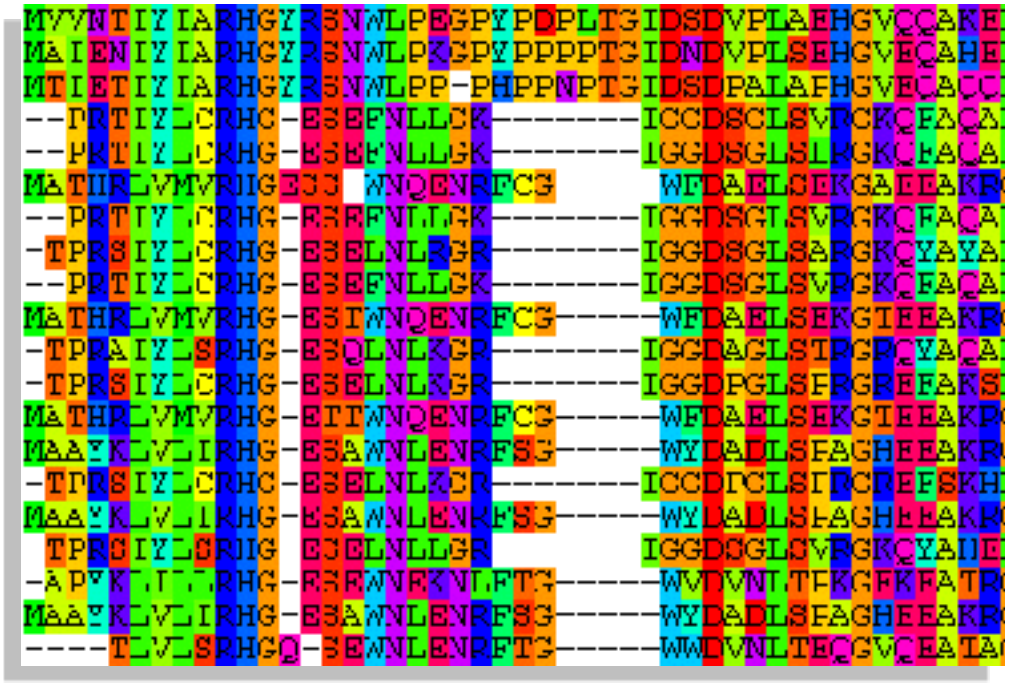
Bioinformatique pour la Biologie Structurale : Quelles informations tirer de l'analyse d'une séquence ?



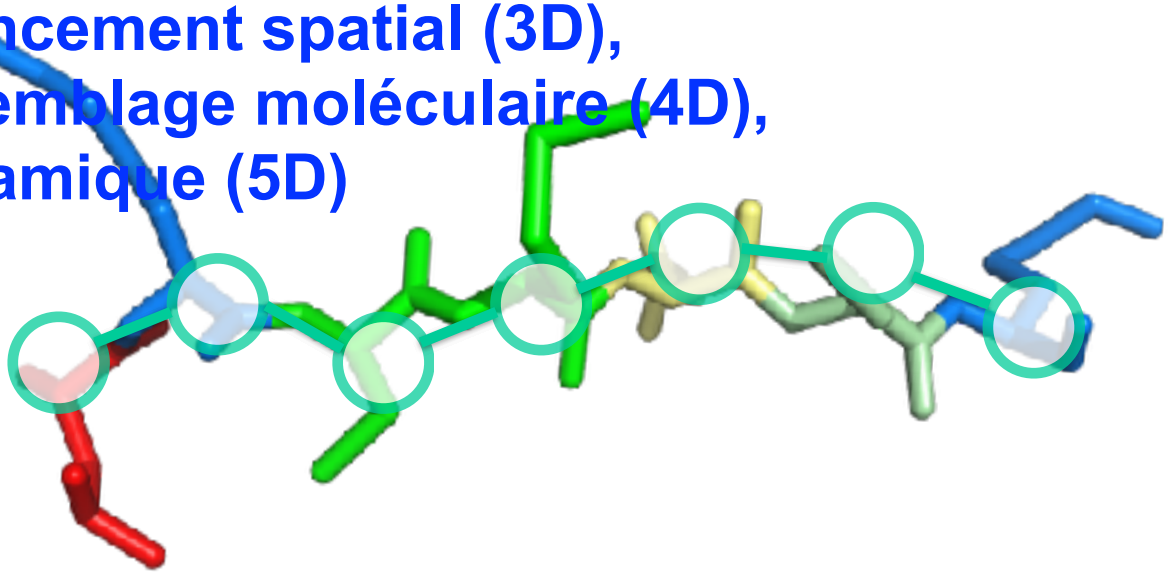
Marie-Hélène Le Du
CEA Saclay, Institut Joliot
91190 Gif sur Yvette Cedex
marie-helene.ledu@cea.fr

Qu’est-ce que la structure d’une macromolécule ?

- Séquence (1D),
- Structures secondaires (2D),
- Agencement spatial (3D),
- Assemblages moléculaires (4D),
- Dynamique (5D)

Qu'est-ce que la structure d'une macromolécule ?

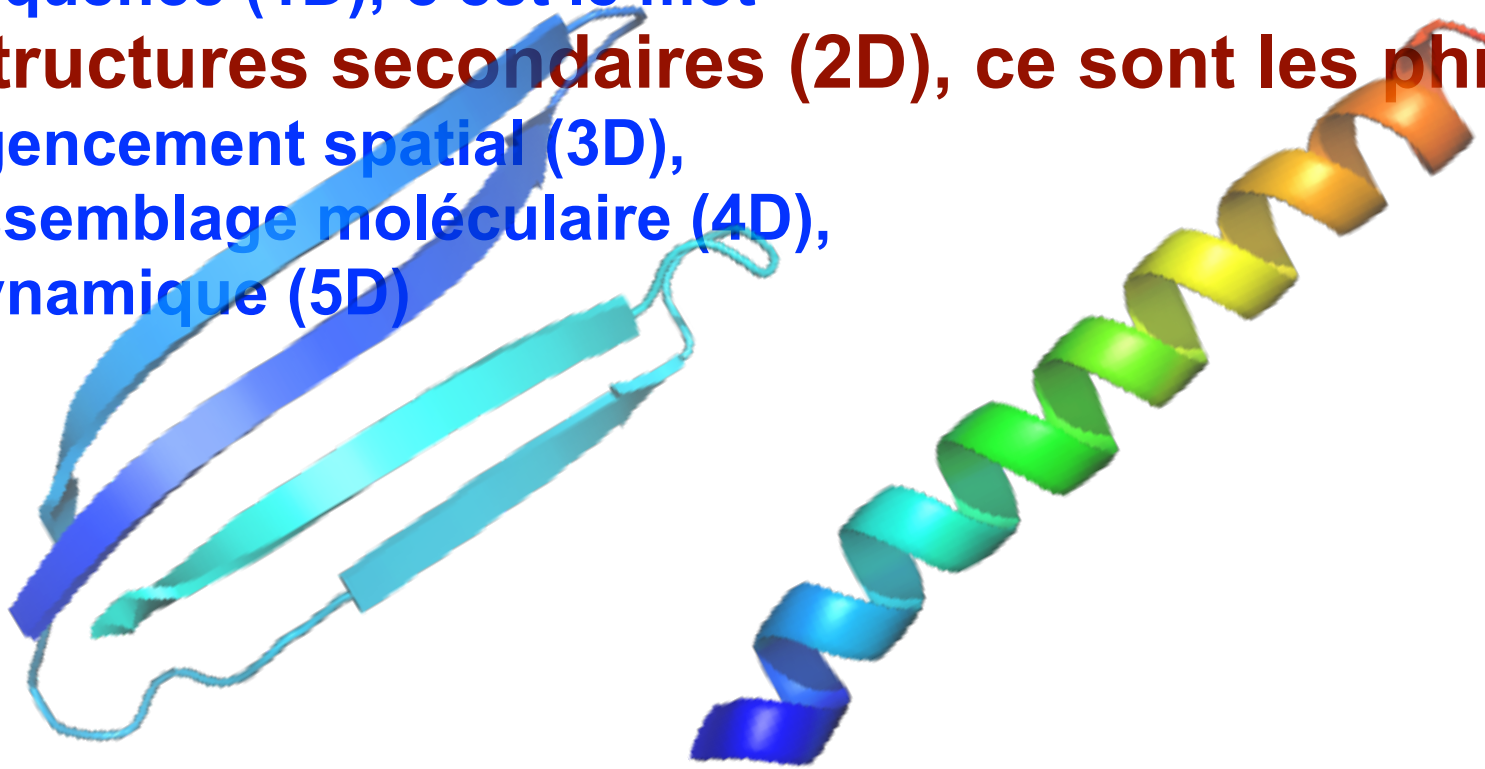
- **Séquence (1D), c'est le mot**
- Structures secondaires (2D),
- Agencement spatial (3D),
- Assemblage moléculaire (4D),
- Dynamique (5D)



Ordre des acides aminés => constitue une chaine polypeptidique avec un squelette, l'enchainement peptique, et des chaines latérales qui dépendent des acides aminés

Qu’est-ce que la structure d’une macromolécule ?

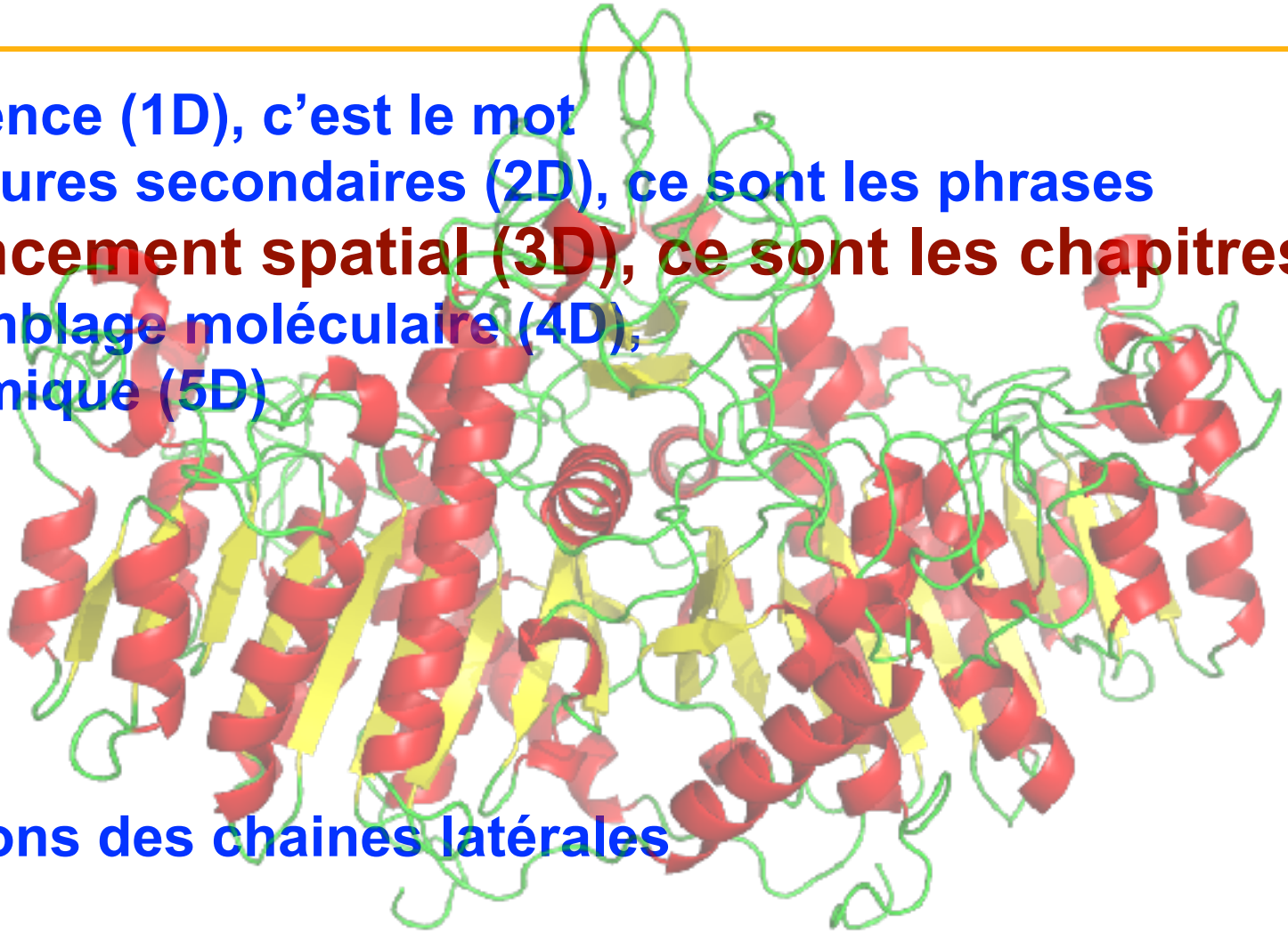
- Séquence (1D), c’est le mot
- **Structures secondaires (2D), ce sont les phrases**
- Agencement spatial (3D),
- Assemblage moléculaire (4D),
- Dynamique (5D)



Interactions du squelette;
exemples: feuillets bêta, hélices alpha

Qu'est-ce que la structure d'une macromolécule ?

- Séquence (1D), c'est le mot
- Structures secondaires (2D), ce sont les phrases
- **Agencement spatial (3D), ce sont les chapitres**
- Assemblage moléculaire (4D),
- Dynamique (5D)

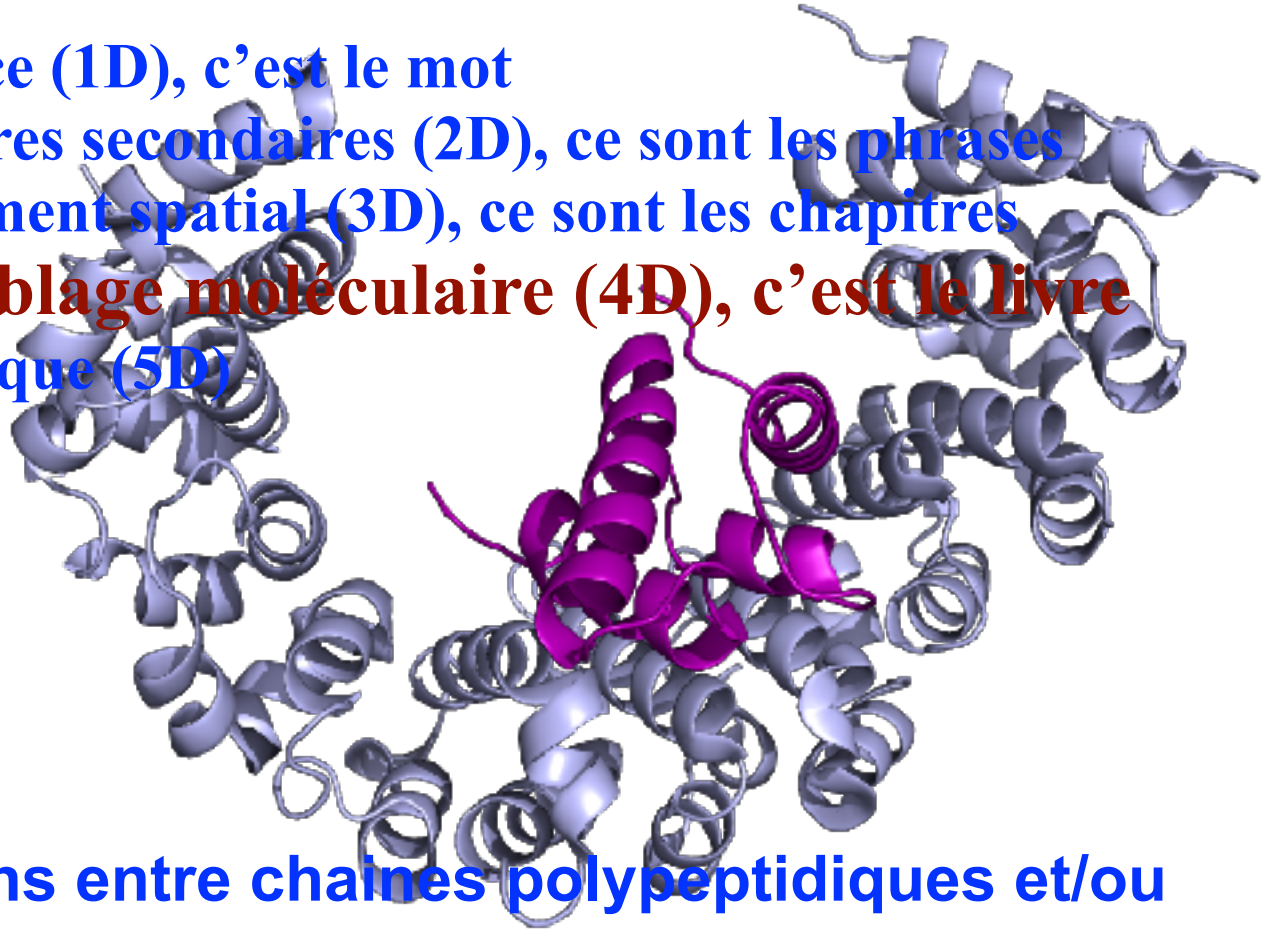


Interactions des chaines latérales

Llinas et al., Protein Sci, 2006

Qu'est-ce que la structure d'une macromolécule ?

- Séquence (1D), c'est le mot
- Structures secondaires (2D), ce sont les phrases
- Agencement spatial (3D), ce sont les chapitres
- **Assemblage moléculaire (4D), c'est le livre**
- Dynamique (5D)

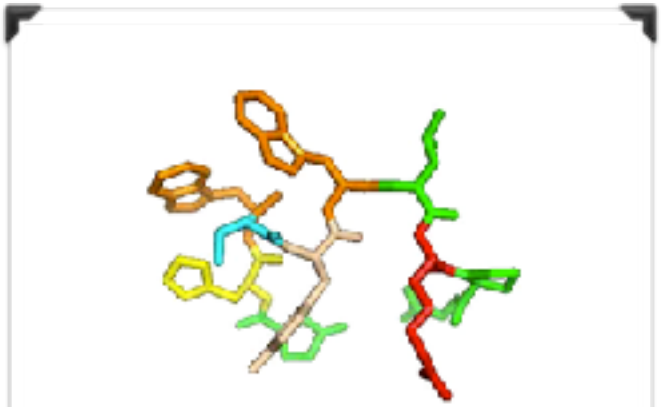


Interactions entre chaines polypeptidiques et/ou nucléotidiques

Barrault et al., PNAS, 2012

Qu'est-ce que la structure d'une macromolécule ?

