Les outils de bioinformatique (du web) pour la biologie structurale

Analyse des séquences et des structures

Claudine MAYER
Université Paris Diderot
Institut Pasteur

université PARIS DIDEROT PARIS 7
USPC Université Sorbonne Paris Cité
Institut Pasteur
Unité de Microbiologie Structurale

Introduction : qu'est-ce qu'une structure ?

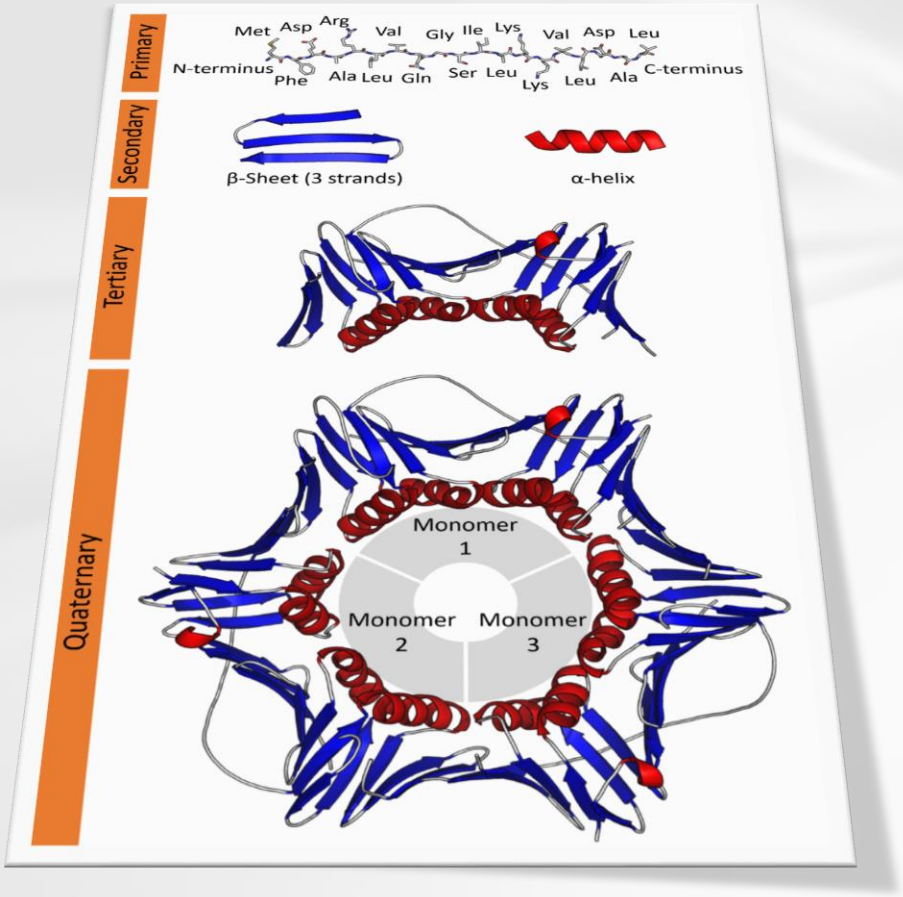
Informations données par la séquence

Prédiction de la structure

Informations données par la structure

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

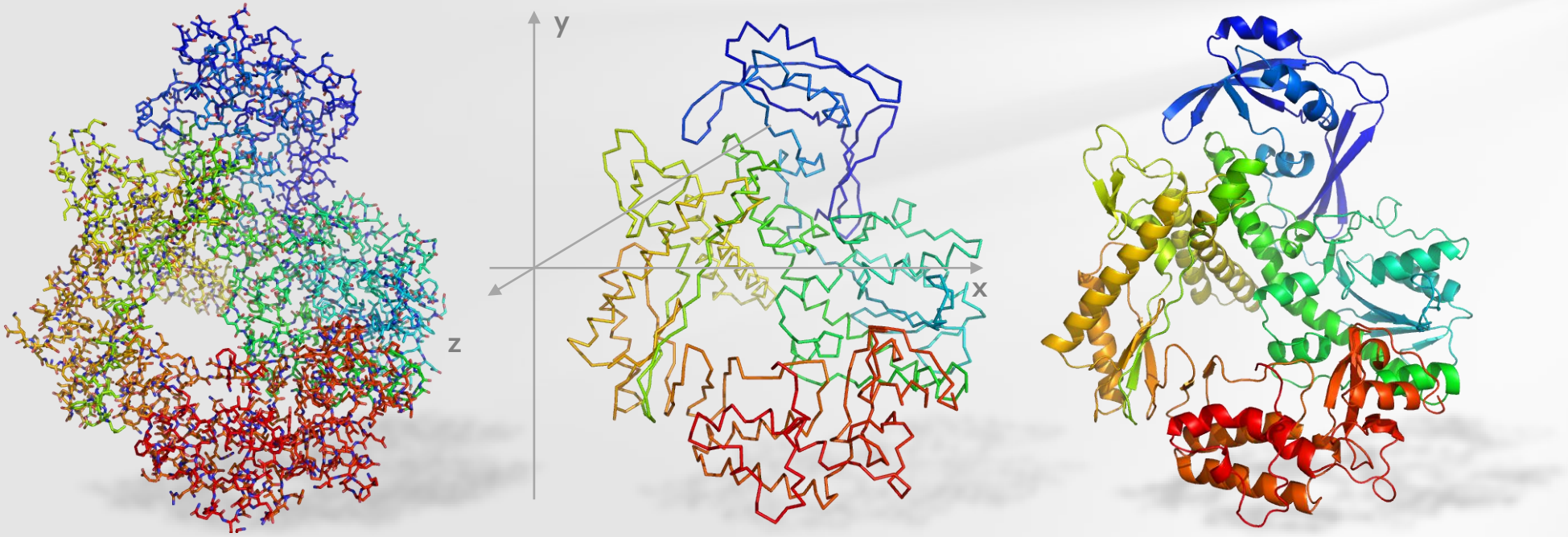
- Séquence (1D)
- Structures secondaires (2D)
- Agencement spatial (3D)
- Assemblages moléculaires (4D)
- Dynamique (5D)



