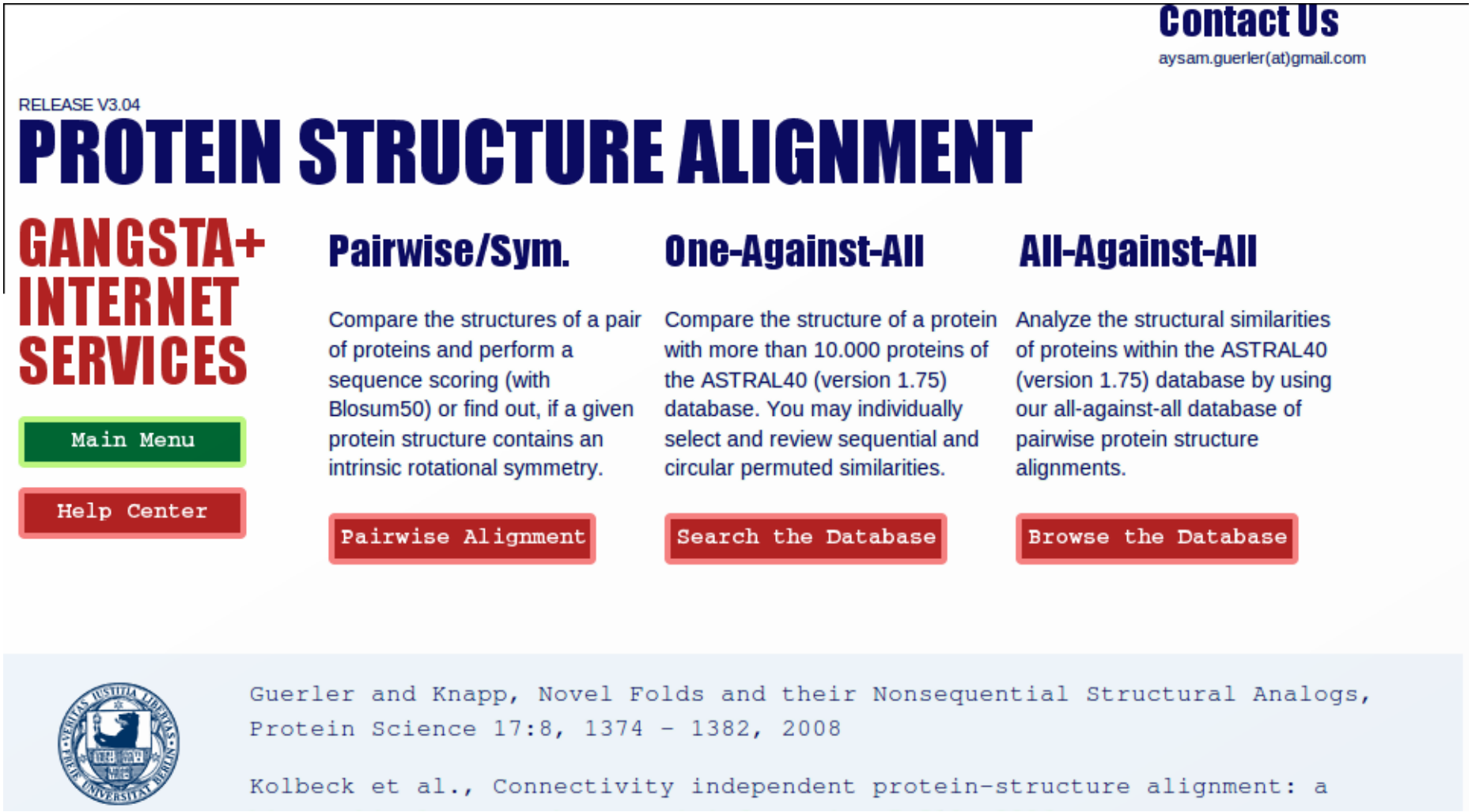
CALCULATE

Calculate structural alignment for **TWO CHAINS** either from the PDB or uploaded by the user.

Updated 2006-11-0

http://cl.sdsc.edu/

Exemple de programme de superposition



http://agknapp.chemie.fu-berlin.de/gplus/

Strap

http://www.bioinformatics.org/strap/

Couteau suisse des fichiers PDB...