- Séquence (1D), c'est le mot
- Structures secondaires (2D), ce sont les phrases
- Agencement spatial (3D), ce sont les chapitres
- Assemblages moléculaires (4D),
- **Dynamique (5D), c'est la saga**



Valéry et al., Nat. Comm, 2015

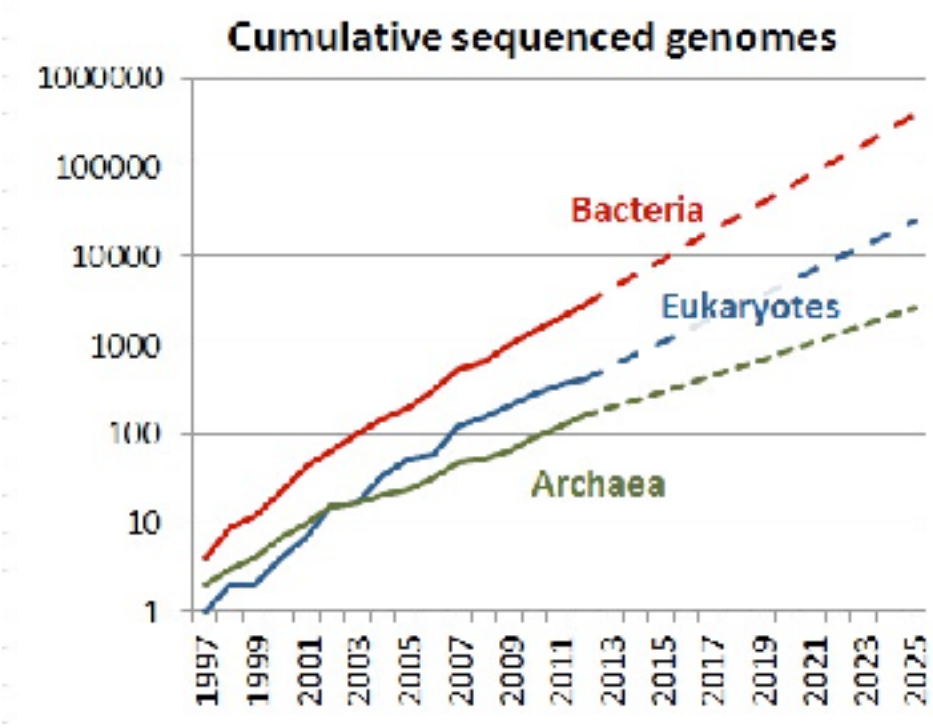
Interactions avec le milieu

La Question du Repliement des Protéines

Dans les conditions physiologiques, une protéine adopte une structure unique

```
KVFGRCELAA AMKRHGLDNY
RGYSLGNWVC AAKFESNFNT
QATNRNTDGS TDYGILQINS
RWWCNDGRTP GSRNLCNIPC
SALLSSDITA SVNCAKKIVS
DGNGMNAWVA WRNRCKGTDV
QAWIRGCRL
```

Contexte de la Prédiction de la Structure des Protéines



Nous sommes à l'ère post-génomique

=> Beaucoup de séquences, peu de fonctions, peu de structures

Contexte de la Prédiction de la Structure des Protéines

Structure:

Méthodes expérimentales :

Cristallographie

RMN

Cryo-EM

=> Quelques structures par jour à travers le monde

… incompatible avec le nombre de gènes séquencés

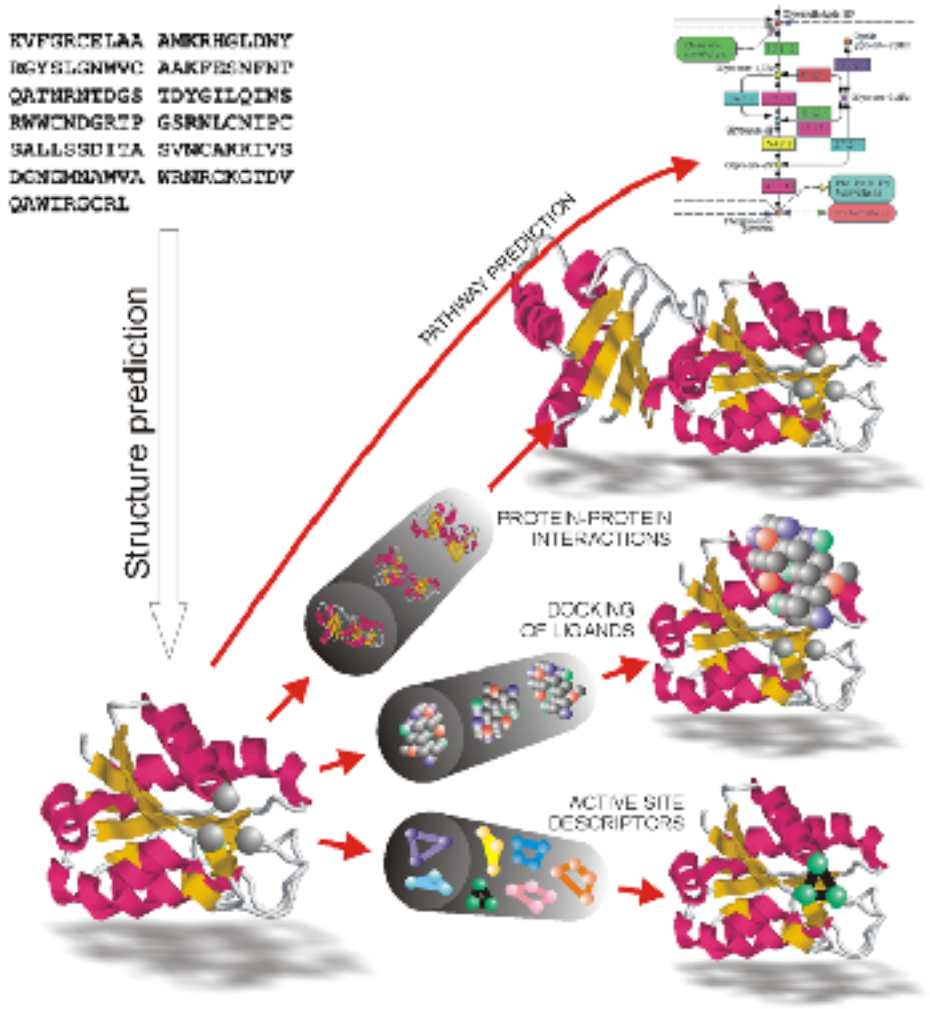
Forte demande en prédiction de structure:

plus de 30,000 gènes chez l'homme.

nombreux génomes seront séquencés au cours des 10 prochaines années.

Pourquoi prédire la structure ?

Pourquoi prédire la structure ?



Et avant une étude de biologie structurale: Pourquoi prédire la structure ?

Délimitation des domaines structurés

Méthodes actuelles de prédiction de structure 3D

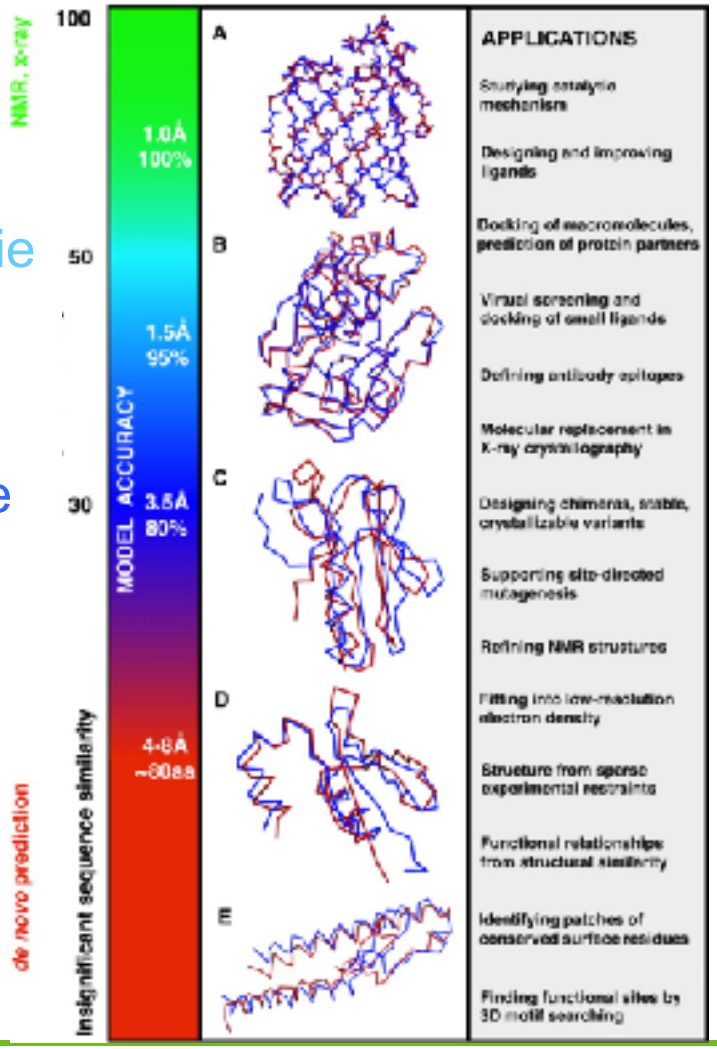
Expérimentales

Modélisation par homologie ou comparative

Beginning of twilight zone

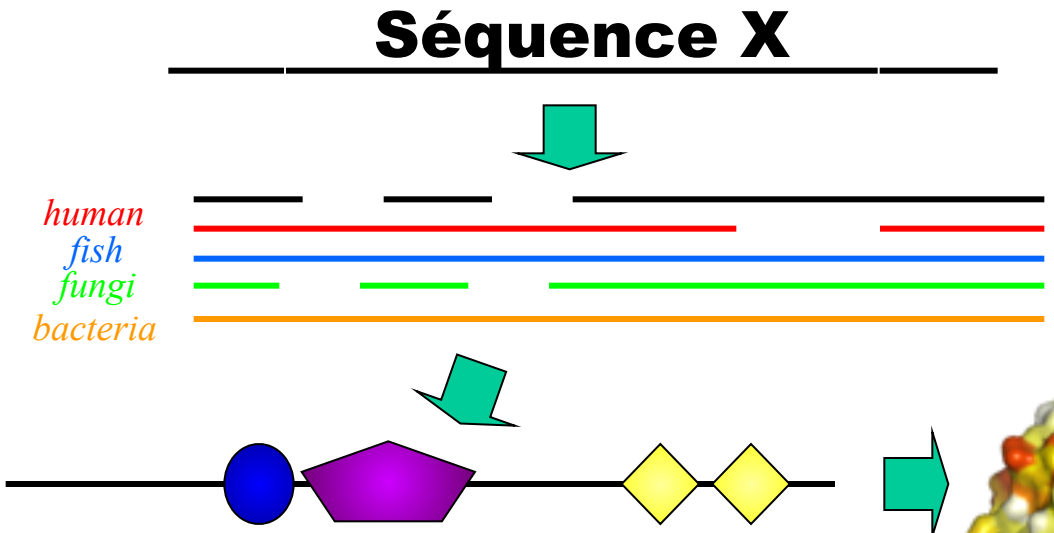
Profile-profile required

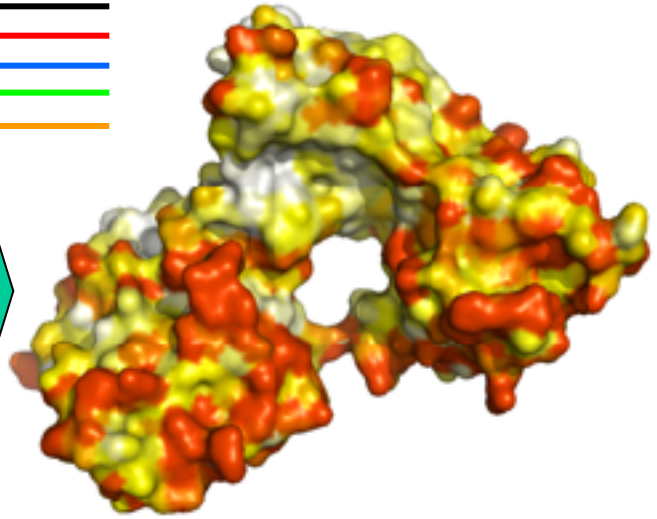
Ab initio



Baker & Sali, *Science* **294**, 2001, pp. 93-96

Quelles informations peut-on extraire d'une analyse de séquence ?





1 - Ma séquence est-elle analogue à une séquence connue ?

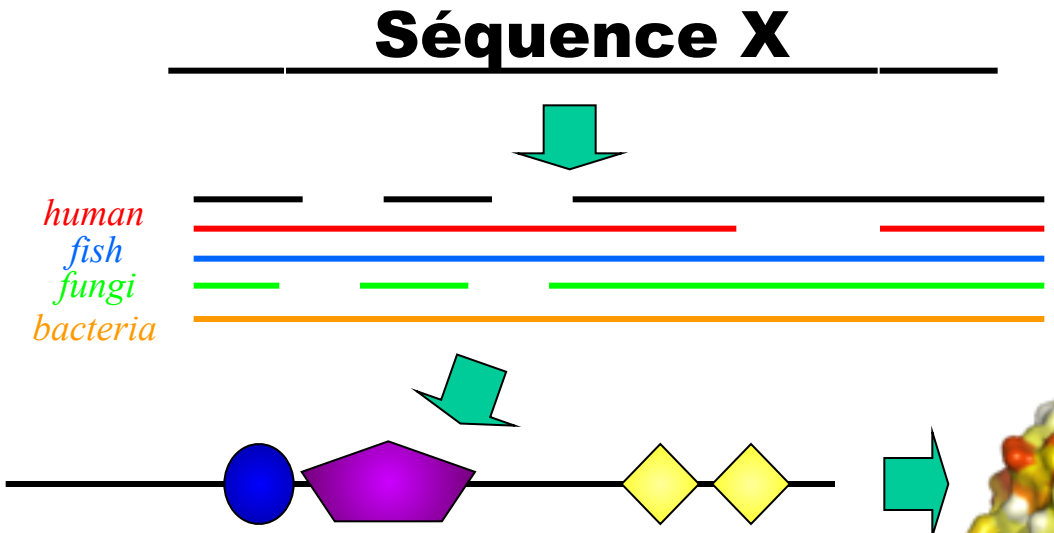
2 - Puis-je effectuer des inférences fonctionnelles ?

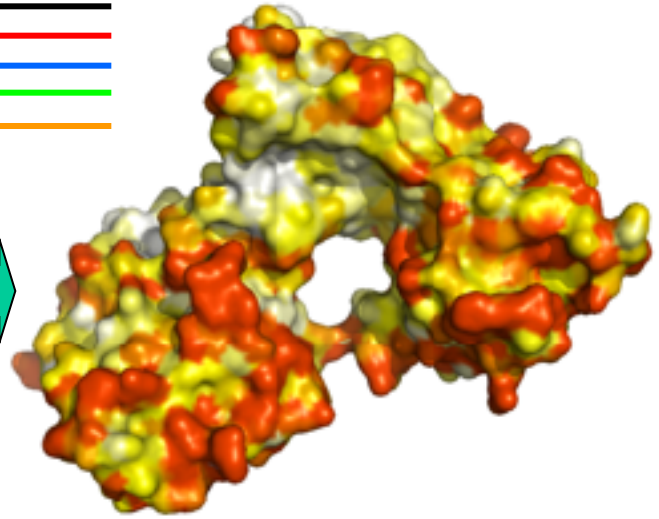
3 - Comment découper ma séquence en modules fonctionnels repliés?
- Quelles sont les régions fonctionnelles et structurales ?

4 - Peut-on prédire son organisation structurale ?

5 - Quelles sont les régions soumises à des pressions de sélection particulières ?

Quelles informations peut-on extraire d'une analyse de séquence ?





1 - Ma séquence est-elle analogue à une séquence connue ?

2 - Puis-je effectuer des inférences fonctionnelles ?

3 - Comment découper ma séquence en modules fonctionnels repliés?
- Quelles sont les régions fonctionnelles et structurales ?

4 - Peut-on prédire son organisation structurale ?

5 - Quelles sont les régions soumises à des pressions de sélection particulières ?