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

A 3-dimensional description of all atoms

Coordinates (x,y,z) of all atoms of the aminoacids that compose the protein
The quality of the structure depends on the resolution of the experimental data provided by the method used to obtain it (X-ray, NMR, cryo-EM)



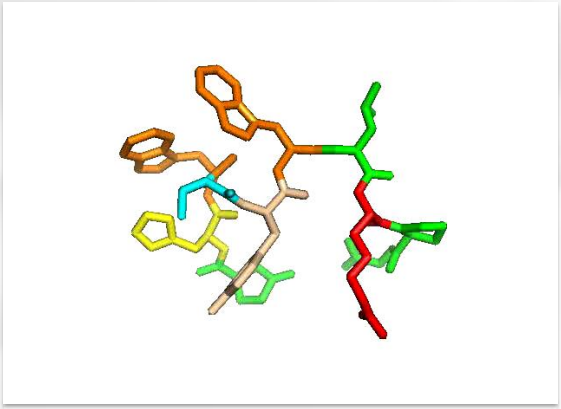
Qu'est-ce que la structure d'une macromolécule ?

Pour reprendre la définition de Marie-Hélène LeDu

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

***Definition** : the arrangement of separate molecules, such as in protein-protein or protein-nucleic acid interactions*

- *Interaction with the environment and with partners within the cell*

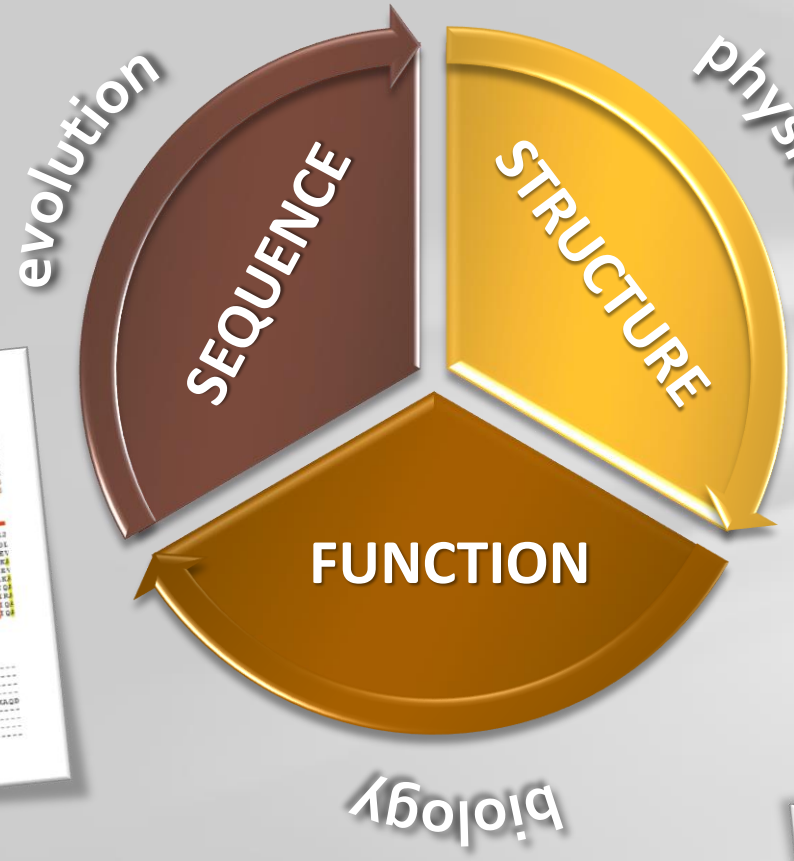


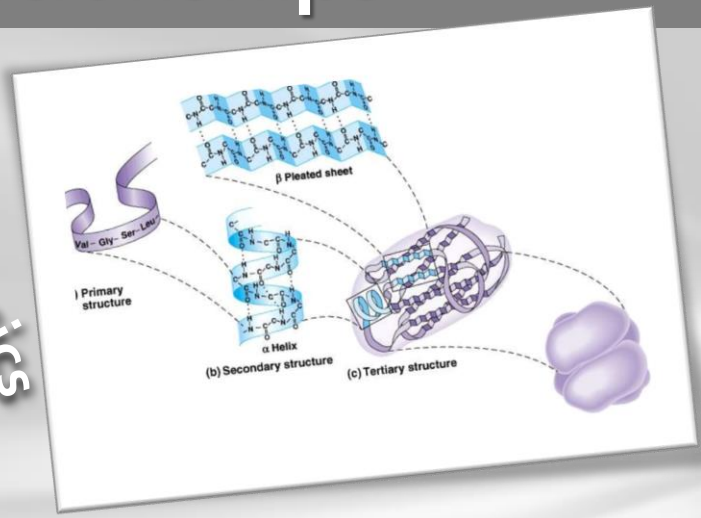
Valéry et al., Nat. Comm, 2015

Atomic view of the histidine environment stabilizing higher-pH conformations of pH-dependent proteins

Sequence-structure-function relationships

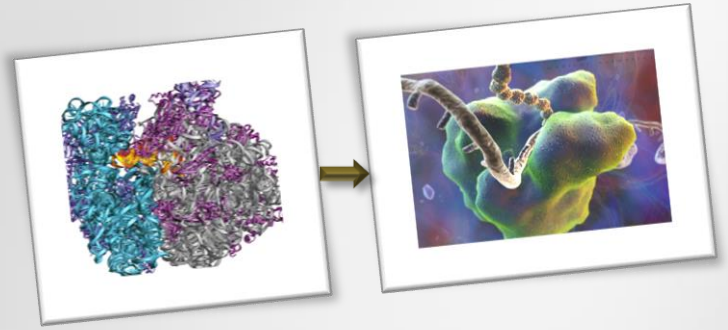
How the sequences evolve based on the phylogenetic distribution