Récupérer la séquence correcte de ma protéine

Type d'analyse	Approche	URL	Commentaires
Choix de la séquence	mots-clés, nom de protéine, fragment de séquence	http://www.uniprot.org/	Sites généraux d'accès aux bases de données de protéomique / génomique,…
		http://www.expasy.org/	
Question biologique	Gene Ontology (GO)	http://geneontology.org/	Gene functions
	CAZY, Cazypedia	https://www.cazypedia.org/index.php/Main_Page	Bases de données spécifiques glycobiologie
Choix de l'organisme	Bibliographie	Pubmed, Isi web of science, Google Scholar,…	

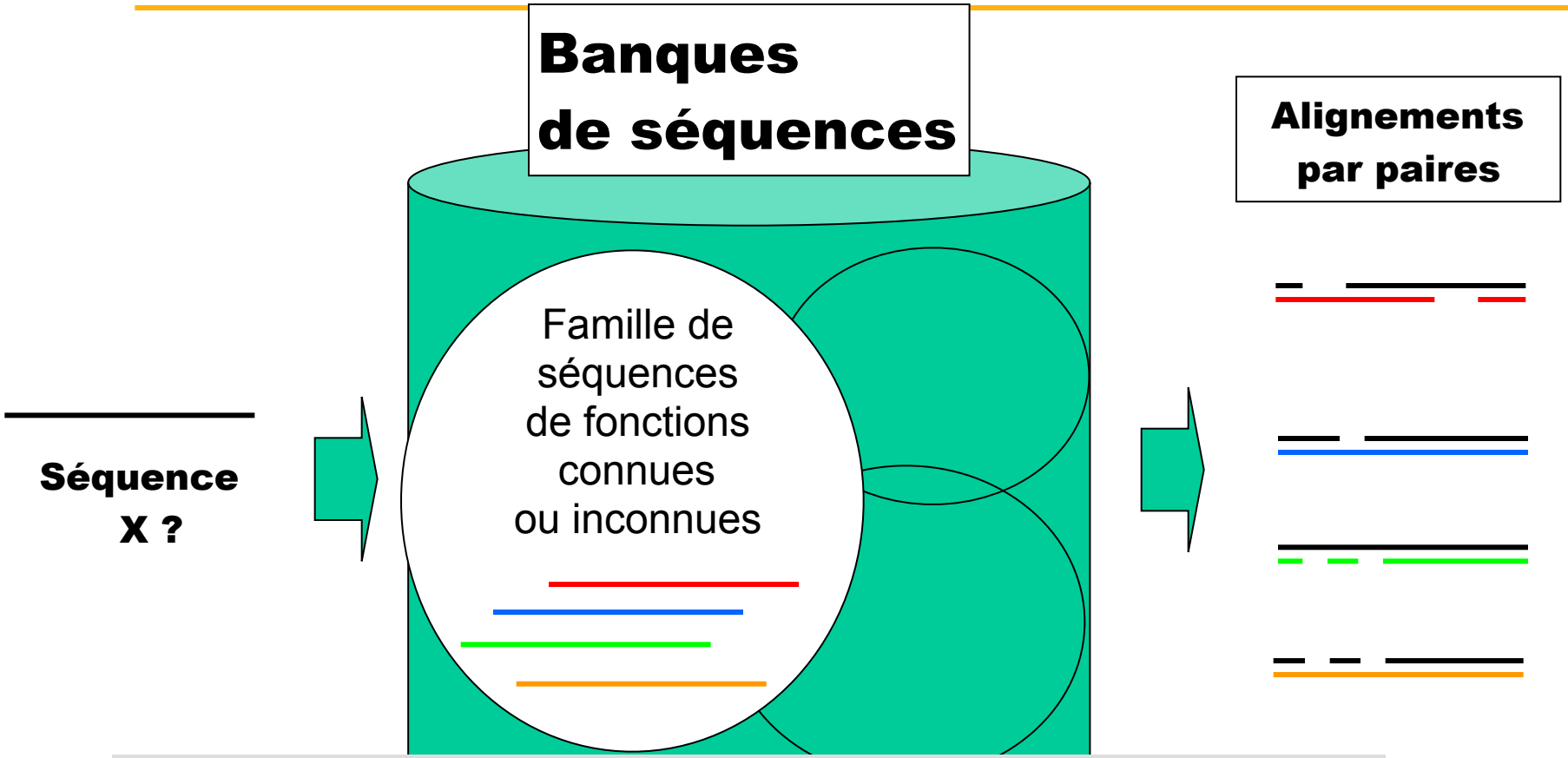
Expasy

Categories
proteomics
genomics
structure analysis
systems biology
evolutionary biology
population genetics
transcriptomics
biophysics
imaging
IT infrastructure
medicinal chemistry
glycomics
Resources A..Z
Links/Documentation
Databases
UniProtKB · functional information on proteins · [more]
UniProtKB/Swiss-Prot · protein sequence database · [more]
STRING · protein-protein interactions · [more]
SWISS-MODEL Repository · protein structure homology models · [more]
PROSITE · protein domains and families · [more]
ViralZone · portal to viral UniProtKB entries · [more]
neXtProt · human proteins · [more]
EMBnet services · bioinformatics tools, databases and courses · [more]
ENZYME · enzyme nomenclature · [more]
GlyTouCan · International glycan structure repository · [more]
HAMAP · UniProtKB family classification and annotation · [more]
MetaNetX · Metabolic Network Repository & Analysis · [more]
MyHits · protein domains database and tools · [more]
PaxDb · protein abundance database · [more]
Protein Spotlight · informally written reviews on proteins · [more]
Rhea · expert curated resource of biochemical reactions · [more]
SugarBind · pathogen sugar-binding · [more]
SwissLipids · knowledge resource for lipid biology · [more]
SwissVar · variants in UniProtKB entries · [more]
TCS · interaction specificity in two-component systems · [more]
UniCarb-DB · glycan mass and structural data · [more]
UniParc (UniProt sequence archive) · UniProt sequence archive · [more]
UniPathway · metabolic pathways for the UniProtKB · [more]
UniRef (UniProt sequence clusters) · UniProtKB sequence clusters · [more]
VenomZone · portal to venom protein UniProtKB entries · [more]
World-2DPAGE Repository · gel-based proteomics data · [more]
Tools
SWISS-MODEL Workspace · structure homology-modeling · [more]
SwissDock · protein ligand docking server · [more]
2ZIP · Prediction of leucine zipper domains · [more]
AACompIdent · protein identification by aa composition · [more]
ALF · simulation of genome evolution · [more]
Alignment tools · Four tools for multiple alignments · [more]
APSSP · Advanced Protein Secondary Structure Prediction · [more]
big-PI · predict GPI modification sites · [more]
Biochemical Pathways · Biochemical Pathways · [more]
BLAST · sequence similarity search · [more]
BLAST (UniProt) · BLAST search on the UniProt web site · [more]
BLAST - NCBI · Biological sequence similarity search · [more]
Blast2Fasta · Blast to Fasta conversion · [more]
CFSSP · Protein secondary structure prediction · [more]
Click2Drug · Directory of computational drug design tools · [more]
Clustal W · Multiple sequence alignment · [more]
Clustal W2 · Multiple sequence alignment program · [more]
COILS · Prediction of Coiled Coil Regions in Proteins · [more]
Compute pI/MW · theoretical pI and Mw computation · [more]
CPHmodels · Protein homology modeling · [more]
Decrease redundancy · Sequence redundancy reduction · [more]
DIALIGN · Local multiple sequence alignment · [more]
DisEMBL · Prediction of disordered protein regions · [more]

Récupérer la séquence correcte de ma protéine

Type d'analyse	Approche	URL	Commentaires
Choix de la séquence	mots-clés, nom de protéine, fragment de séquence	http://www.uniprot.org/	Sites généraux d'accès aux bases de données de protéomique / génomique,…
		http://www.expasy.org/	
Question biologique	Gene Ontology (GO)	http://geneontology.org/	Gene functions
	CAZY, Cazypedia	https://www.cazypedia.org/index.php/Main_Page	Bases de données spécifiques glycobiologie
Choix de l'organisme	Bibliographie	Pubmed, Isi web of science, Google Scholar,…	

Recherche d'homologues
Comparaison séquence/séquence



OUTILS : Blast (http://www.ncbi.nlm.nih.gov/BLAST/)

Basic Local Alignment Search Tool

Méthodes actuelles de prédiction de structure 3D

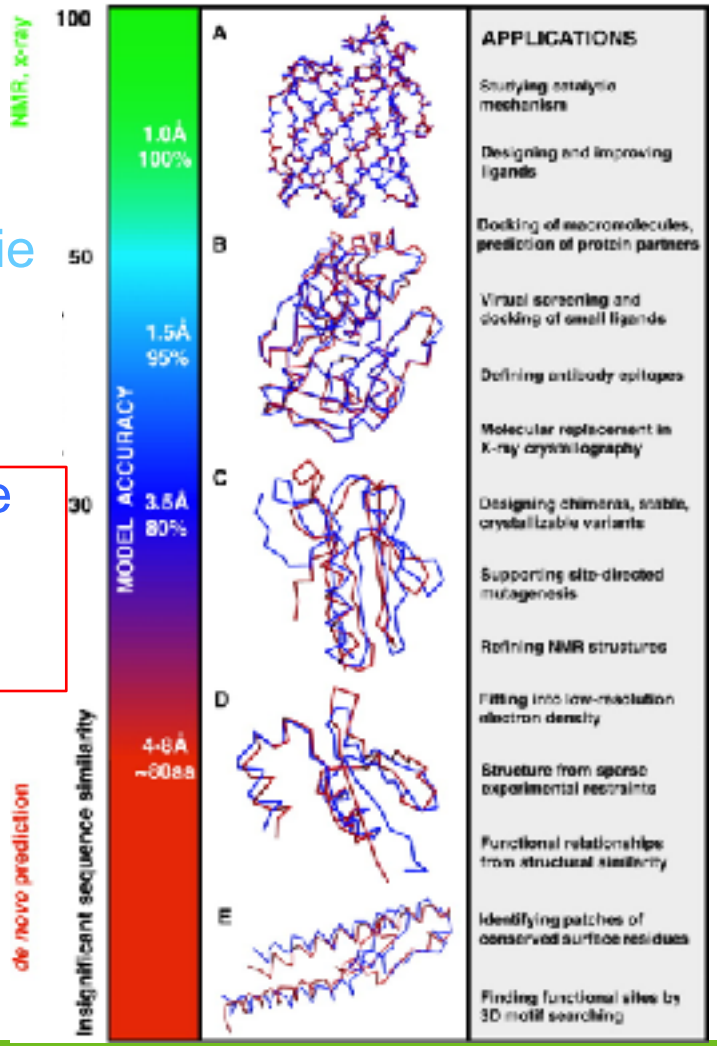
Expérimentales

Modélisation par homologie ou comparative (blast suffisant)

Beginning of twilight zone

Profile-profile required

Ab initio



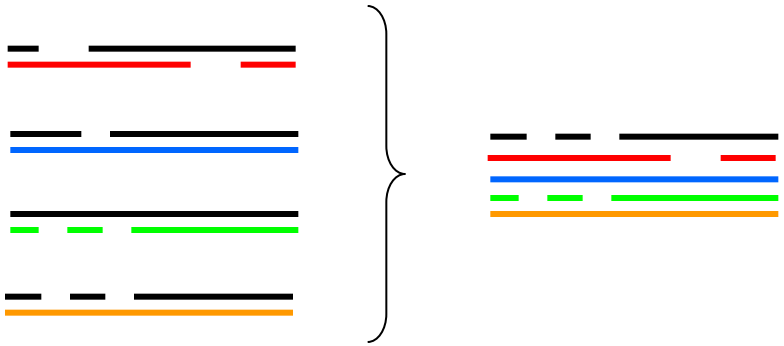
Baker & Sali, *Science* **294**, 2001, pp. 93-96

Combiner les alignments entre paires ➔ alignements multiples

Importance de l'histoire évolutive d'une famille de protéines

"pairwise alignment whispers... multiple alignment shouts out loud." (Hubbard et al., 1996)

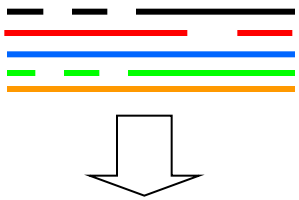
Alignements par paires



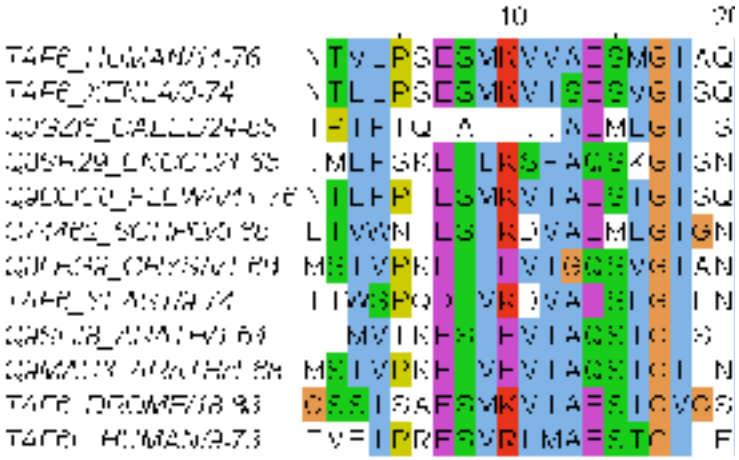
OUTILS : * Clustalw (http://www.ebi.ac.uk/Tools/clustalw/index.html/)
*** MUSCLE (http://www.drive5.com/muscle/)
*** MAFFT (http://mafft.cbrc.jp/alignment/server/)

Extraction des informations des alignements multiples → Notion de profil

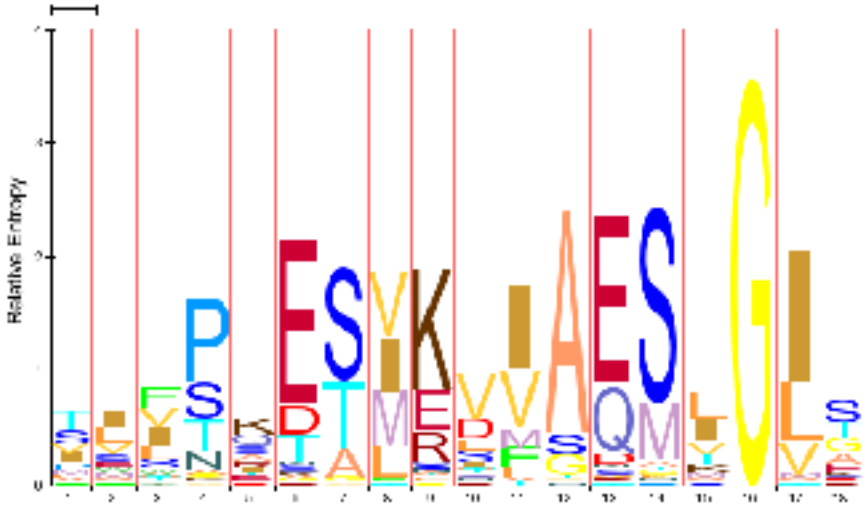
Détection des séquences reliées à la séquence cible.



Alignement Multiple



Profil ou PSSM (position specific scoring matrix)



PSI-BLAST : méthode itérative pour construire les profils

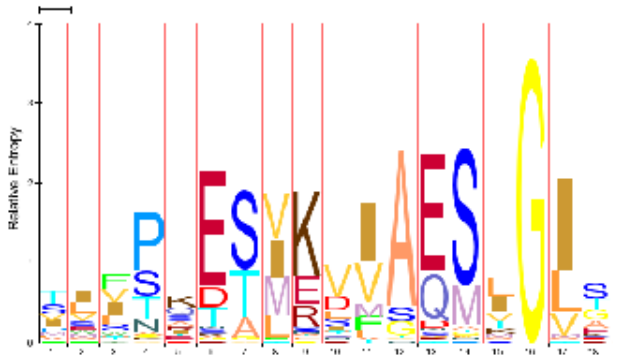
Séquence X

Blast

Base de donnée Sequences

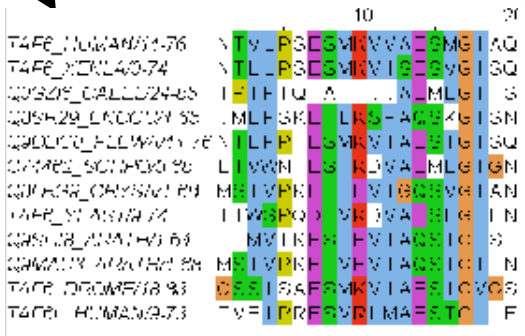
PSI-Blast

Profil



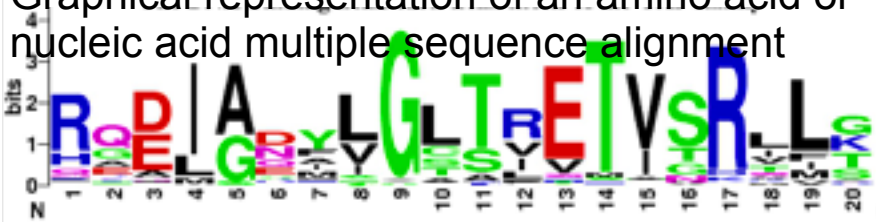
Itérations jusqu'à convergence

Alignement Multiple

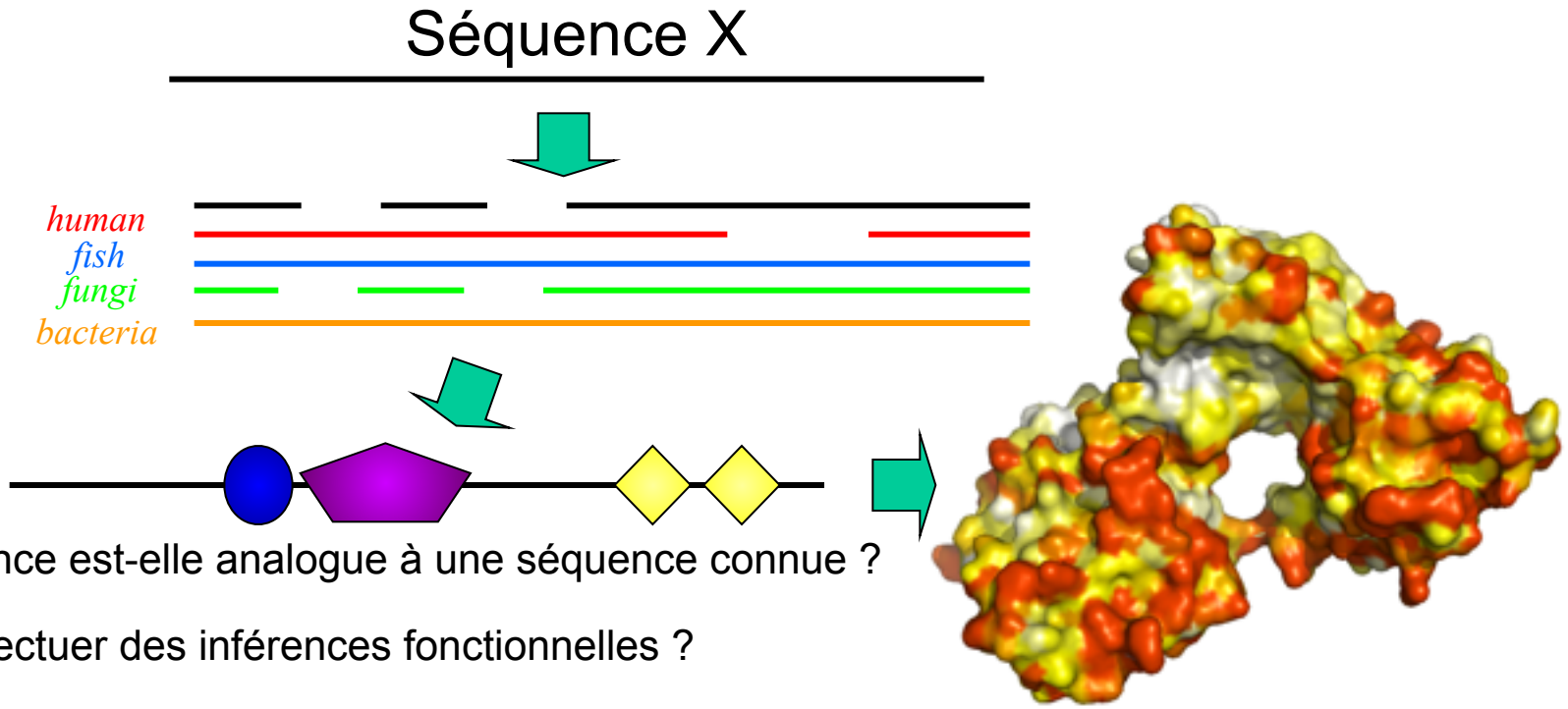


OUTILS : PSI-Blast (http://www.ncbi.nlm.nih.gov/BLAST/)

Visualisation des alignements multiples

Logiciel	URL	
Jalview	http://www.jalview.org/	Logiciel pour l'édition et la visualisation d'alignement de séquence, analyse phylogénique, analyse structurale
Geneious	http://www.geneious.com/	Logiciel pour l'édition et la visualisation d'alignement de séquence, analyse phylogénique, analyse structurale (payant mais free trial)
Alscript	http://www.compbio.dundee.ac.uk/software.html	Format multiple sequence alignments in PostScript for publications
ESPript	http://espript.ibcp.fr/ESPript/ESPript/	Easy Sequencing in PostScript': met en valeur les similarités et l'information de structures secondaires à partir d'alignements de séquences pour l'analyse et la publication
WebLogo	http://weblogo.berkeley.edu/logo.cgi	Graphical representation of an amino acid or nucleic acid multiple sequence alignment 

Quelles informations peut-on extraire d'une analyse de séquence ?



1 - Ma séquence est-elle analogue à une séquence connue ?

2 - Puis-je effectuer des inférences fonctionnelles ?

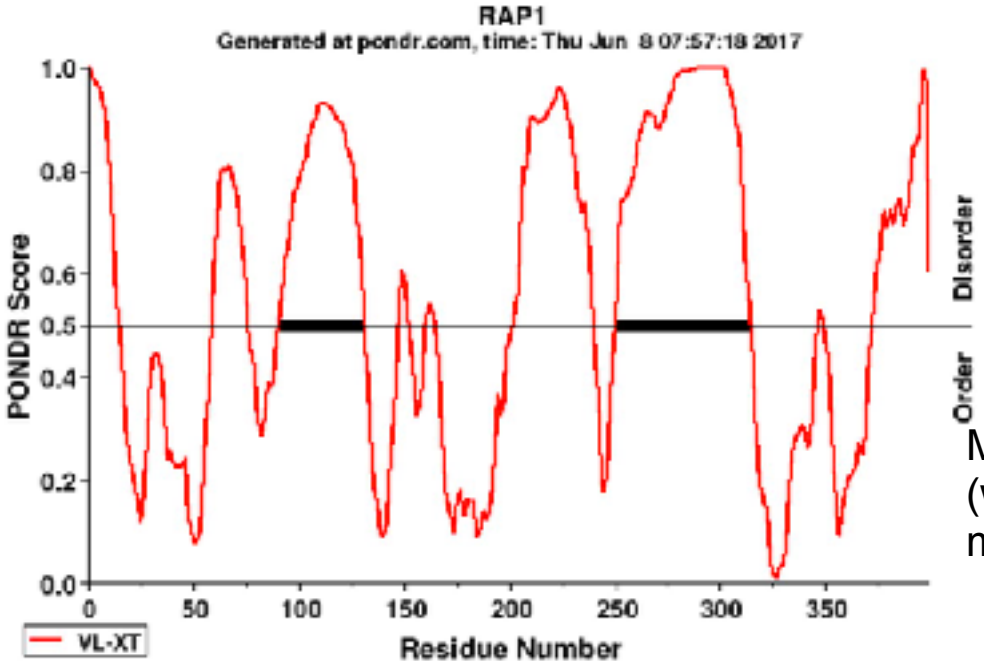
3 - Comment découper ma séquence en modules fonctionnels repliés?
- Quelles sont les régions fonctionnelles et structurales ?

4 - Peut-on prédire son organisation structurale ?

5 - Quelles sont les régions soumises à des pressions de sélection particulières ?

Recherche des régions ordonnées / désordonnées

PONDR® Predictor of Natural Disordered Regions : http://www.pondr.com/



Estimation des régions désordonnées ou ordonnées

Molecular Kinetics (www.molecularkinetics.com; main@molecularkinetics.com)

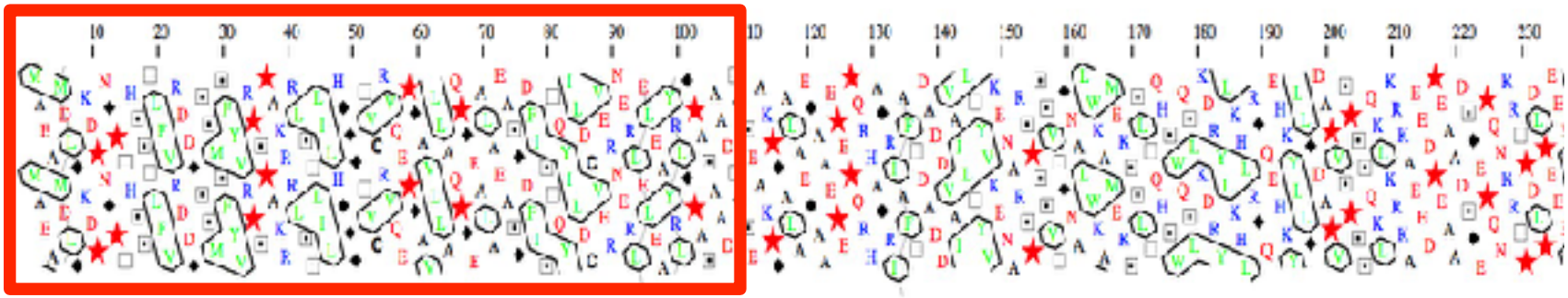
PONDRs are typically feedforward neural networks that use sequence attributes taken over windows of 9 to 21 amino acids. These attributes, such as the fractional composition of particular amino acids, hydropathy, or sequence complexity, are averaged over these windows and the values are used to train the neural network during predictor construction.

Recherche des régions ordonnées / désordonnées

HCA 1.0.2

Hydrophobic Cluster Analysis.

http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::HCA

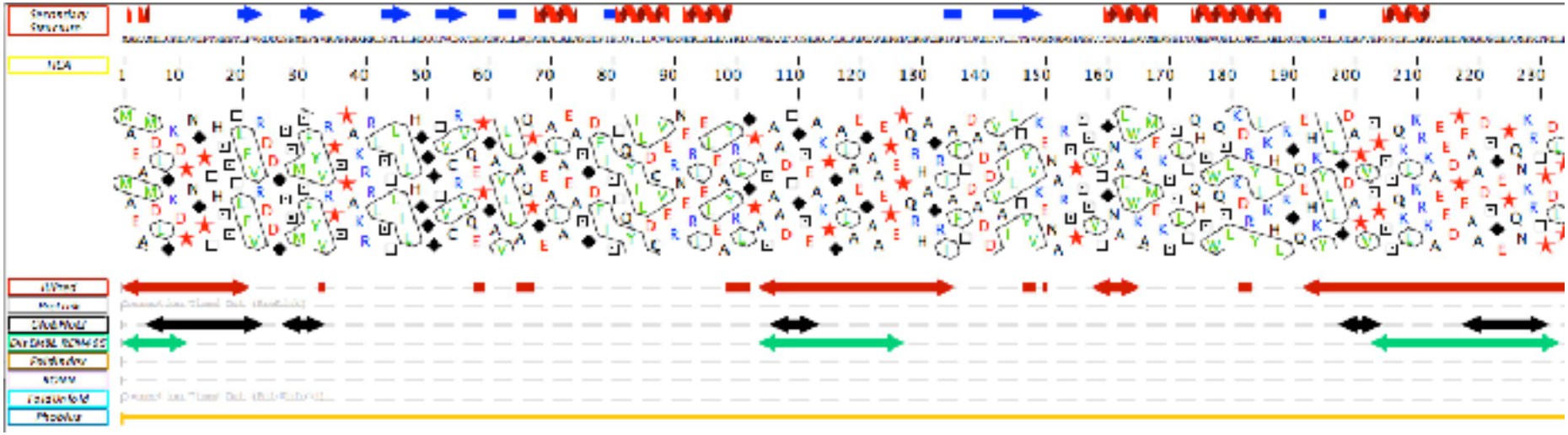


Domaine replié

Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell Mol Life Sci. 1997 Aug;53(8):621-45. Review.
Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, HenrissatB, Mornon JP.

Recherche des régions ordonnées / désordonnées

MeDor (Metaserver of Disorder) : http://www.vazymolo.org/MeDor/index.html



"MeDor: a metaserver for predicting protein disorder." Philippe Lieutaud, Bruno Canard and Sonia Longhi. BMC Genomics. 2008 Sep 16;9 Suppl 2:S25.

Exploitation des bibliothèques d'alignements multiples

→ Identification rapide de la composition en domaines

Séquence X

Existe-t-il des régions conservés ?

Base de donnée d'alignements (profils)

http://smart.embl-heidelberg.de/

- Identification de domaines
- Recherche par architecture analogue
- Interface avec les cartes interactions

http://pfam.xfam.org/

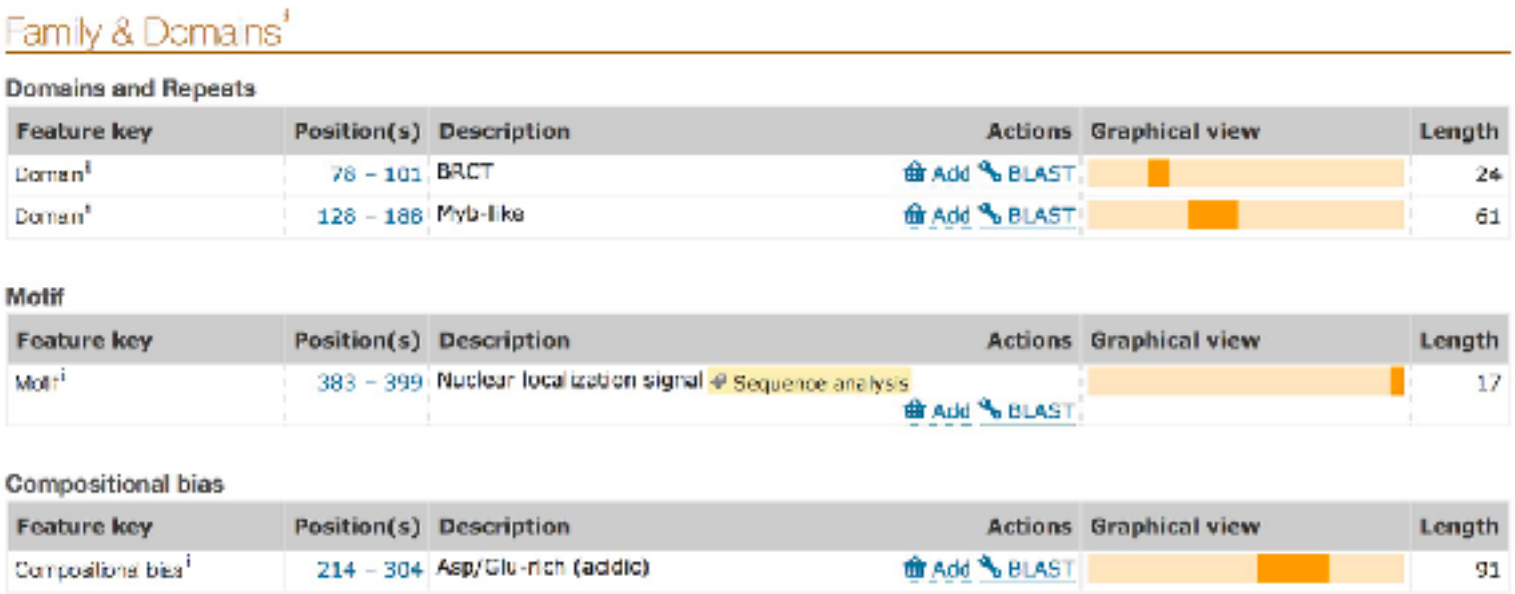
CDD, conserved domain database au NCBI

Exemple hRAP1

```
>gi|8102033|gb|AAF72711.1|AF262988_1 TRF2-interacting telomeric RAP1 protein [Homo
sapiens]
MAEAMDLGKDPNGPTHSSTLFVRDDGSSMSFYVRPSPAKRRLSTLILHGGGTVCRVQEPGAVLLAQPGEA
LAEASGDFISTQYILDCVERNERLELEAYRLGPASAADTGSEAKPGALAEGAAEPEPQRHAGRIAFTDAD
DVAILTYVKENARSPSSVTGNALWKAMEKSSLTQHSWQSLKDRYLKHLRGQEHKYLLGDAPVSPSSQKLK
RKAEEDPEAADSGEPQNKRTPDLPEEEYVKEEIQENEEAVKKMLVEATREFEEVVVDESPPDFEIHITMC
DDDPPTPEEDSETQPDEEEEEEEEKVSQPEVGAAIKIIRQLMEKFNLDLSTVTQAFLKNSGELEATSAFL
ASGQRADGYPIWSRQDDIDLQKDDEDTREALVKKFGAQNVARRIEFRKK
```

Exemple hRAP1

1/ Uniprot



Family & Domains

Domains and Repeats

Feature key	Position(s)	Description	Actions	Graphical view	Length
Domain	78 – 101	BRCT	Add BLAST		24
Domain	128 – 188	Myb-like	Add BLAST		61

Motif

Feature key	Position(s)	Description	Actions	Graphical view	Length
Motif	383 – 399	Nuclear localization signal Sequence analysis	Add BLAST		17

Compositional bias

Feature key	Position(s)	Description	Actions	Graphical view	Length
Compositional bias	214 – 304	Asp/Glu-rich (acidic)	Add BLAST		91

```
>gi|8102033|gb|AAF72711.1|AF262988_1 TRF2-interacting telomeric RAP1
protein [Homo sapiens]
MAEAMDLGKDPNGPTHSSTLFVRDDGSSMSFYVRPSPAKRRLSTLILHGGGTVCRVQEPGAVLLAQPGEA
LAEASGDFISTQYILDCVERNERLELEAYRLGPASAADTGSEAKPGALAEGAAEPEPQRHAGRIAFTDAD
DVAILTYVKENARSPSSVTGNALWKAMEKSSLTQHSWQSLKDRYLKHLRGQEHKYLLGDAPVSPSSQKLK
RKAEEDPEAADSGEPQNKRTPDLPEEEYVKEEIQENEEAVKKMLVEATREFEEVVVDESPPDFEIHITMC
DDDPPTPEEDSETQPDEEEEEEEEKVSQPEVGAAIKIIRQLMEKFNLDLSTVTQAFLKNSGELEATSAFL
ASGQRADGYPIWSRQDDIDLQKDDEDTREALVKKFGAQNVARRIEFRKK
```

Exemple hRAP1

1b/ Expasy => protparam

Number of amino acids: 399
Molecular weight: 44259.9
Theoretical pI: 4.64

Amino acid composition:

Ala (A)	40	10.0%
Arg (R)	23	5.8%
Asn (N)	9	2.3%
Asp (D)	30	7.5%
Cys (C)	3	0.8%
Gln (Q)	18	4.5%
Glu (E)	51	12.8%
Gly (G)	23	5.8%
His (H)	7	1.8%
Ile (I)	14	3.5%
Leu (L)	32	8.0%
Lys (K)	25	6.3%
Met (M)	7	1.8%
Phe (F)	11	2.8%
Pro (P)	25	6.3%
Ser (S)	30	7.5%
Thr (T)	19	4.8%
Trp (W)	3	0.8%
Tyr (Y)	8	2.0%
Val (V)	21	5.3%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%

Total number of negatively charged residues (Asp + Glu): 81
Total number of positively charged residues (Arg + Lys): 48

Atomic composition:

Carbon	C	1917
Hydrogen	H	3033
Nitrogen	N	537
Oxygen	O	646
Sulfur	S	10

Formula: $C_{1917}H_{3033}N_{537}O_{646}S_{10}$
Total number of atoms: 6143

Extinction coefficients:
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.
Ext. coefficient 28545
Abs 0.1% (=1 g/l) 0.645, assuming all pairs of Cys residues form cystines

Ext. coefficient 28420
Abs 0.1% (=1 g/l) 0.642, assuming all Cys residues are reduced

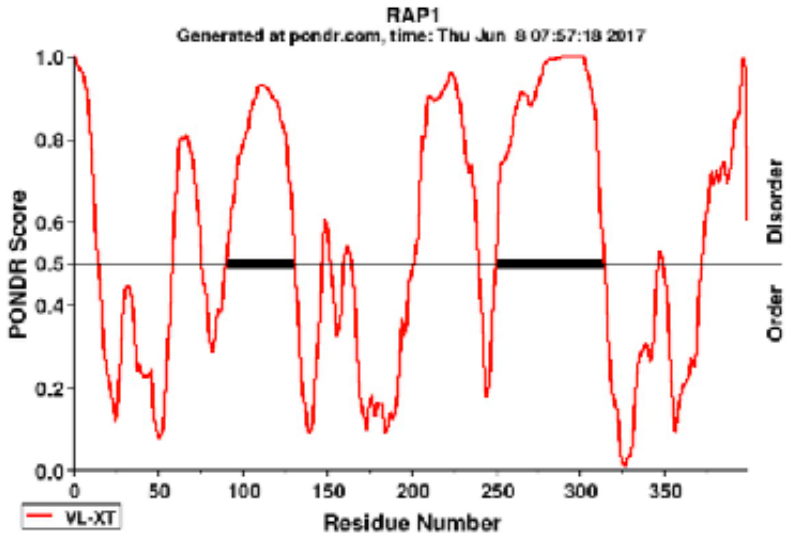
Estimated half-life:
The N-terminal of the sequence considered is M (Met).
The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
>20 hours (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).