Structure determination based on physical methods Sequence determines structure

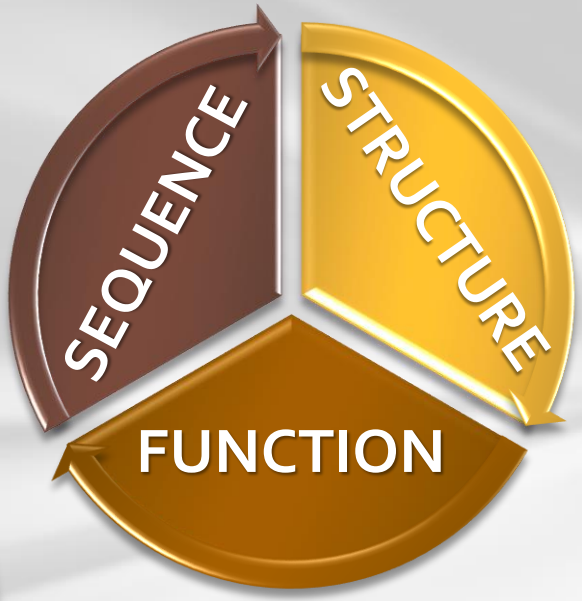
Understand the biological questions in a 3D perspective



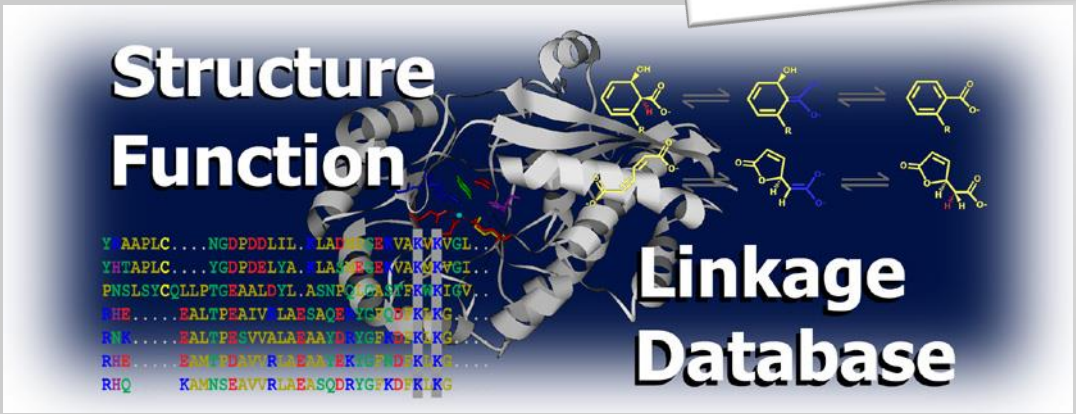
Sequence determines structure

The aminoacid sequence of a protein determines its three-dimensional structure

Proteins that share sequence identity > 25 % are structurally similar



Homology modeling

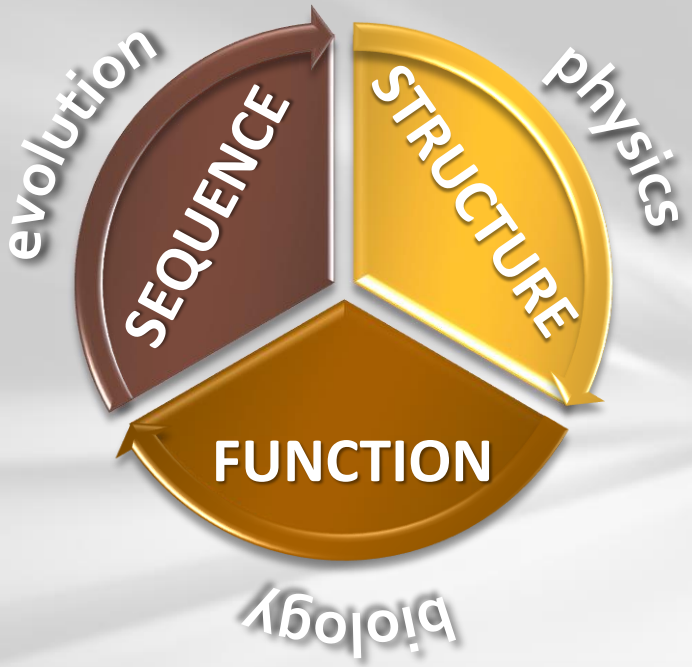