Instability index:
The instability index (II) is computed to be 57.42
This classifies the protein as unstable.

Aliphatic index: 70.25
Grand average of hydropathicity (GRAVY): -0.763

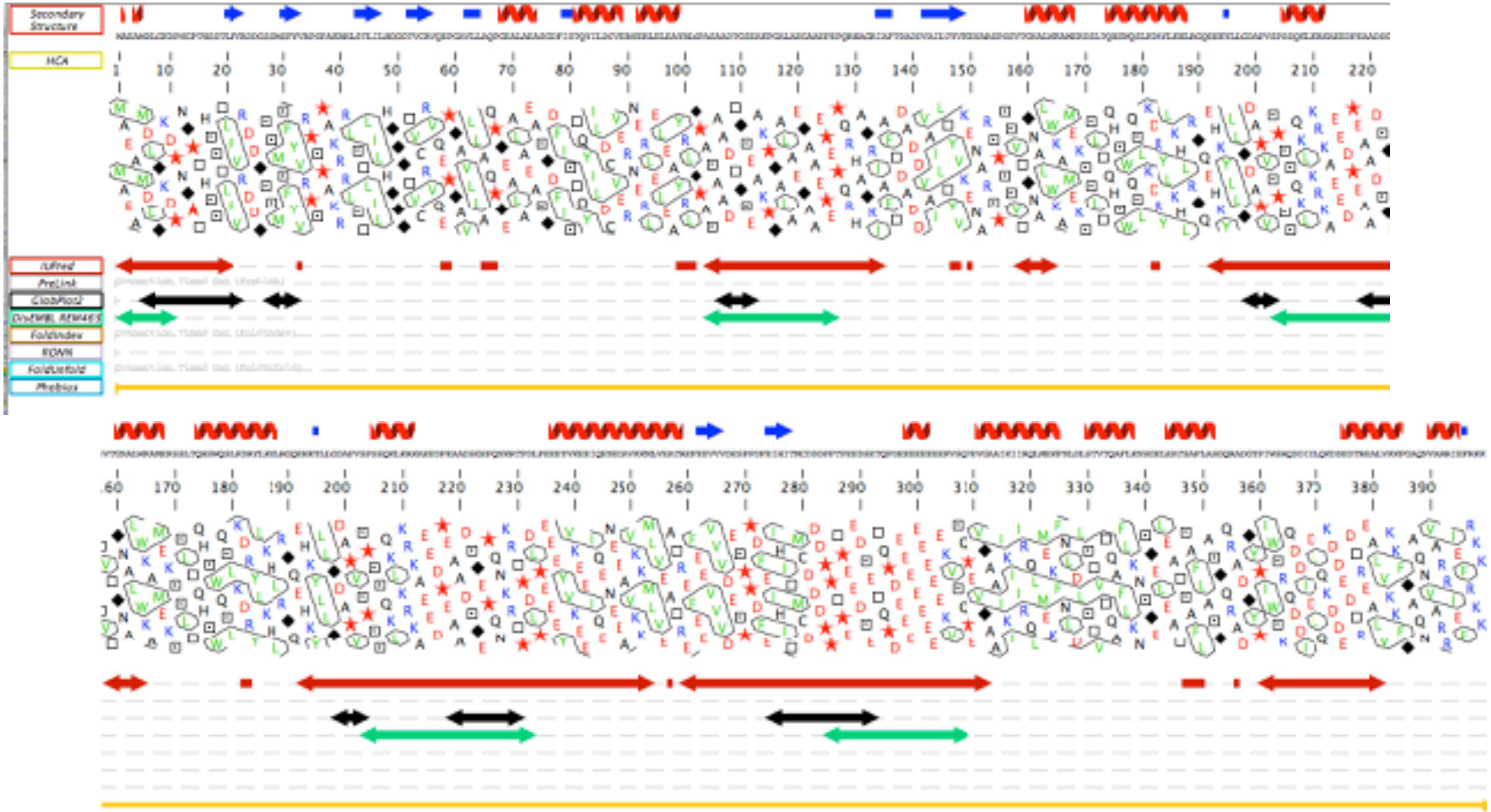
Exemple hRAP1

2/ Psi-blast // PONDR // MEDOR



Exemple hRAP1

2/ Psi-blast // PONDR // MEDOR



Exemple hRAP1

>gi|8102033|gb|AAF72711.1|AF262988_1 TRF2-interacting telomeric RAP1 protein [Homo sapiens]
MAEAMDLGKDPNGPTHSSTLFVRDDGSSMSFYVRPSPAKRRLSTLILHGGGTVCRVQEPGAVLLAQPGEA
LAEASGDFISTQYILDCVERNERLELEAYRLGPASAADTGSEAKPGALAEGAAEPEPQRHAGRIAFTDAD
DVAILTYVKENARSPSSVTGNALWKAMEKSSLTQHSWQSLKDRYLKHLRGQEHKYLLGDAPVSPSSQKLK
RKAEEDPEAADSGEPQNKRTPDLPEEEYVKEEIQENEEAVKKMLVEATREFEEVVVDESPPDFEIHITMC
DDDPPTPEEDSETQPDEEEEEEEEKVSQPEVGAAIKIIRQLMEKFNLDLSTVTQAFLKNSGELEATSAFL
ASGQRADGYPIWSRQDDIDLQKDDEDTREALVKKFGAQNVARRIEFRKK

En première approche => Découpage en 3 domaines

A confirmer par une protéolyse ménagée

Méthodes actuelles de prédiction de structure 3D

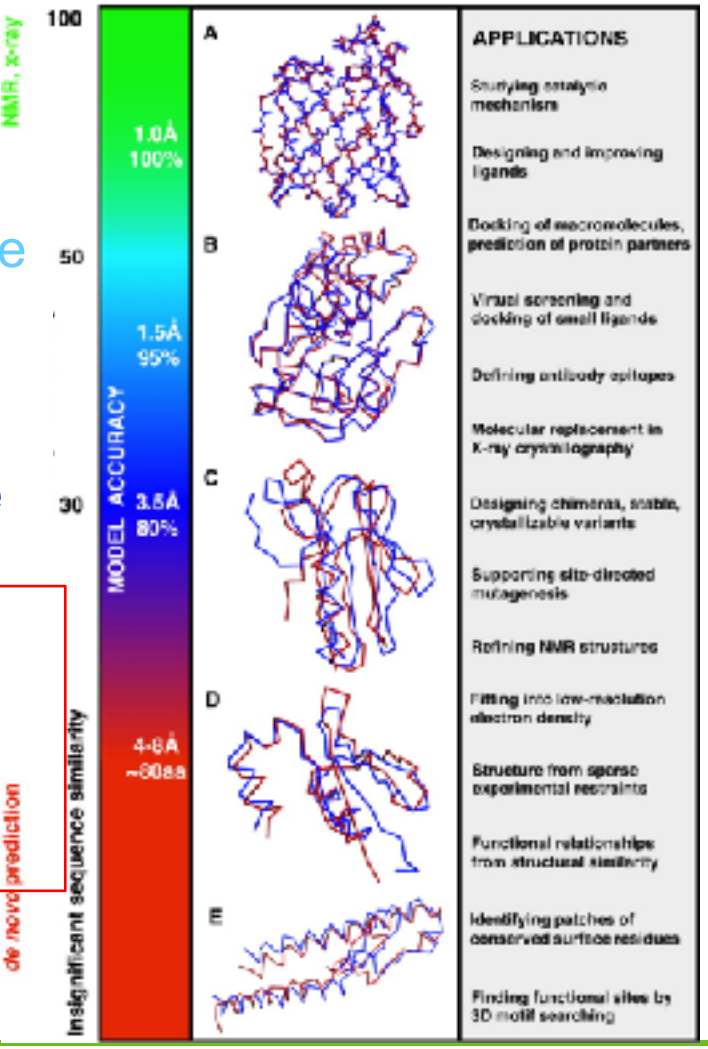
Expérimentales

Modélisation par homologie ou comparative (blast suffisant)

Beginning of twilight zone

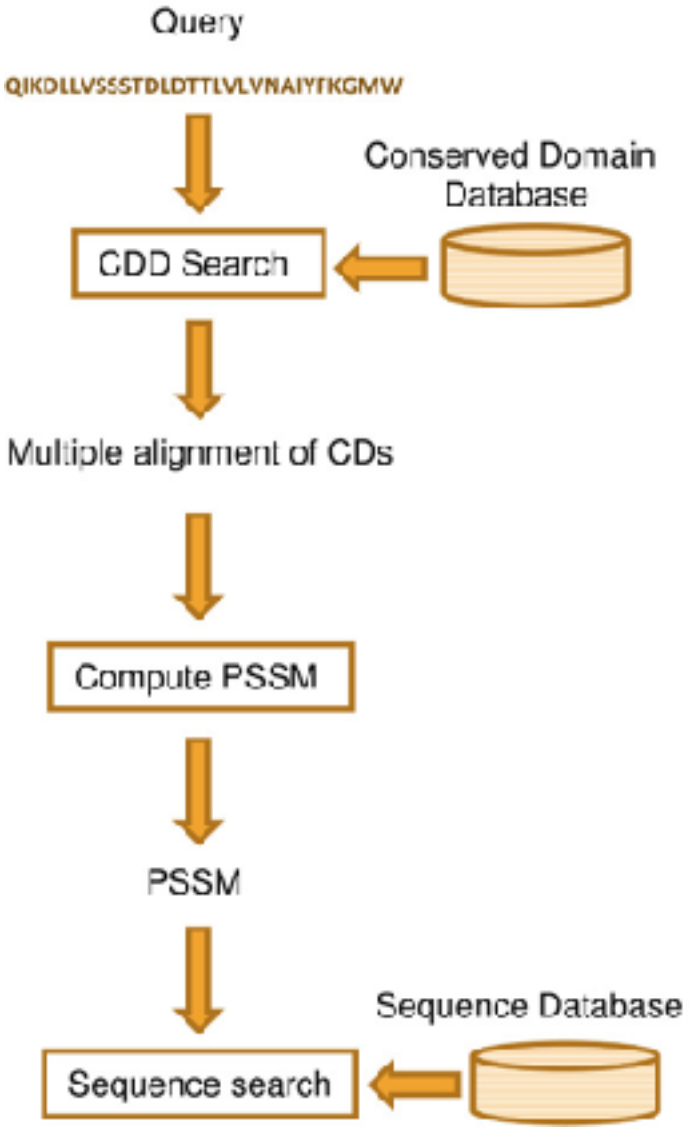
Profile-profile required

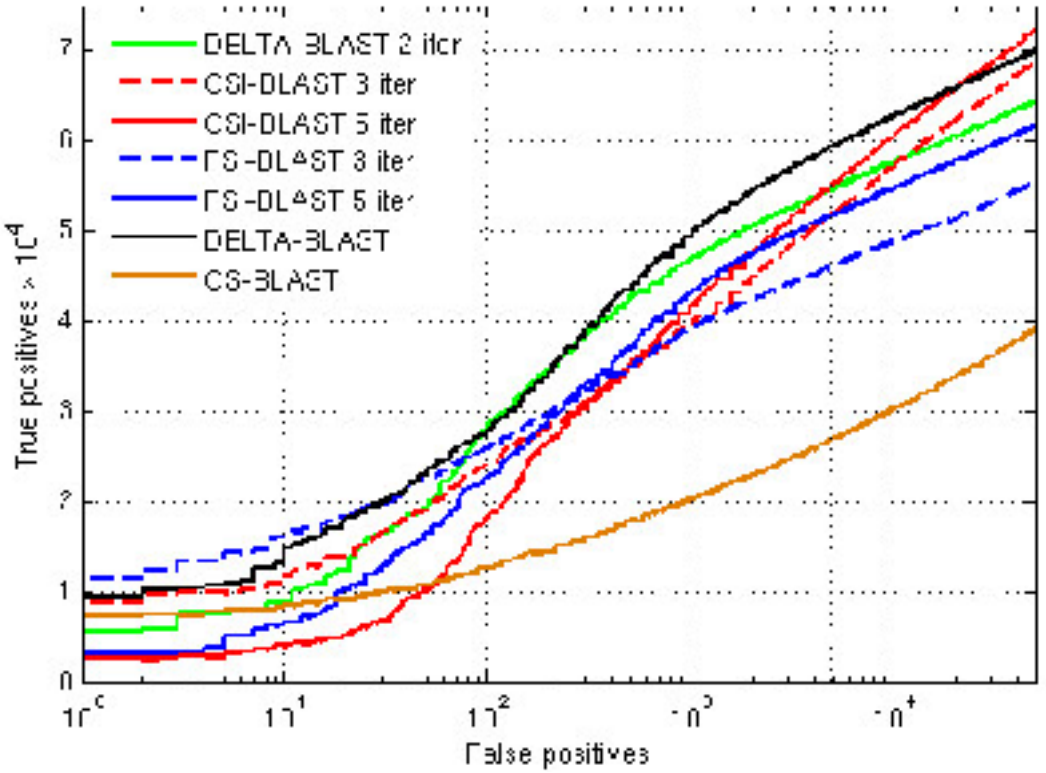
Ab initio



Baker & Sali, *Science* **294**, 2001, pp. 93-96

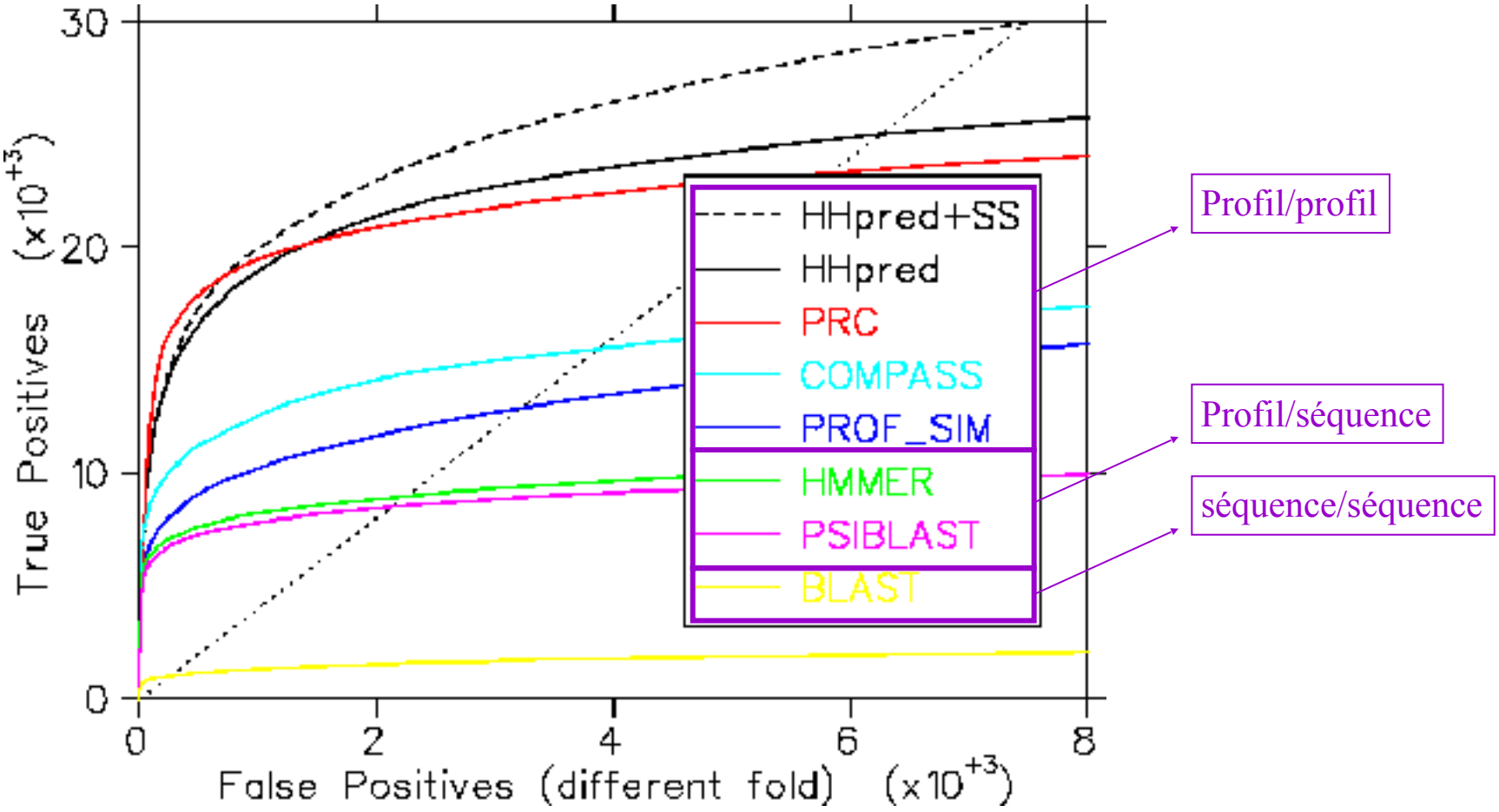
Delta-Blast : Exploitation des domaines pour l'amélioration des recherches d'homologies lointaines





Performances comparées

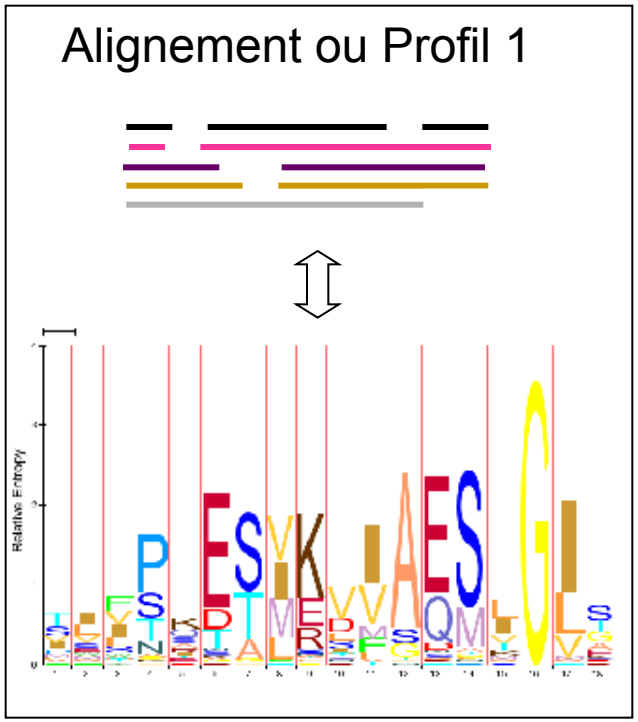
Gain très important en spécificité
grâce aux comparaisons profil/profil



J Söding. Bioinformatics 2005. 21(7):951-60

HHpred+SS : méthode comparaison profils HMM/HMM
+ prédiction de structure secondaire par PsiPred

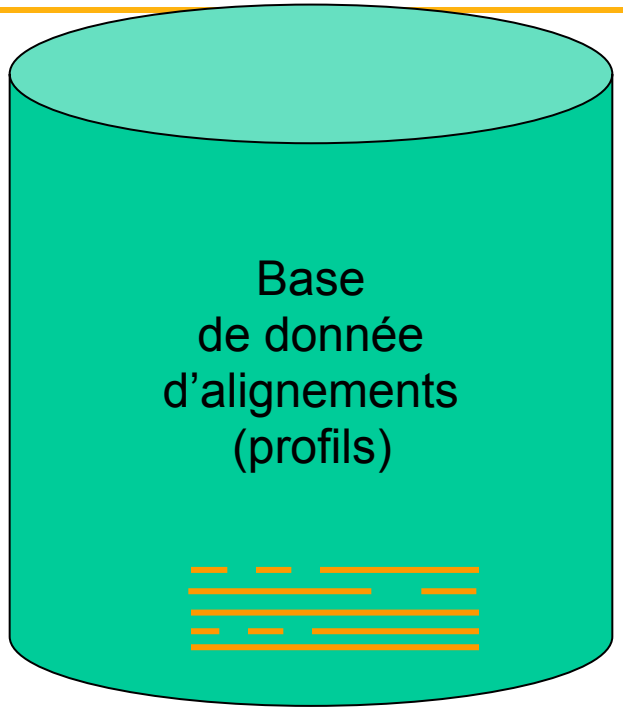
Les développements méthodologiques pour l'analyse de séquence à très haute divergence



Outils de comparaison jusqu'à 2005

ex : PSI-Blast

?



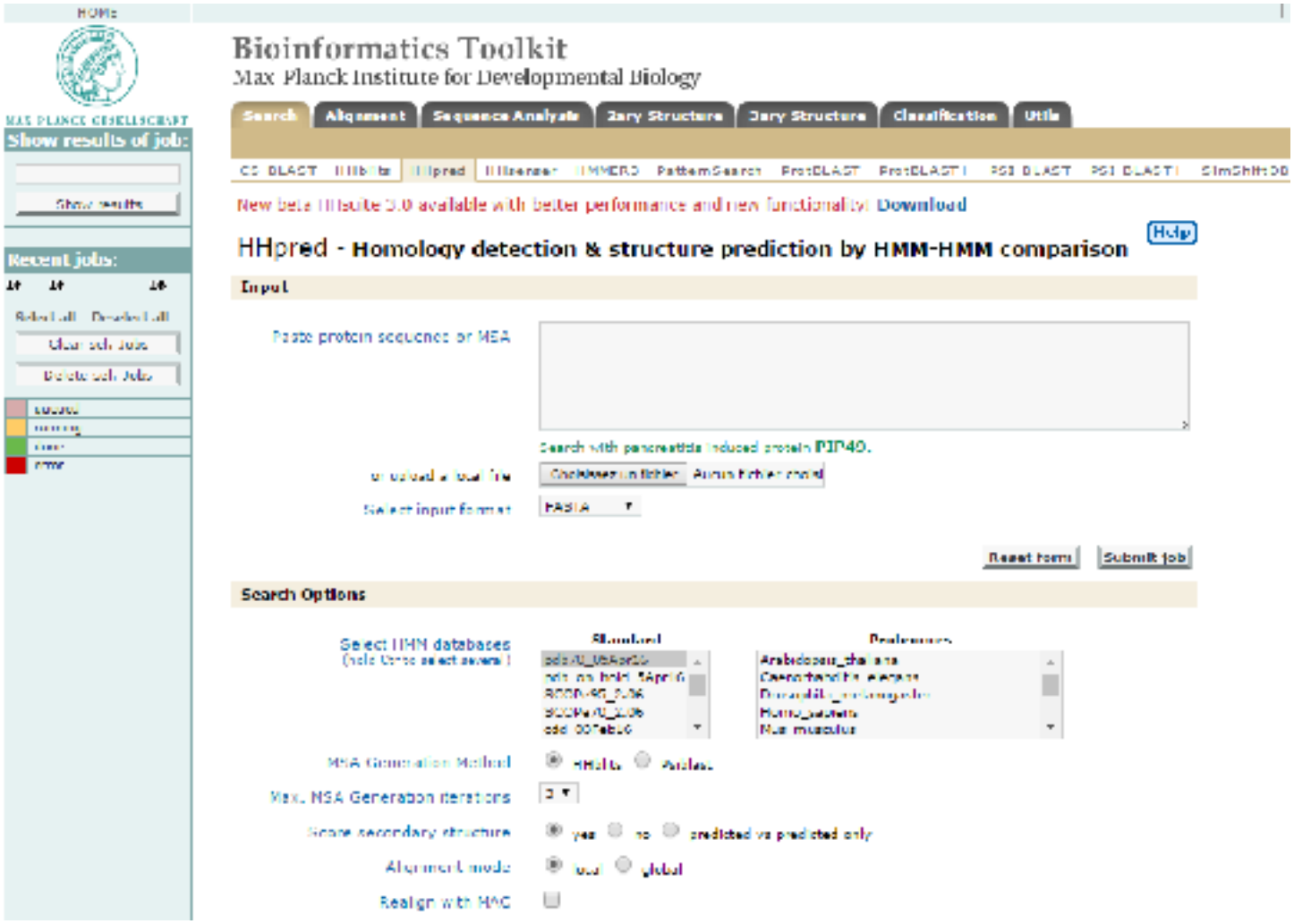
Exemples :

HHpred
Soding J, Bioinformatics, 2005
http://toolkit.tuebingen.mpg.de/hhpred

Très bon serveur pour l'analyse de séquences

http://toolkit.tuebingen.mpg.de

Soding J, Biegert A, Lupas AN. Nucleic Acids Res. (2005) 33(Web Server issue):W244-8



Dissection d'un output HHpred

Switch : vue par profil

Match pdb

%identity Query vs Target

% Probabilité
fiabilité du match
Sécurité → P>90%
Possible → P>50%

Q=Query
ss_pred=prediction
struct. secondaire

T=Target
ss_pred=prediction
struct. secondaire
Ss_dssp=struct. sec.
observes dans pdb

E=Feuillet beta (e=moins fiable)
H=Hélices
C=désordonné

Majuscule > 50% conserve dans la famille
Minuscule = modérémment conservé

Dissection d'un output HHpred

Switch : vue par profil

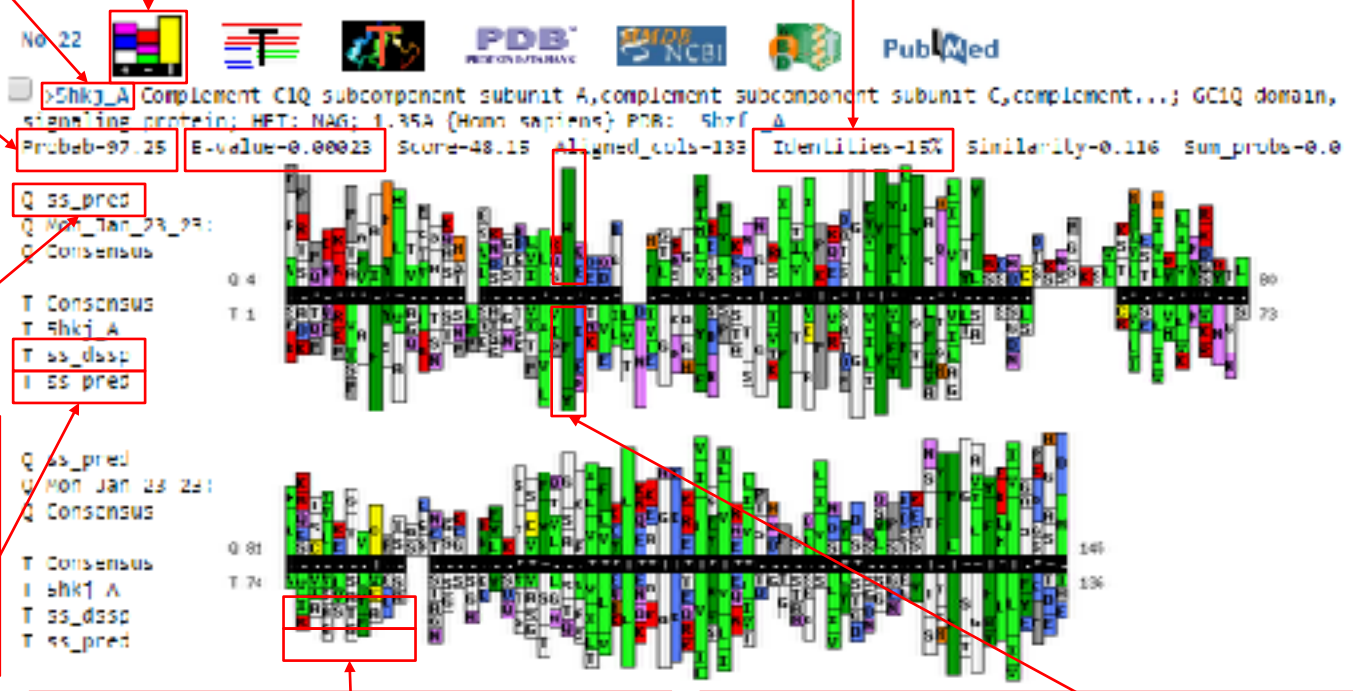
Match pdb

%identity Query vs Target

% Probabilité
fiabilité du match
Sécurité → P>90%
Possible → P>50%

Q=Query
ss_pred=prediction
struct. secondaire

T=Target
ss_pred=prediction
struct. secondaire
Ss_dssp=struct. sec.
observes dans pdb



E=Feuillet beta (e=moins fiable)
H=Hélices
C=désordonné

Majuscule > 50% conservé dans la famille
Minuscule = modérément conservé

Prédiction de Structures Secondaires

Quelques serveurs qui regroupent
un ensemble d'outils de prédiction

Jpred
http://www.compbio.dundee.ac.uk/jpred/

The PredictProtein Server
https://www.predictprotein.org/

PBIL
https://prabi.ibcp.fr/htm/index.php

Considérations pratiques

- La modélisation par homologie permet d'obtenir des **modèles** de très bonne qualité

  1 Å de précision sur la position des atomes correspond à:
  - Structure *X-ray à 2.5 Å de résolution et un facteur R de 25%*
  - *RMN avec 10 contraintes de distances inter-protons par résidus*

- La modélisation par homologie réalisée avec des structures dont les identités de séquence sont >40% identité donne des résultats équivalents.

MAIS

- La qualité des modèles est dépendante de l'alignement de séquence
  – *importance d'un ajustement manuel éventuel de cet alignement*

- L'emploi de templates multiples améliore grandement la qualité des alignements, donc des modèles obtenus.

Exemple hRAP1

>gi|8102033|gb|AAF72711.1|AF262988_1 TRF2-interacting telomeric RAP1 protein [Homo sapiens]
MAEAMDLGKDPNGPTHSSTLFVRDDGSSMSFYVRPSPAKRRLSTLILHGGGTVCRVQEPGAVLLAQPGEA
LAEASGDFISTQYILDCVERNERLELEAYRLGPASAADTGSEAKPGALAEGAAEPEPQRHAGRIAFTDAD
DVAILTYVKENARSPSSVTGNALWKAMEKSSLTQHSWQSLKDRYLKHLRGQEHKYLLGDAPVSPSSQKLK
RKAEEDPEAADSGEPQNKRTPDLPEEEYVKEEIQENEEAVKKMLVEATREFEEVVVDESPPDFEIHITMC
DDDPPTPEEDSETQPDEEEEEEEEKVSQPEVGAAIKIIRQLMEKFNLDLSTVTQAFLKNSGELEATSAFL
ASGQRADGYPIWSRQDDIDLQKDDEDTREALVKKFGAQNVARRIEFRKK

MAEAMDLGKDPNGPTHSSTLFVRDDGSSMSFYVRPSPAKRRLSTLILHGGGTVCRVQEPGAVLLAQPGEA
LAEASGDFISTQYILDCVERNERLELEAYRLGPASAADTGS

GSEAKPGALAEGAAEPEPQRHAGRIAFTDAD
DVAILTYVKENARSPSSVTGNALWKAMEKSSLTQHSWQSLKDRYLKHLRGQEHKYLLGDAPV

PDEEEEEEEEKVSQPEVGAAIKIIRQLMEKFNLDLSTVTQAFLKNSGELEATSAFL
ASGQRADGYPIWSRQDDIDLQKDDEDTREALVKKFGAQNVARRIEFRKK

Exemple hRAP1

MAEAMDLGKDPNGPTHSST : zone d'incertitude
=> ajustement/amélioration des alignements
=> clonages multiples
=> protéolyse ménagée

MAEAMDLGKDPNGPTHSSTLFVRDDGSSMSFYVRPSPAKRRLSTLILHGGGTVCRVQEPGAVLLAQPGEA
LAEASGDFISTQYILDCVERNERLELEAYRLGPASAADTGS

GSEAKPGALAEGAAEPEPQRHAGRIAFTDAD
DVAILTYVKENARSPSSVTGNALWKAMEKSSLTQHSWQSLKDRYLKHLRGQEHKYLLGDAPV

PDEEEEEEEEKVSQPEVGAAIKIIRQLMEKFNLDLSTVTQAFLKNSGELEATSAFL
ASGQRADGYPIWSRQDDIDLQKDDEDTREALVKKFGAQNVARRIEFRKK

Et après la structure ?

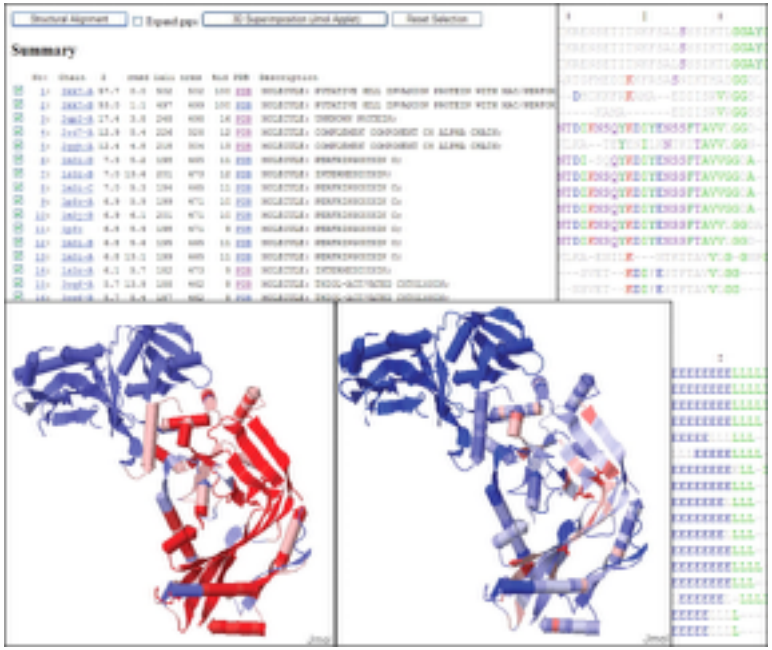
Comparaison 3D: Server DALI

Recherche des homologues tridimensionnels

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

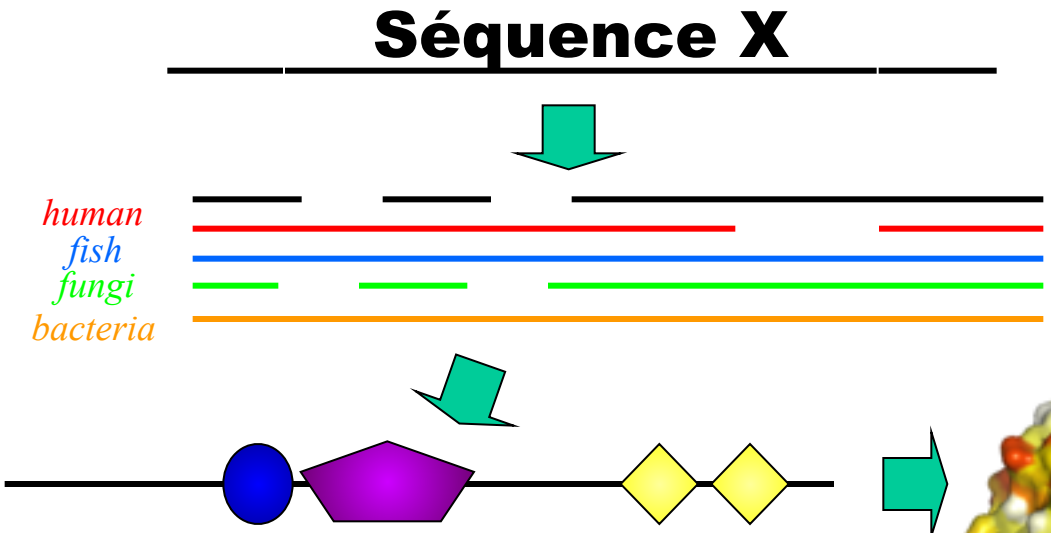
Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G (1992) A database of protein structure families with similar folding motifs. Protein Science 1:1691-1698

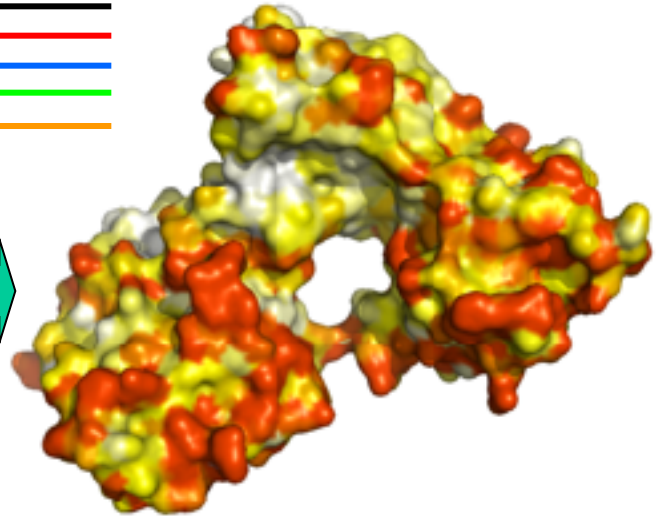
A partir d'une structure 3D, permet de
- rechercher des repliements similaires
- retrouver des fonctions associées



From: Dali server: conservation mapping in 3D
Nucleic Acids Res. 2010;38(suppl_2):W545-W549. doi:10.1093/nar/gkq366

Quelles informations peut-on extraire d'une analyse de séquence ?