… And structure is highly correlated to function

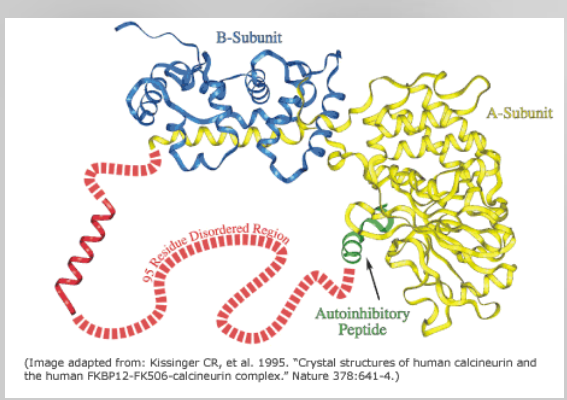
http://sfld.rbvi.ucsf.edu/django/

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

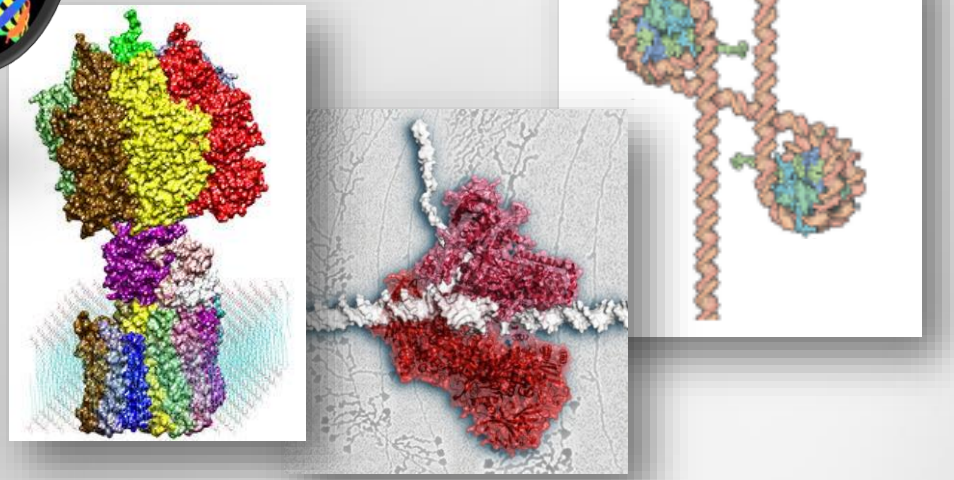
The 3D structure of a protein defines not only its size and its shape but also its function



Exception
IDPs & IDRs

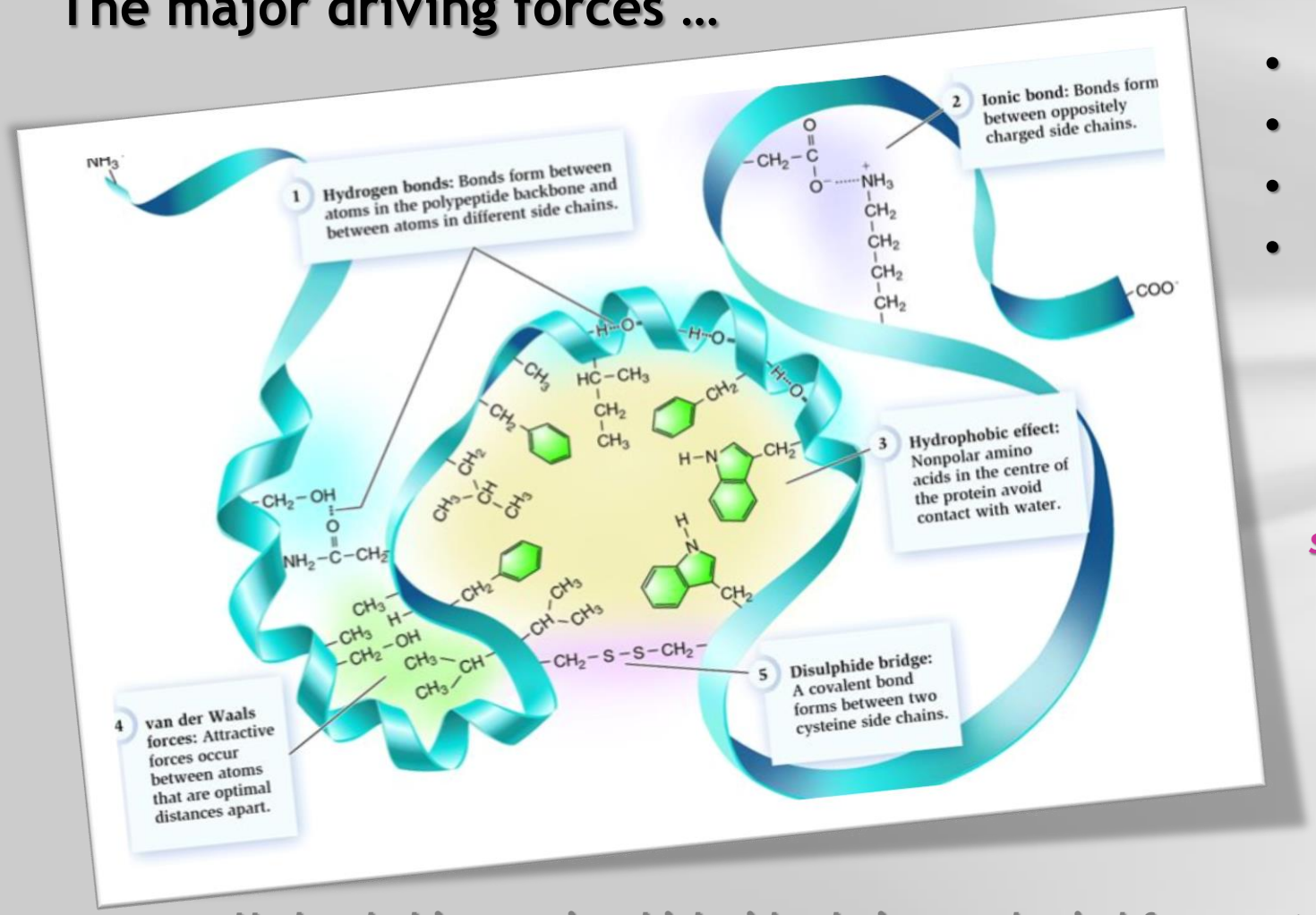


(Image adapted from: Kissinger CR, et al. 1995. "Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex." Nature 378:641-4.)



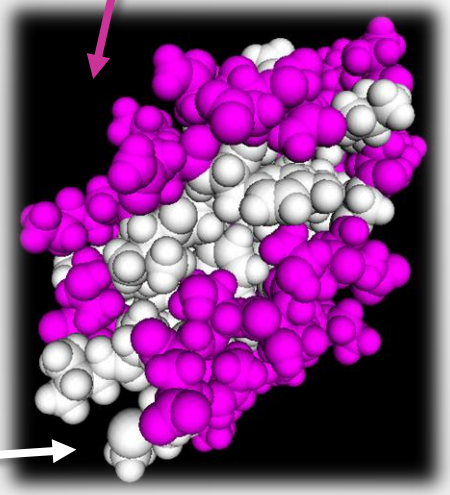
Factors determining protein folding

The major driving forces …



- Hydrophobic effect
- H-bonds
- Conformational entropy
- Ionic interactions…

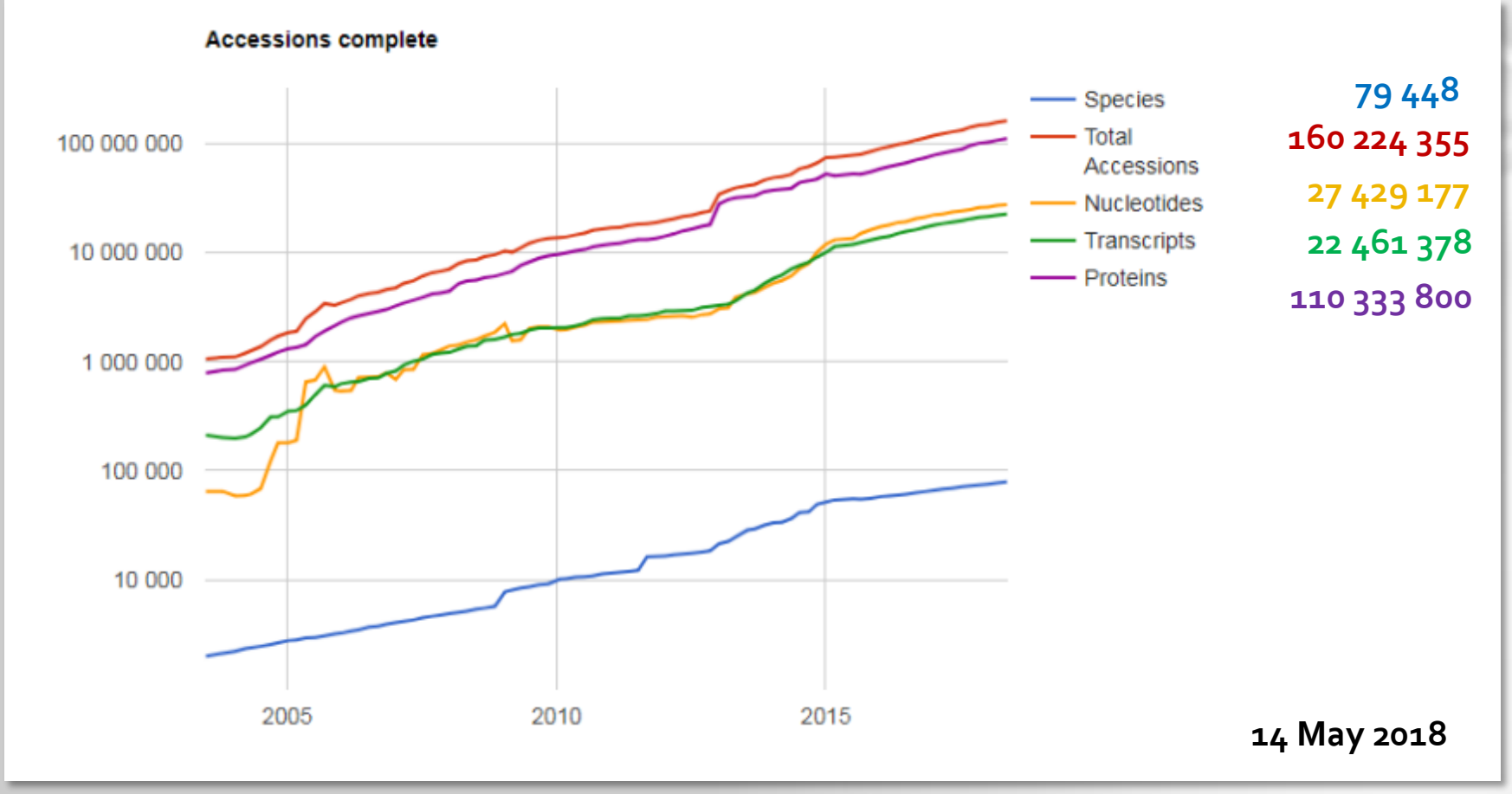
Charged and polar side chains are situated on the solvent-exposed surface where they interact with surrounding water molecules



Hydrophobic core in which side chains are buried from water