1 - Ma séquence est-elle analogue à une séquence connue ?

2 - Puis-je effectuer des inférences fonctionnelles ?

3 - Comment découper ma séquence en modules fonctionnels repliés?
- Quelles sont les régions fonctionnelles et structurales ?

4 - Peut-on prédire son organisation structurale ?

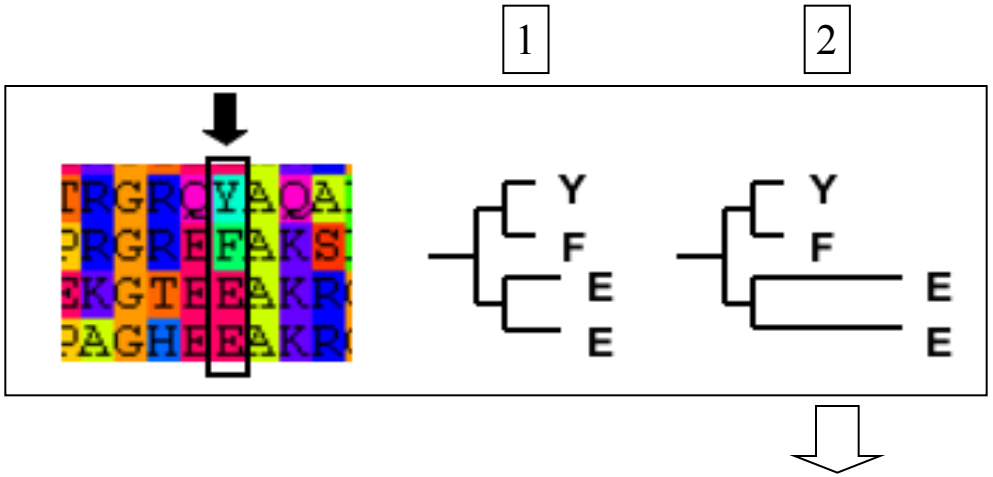
5 - Quelles sont les régions soumises à des pressions de sélection particulières ?

Analyse des conservations automatique
Une des meilleures sensibilités : ConSurf

http://consurf.tau.ac.il/

N. Ben Tal, Nucleic Acids Research, 2005, Vol. 33, Web Server issue W299–W302

Méthode dont la sensibilité est accrue par une prise en compte des vitesses d'évolution à chaque position
(cf publi Rate4site Pupko et al Bioinfo 2002)

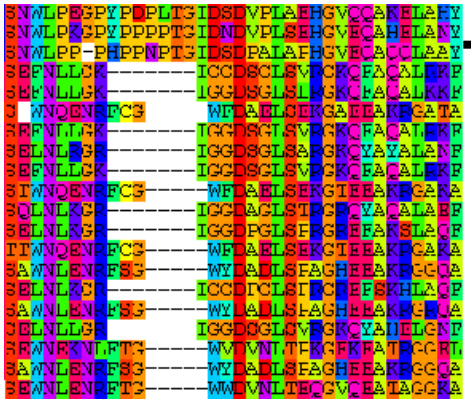


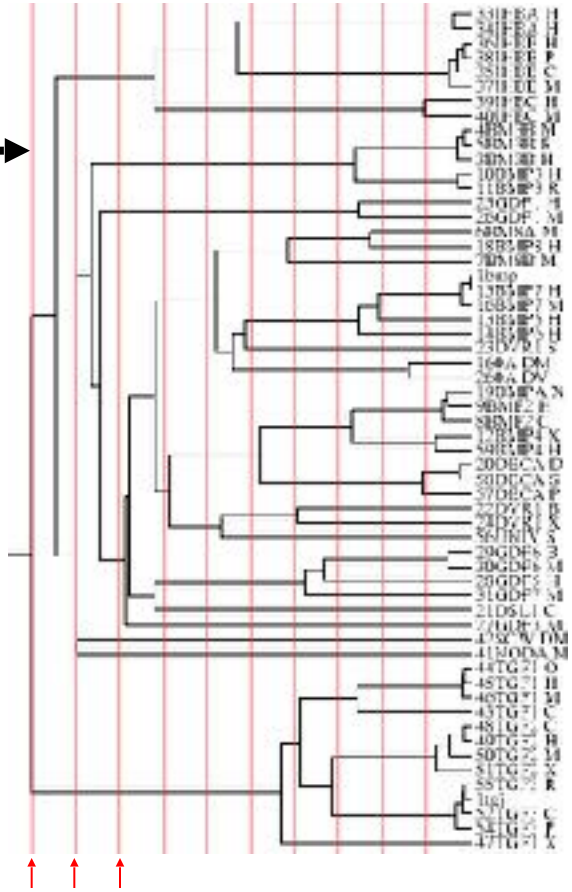
Pression de sélection plus importante sur le « E » dans le cas 2

Score de conservation dans la colonne des facteurs B du fichier pdb

Analyse des traces évolutives

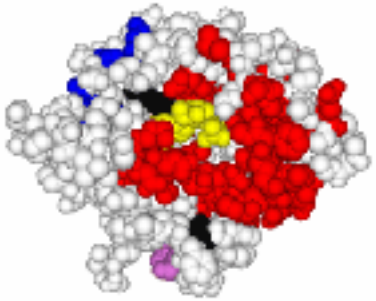
http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/
http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html





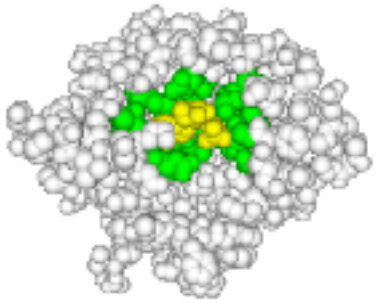
Analyse de l'arbre phylogénétique
par partition de classes

« Evolutionnary Traces »



Patchs colorés
= Classes d'acides aminés

Epitope fonctionnel



Yao, H., …and O. Lichtarge. 2003.
J Mol Biol. 326:255-261.

Principe de la CoEvolution d'Interface :

Analyse Statistique → Extraction des caractéristiques→Score de Docking

Database of Interfaces with Evolutionary information

InterEvol:

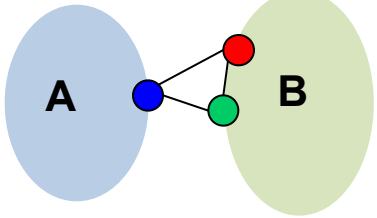
~18,000 non-redundant interfaces
among which
~4,000 heteromeric interfaces

G. Faure, et al Nucleic Acids Res. (2012)

http://biodev.cea.fr/interevol

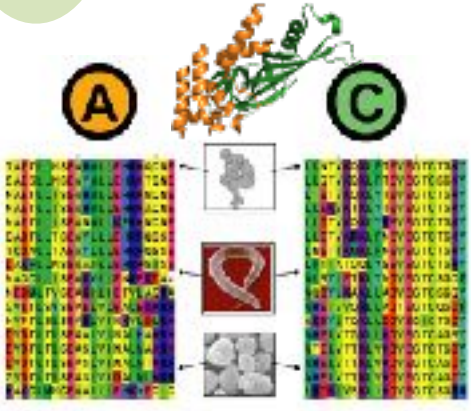
InterEvScore : Discriminate co-evolved interfaces

Contacts multi-domaines



Patches apolaires

Information d'évolution

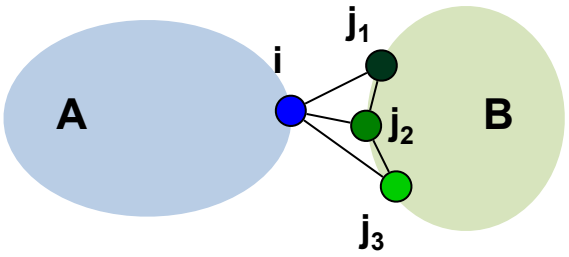


J. Andreani, et al , Plos Comp Biol (2012) ; Bioinformatics (2013)

Development of InterEvScore, a docking score taking evolution into account

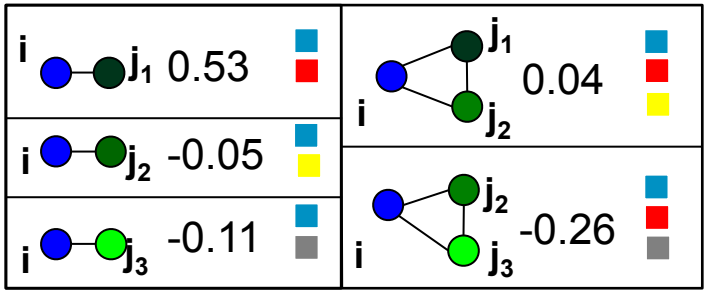
Score inter-molecular interface contacts

For every possible couple or triad

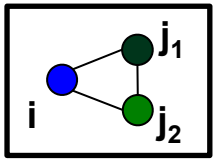


Find the most favorable environment …

Contact propensities derived from InterEvol *(statistics on 1,289 interfaces)*

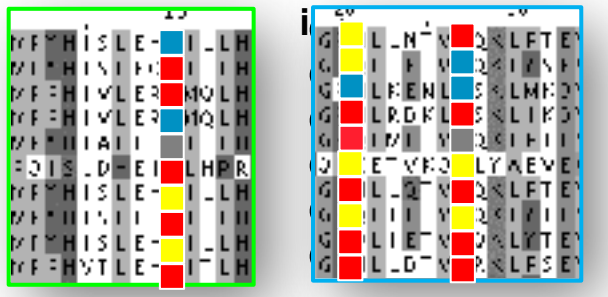


…taking into account evolution



Alignment for A

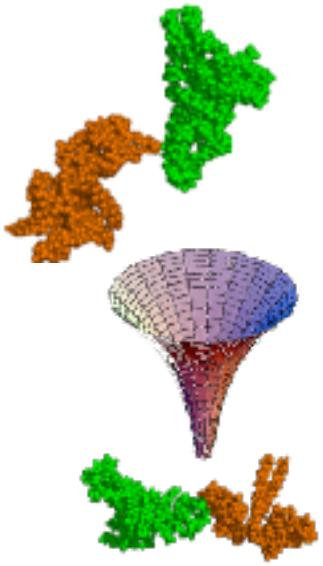
Alignment for B

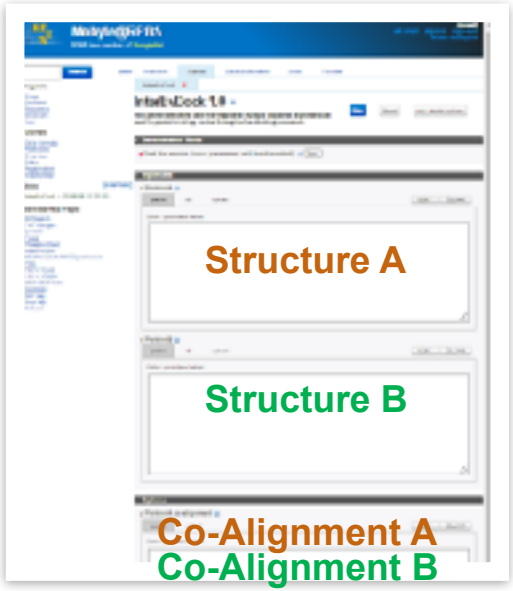


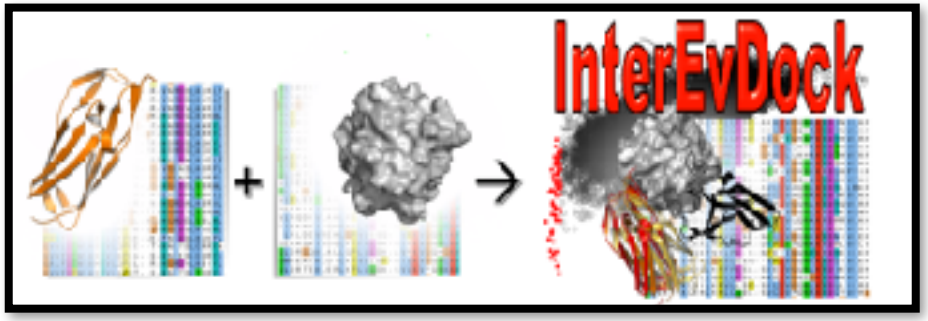
H. sapiens
M. musculus
D. rerio
⋮
S. cerevisiae

> 10 sequences

InterEvDock server to account for Co-evolution Information in Docking







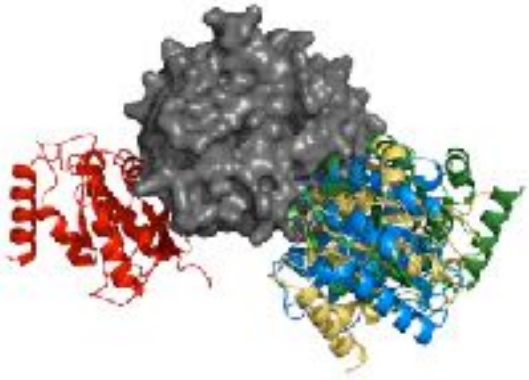
Returns 10 models (~1h) selected from
a Consensus among the top solutions of :

- **InterEvScore** (residue based / coevolution),
- **SOAP-PP** (statistical atomic based) (*A. Sali's lab*),
- **FRODOCK** (rigid-body + physics based) (*P. Chacon's lab*)

J. Yu et al, Nucleic Acids Research (2016)

Running at RPBS server *(coll. P. Tufféry)*
http://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock/

InterEvScore success rates for individual test cases

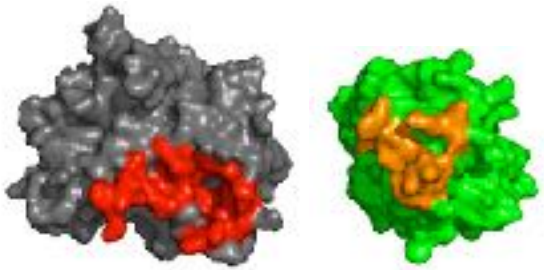


- Can we predict the **binding mode**?

With 10 models per test case

46 % success

(25 successful cases out of 54 test cases)



- Can we predict **binding sites**?