Minimizing the number of hydrophobic side chains exposed to water is the principal driving force behind the folding process

Available sequence data in RefSeq

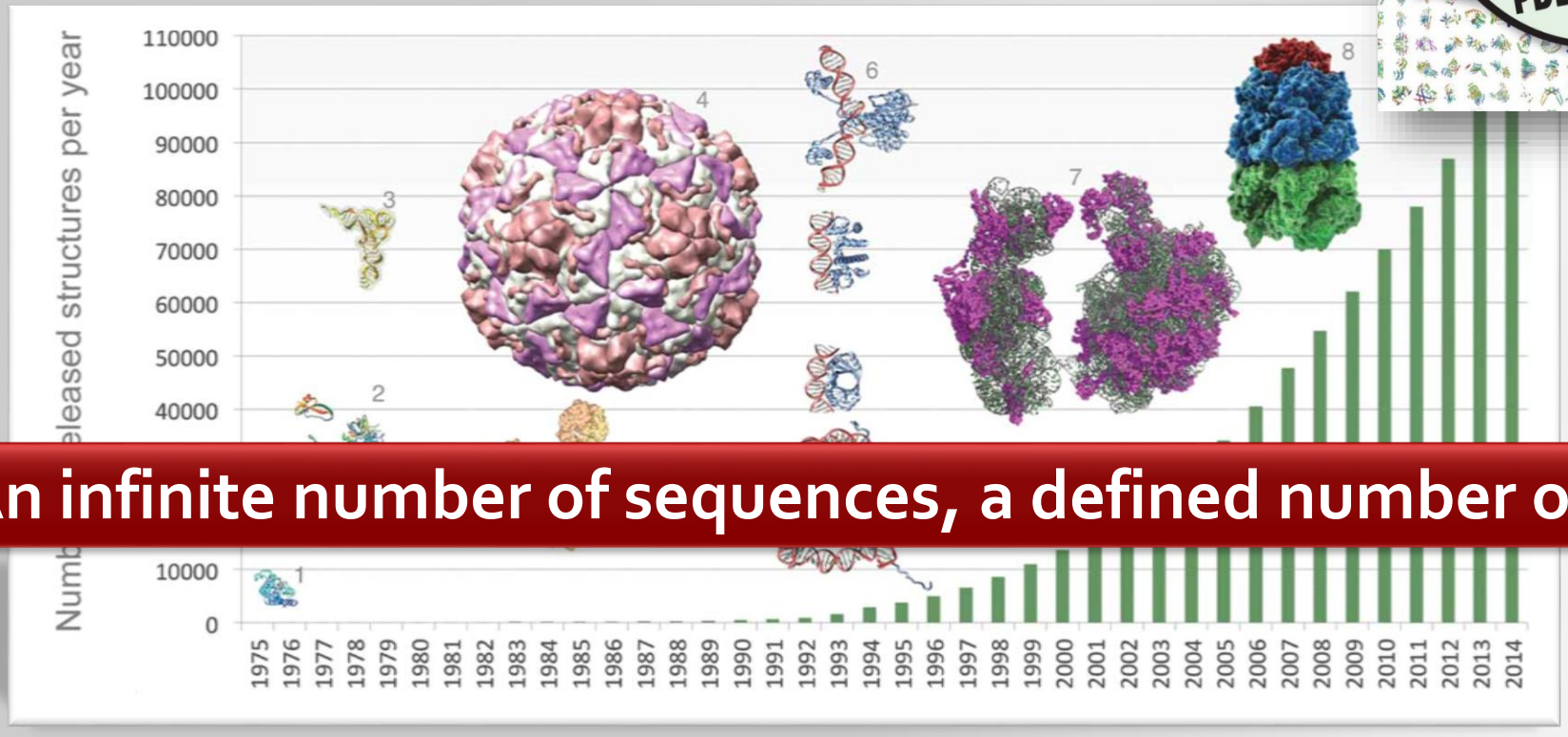
The Big Data era: The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, *non-redundant*, *well-annotated* set of sequences including genomic DNA, transcripts, and proteins



Available structural data in the PDB

140 591 structures in the PDB : most of the protein can be structurally studied e.g. the structure exists or it can be modelled

RCSB PDB
PROTEIN DATA BANK

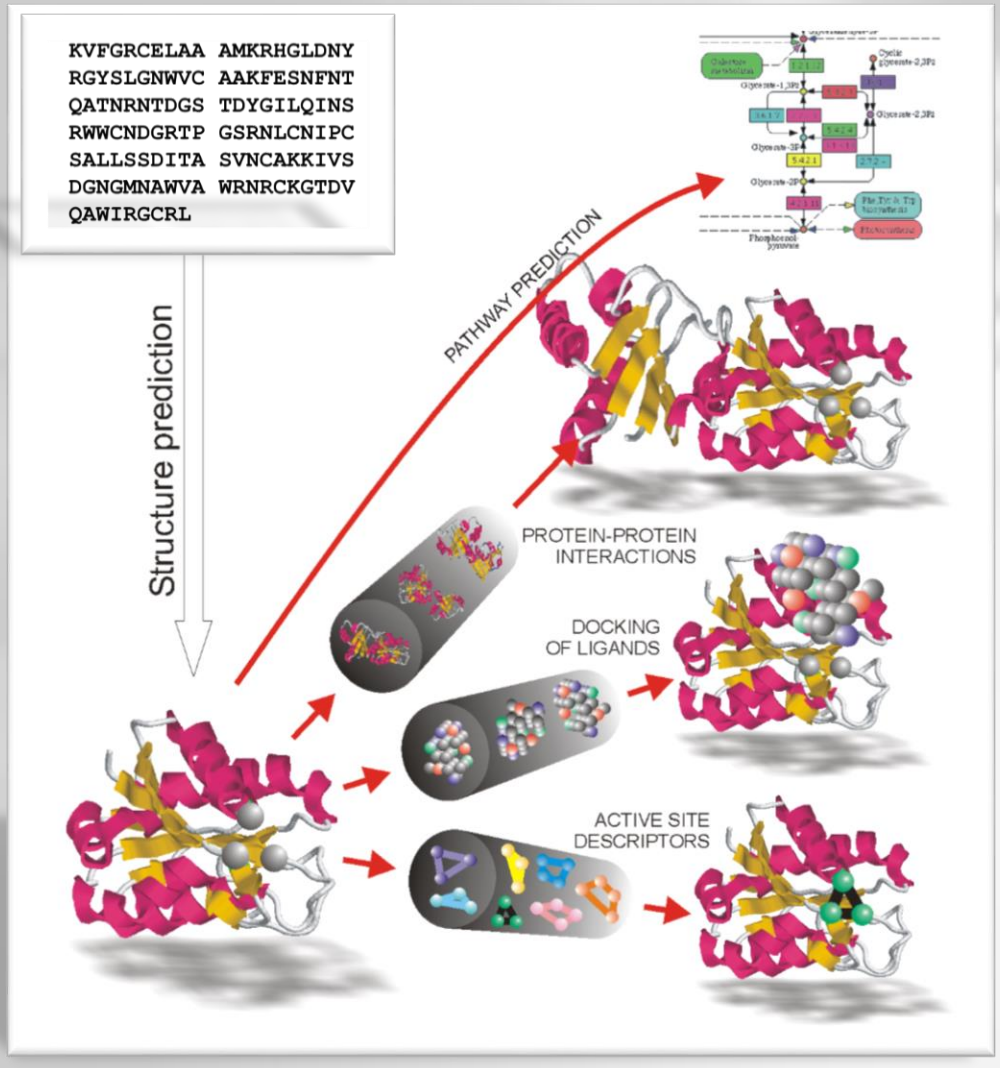


An infinite number of sequences, a defined number of folds

How to predict a structure

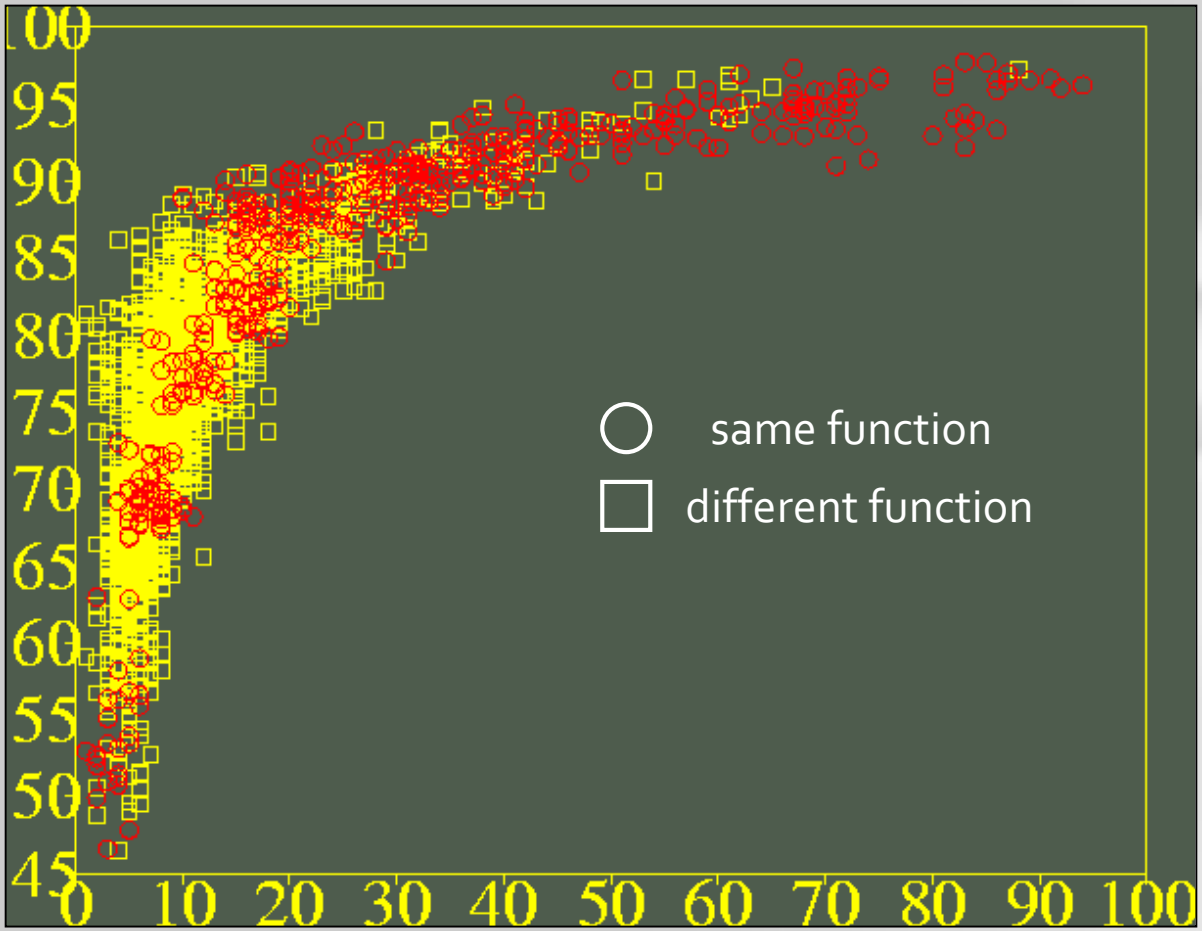
Why predicting a structure?

In a first approach, depending on the biological question, numerous sequences can be structurally modelled