With 10 models per test case:

65 % success (35 cases) for both partners

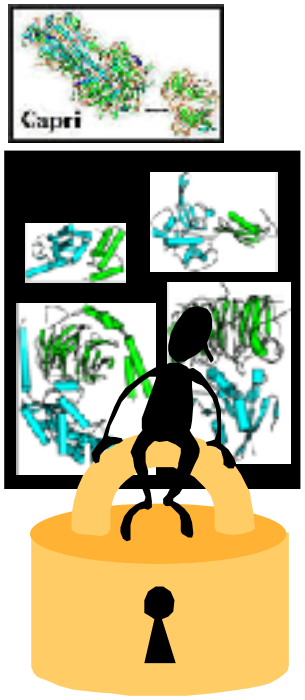
91 % success (49 cases) for at least one partner

out of 54 test cases

Running at the Ressource Parisienne en Bioinformatique Structurale *(Yu et al NAR 2016)*

http://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock/

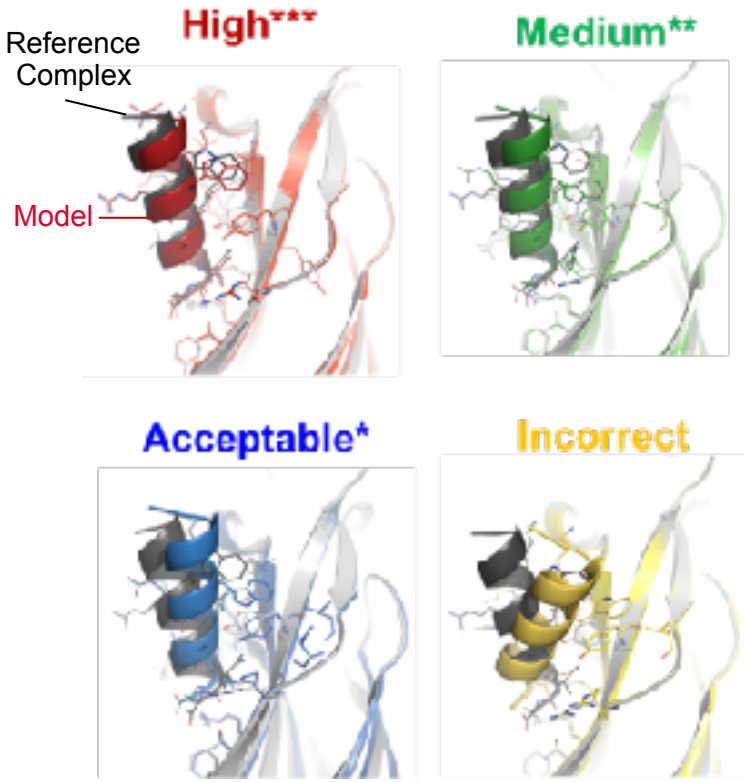
THE « CAPRI » INTERNATIONAL CHALLENGE : CRITICAL ASSESSMENT OF PREDICTION OF INTERACTIONS



- **Blind test for the different predictive strategies**

  About every six months structures of individual binding partners are released to the community.

  → X-ray structure of every query complex is kept secret.



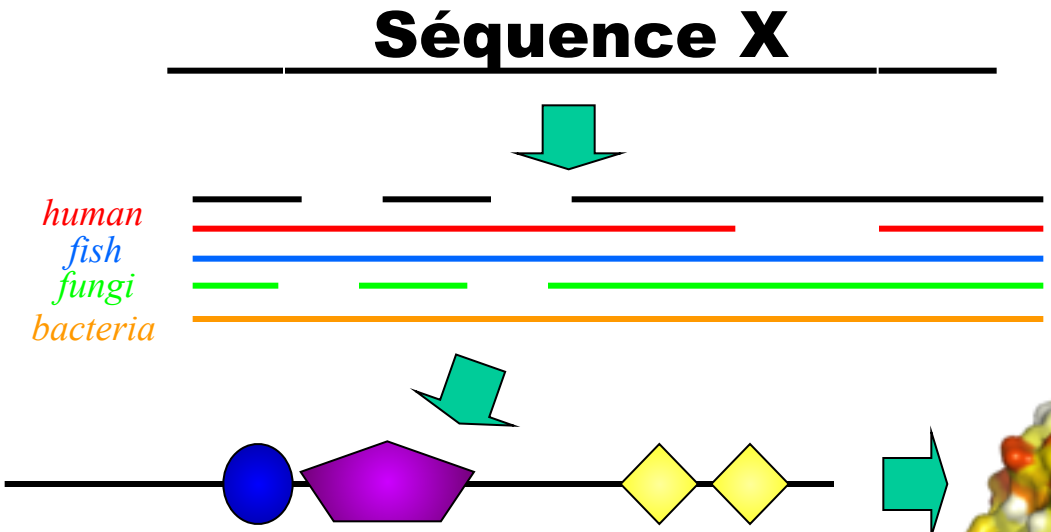
THE THREE MOST RECENT CAPRI EVALUATION MEETINGS.

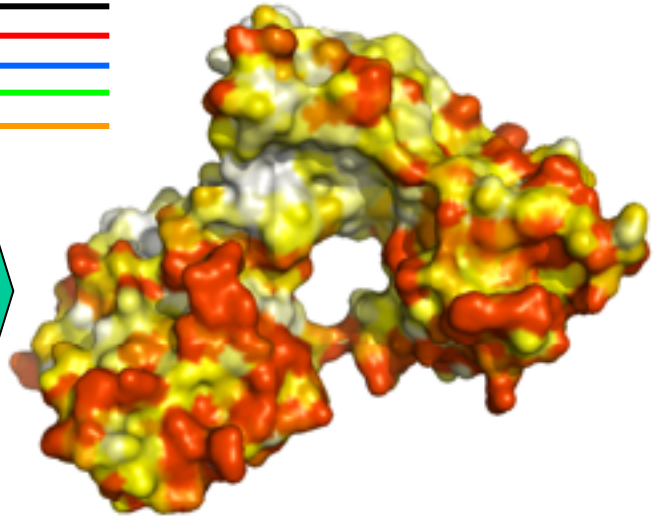
	CAPRI evaluation meeting, year and targets					
	2009, 12 targets		2013, 14 targets		2016, 20 targets	
Rank	Group	Success	Group	Success	Group	Success
1	Vajda/Kozakov	6/4***/2**	Bonvin	9/1***/3**	**Guerois**	**10/1***/8****
2	Zacharias	6/4***/1**	Bates	8/2**	Zacharias	10/3***/2**
3	Zou	6/3***/2**	Vakser	7/1***	ClusPro	9/3**
4	Eisenstein	6/3***/1**	Kozakov/ Vajda	6/2***/3**	Kozakov/ Vajda	8/3***/2**
5	Wolfson	6/3***/1**	Shen	6/1***/3**	Seok	8/3***/2**
6	Weng	6/2***/2**	Fernandez-Recio	6/1***/3**	Fernandez-Recio	7/1***/3**
7	Zhou	6/2***/2**	ClusPro	6/4**	Zou	7/1***/2**
8	Bonvin	6/1***/4**	Zou	6/1***/2**	Weng	6/1***/4**
9	ClusPro	5/1***/3**	Zacharias	6/1***	Vakser	6/2***/2**
10	Fernandez-Recio	5/2**	Eisenstein	5/1***/2**	Bates	6/3**

Kozakov D et al (2017) Nat. Protocols

Et en absence de structure ?

Quelles informations peut-on extraire d'une analyse de séquence ?





1 - Ma séquence est-elle analogue à une séquence connue ?

2 - Puis-je effectuer des inférences fonctionnelles ?

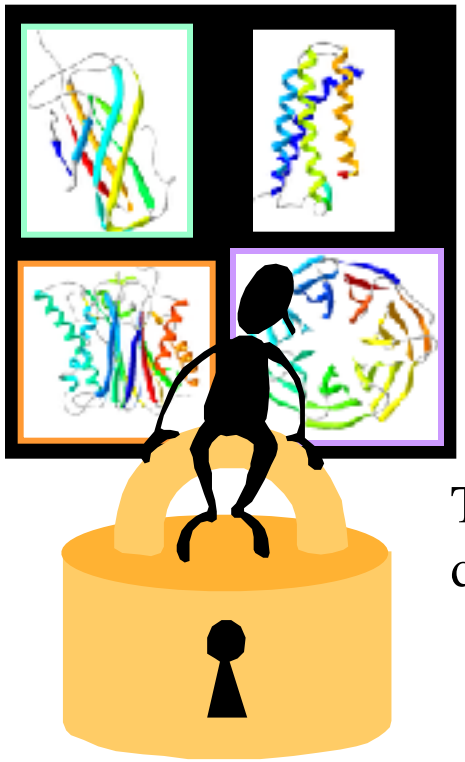
3 - Comment découper ma séquence en modules fonctionnels repliés?
- Quelles sont les régions fonctionnelles et structurales ?

4 - Peut-on prédire son organisation structurale ?

5 - Quelles sont les régions soumises à des pressions de sélection particulières ?

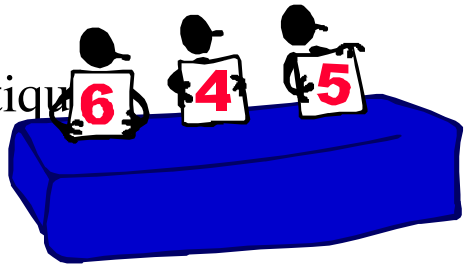
Comment évaluer les meilleures méthodes : CASPn

http://predictioncenter.org/

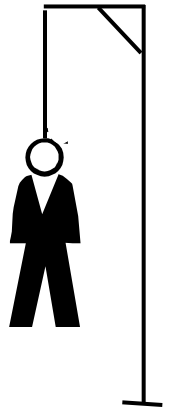


Tous les deux ans, un ensemble de nouvelles structures sont gardées secrètes durant la prédiction.

Classement, analyse critique des points forts et faibles de chaque méthode.

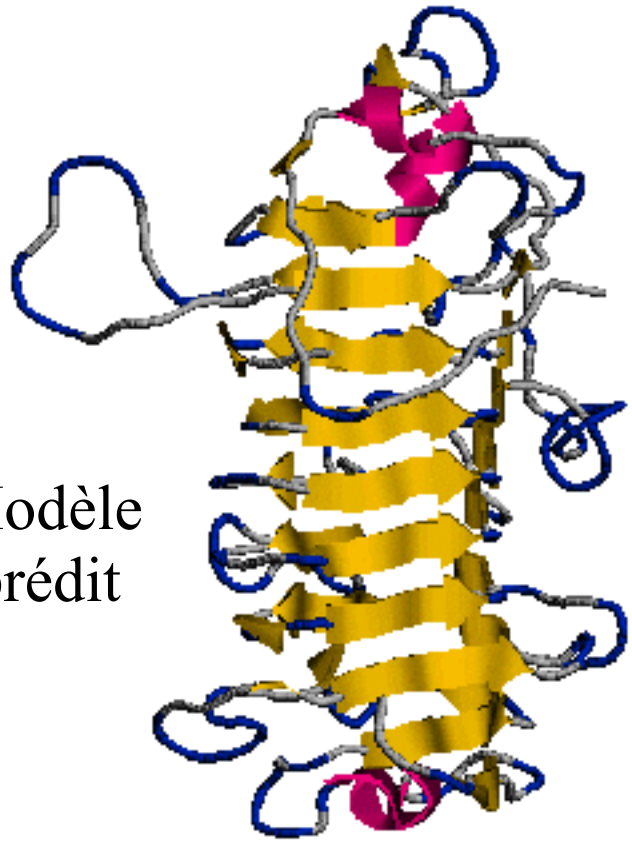


Test en aveugle des différents algorithmes



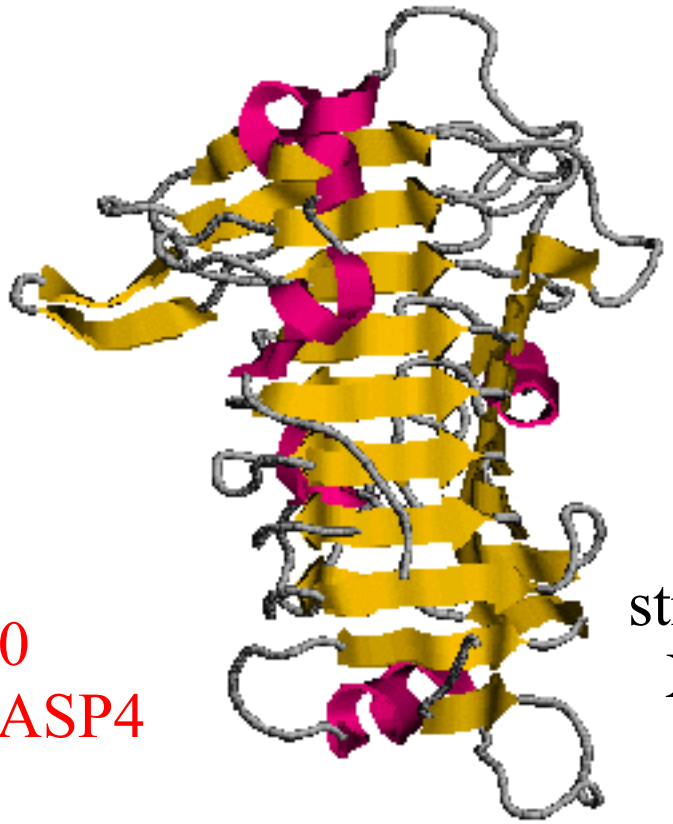
Exemple de prédiction

Pas d'homologie de séquence évidente…



Modèle prédit

cible T0100 De CASP4



structure X-ray

Résultats CASP11 (2014):

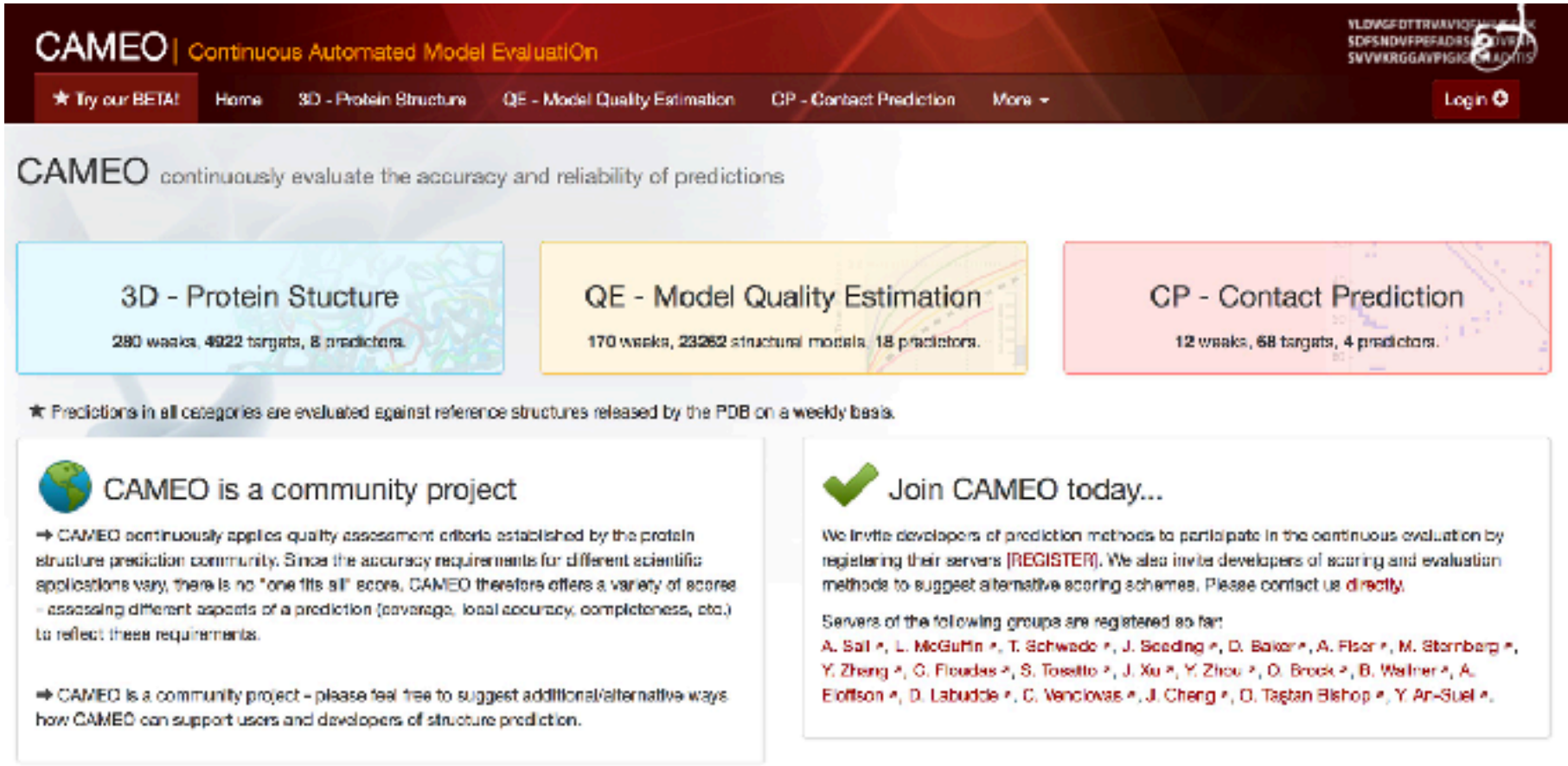
http://predictioncenter.org/casp11/results.cgi

Best servers for template-based modelling

- {Zhang-Server, QUARK} (**I-TASSER**)
- **nns** (Lee group)
- **RosettaServer** – gets much better with best models
- MyProteinMe (Skwark group)

- **HHpred** ranked best in CASP9 (2010) and has been integrated into other servers/approaches

Controle continu avec CAMEO



http://cameo3d.org/

Analyse des séquences protéiques en absence d'homologues détectables

- ***Analyse des peptides signaux***
  G. von Heijne SignalP http://www.cbs.dtu.dk/services/SignalP
- ***Low complexity***
- ***Domaines Coiled-coils***
  A. Lupas COILS (http://www.ch.embnet.org/software/COILS_form.html)
  B. Berger Multicoil (http://groups.csail.mit.edu/cb/multicoil/cgi-bin/multicoil.cgi)
  M. Delorenzi Marcoil (http://bcf.isb-sib.ch/webmarcoil/webmarcoilC1.html)

  Analyses combinées dans SMART : http://smart.embl-heidelberg.de/
  Nombreux outils regroupés dans le Bioinformatics Toolkit
  http://toolkit.tuebingen.mpg.de/

- ***Segments transmembranaires***
  cf diapo suivante
- ***Structures Secondaires***
  cf diapos suivantes
- **Reconnaissance de repliement Threading**

Sources d'inspiration pour le cours et pour aller plus loin

- **Sur les alignements et l'évolution structurale des protéines**
  http://www.people.virginia.edu/~wrp/papers/ismb2000.pdf

- **Très bon chapitre d'ouvrage par Koonin accessible on-line**
  http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=sef.chapter.166

MERCI MERCI MERCI:

Jessica Andreani : jessica.andreani@cea.fr
Raphaël Guerois : guerois@cea.fr
CEA Saclay, Institut Joliot
91190 Gif sur Yvette Cedex

RéNaFoBiS

Grand tournoi bb-foot.

Regardez cet air combatif !!!