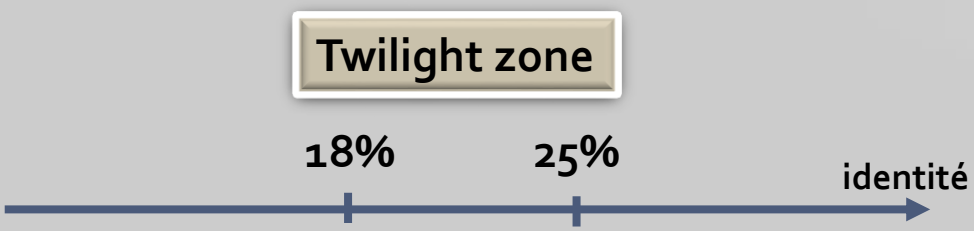
How to predict a structure

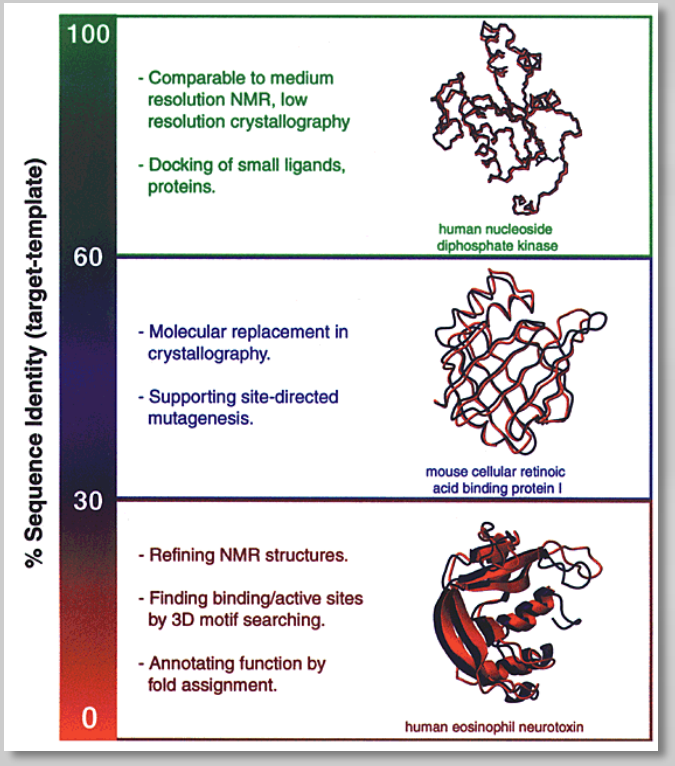
Before predicting the structure



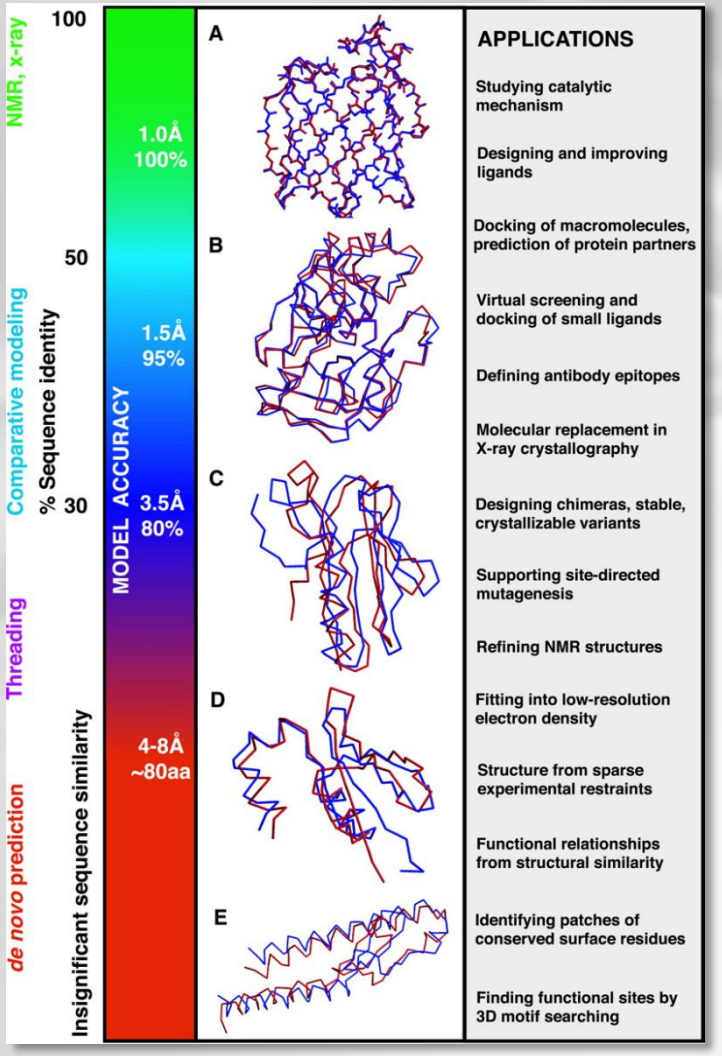
Proteins that share sequence identity are structurally similar

How to predict a structure





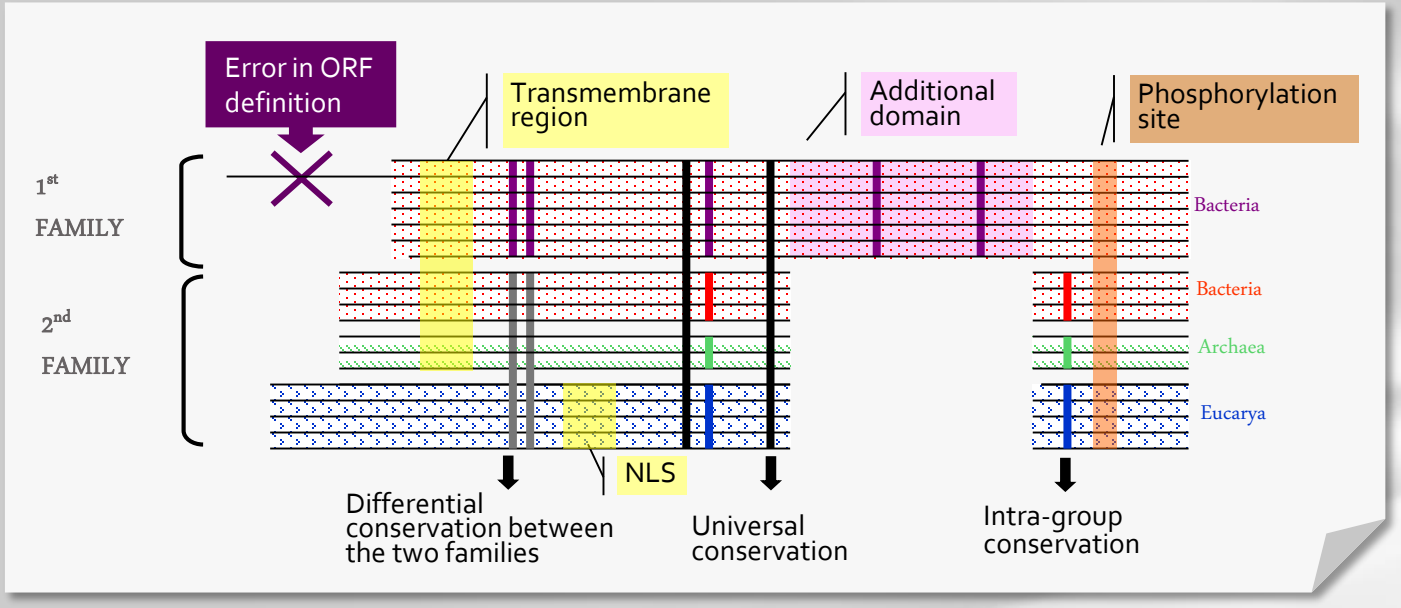
Sali, A. & Kuriyan, J. *Trends Biochem. Sci.* 22, M20–M24 (1999)



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

Multiple alignment of complete sequences

MACS



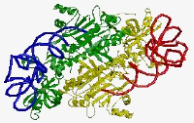
domain organization, structural motifs
key functional residues, ORF definition
localization signals, conservation pattern, ...

Lecompte *et al* Gene. 270, 17-30 (2001)

Functional genomics

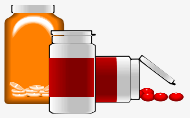
Evolutionary studies

Structure modeling



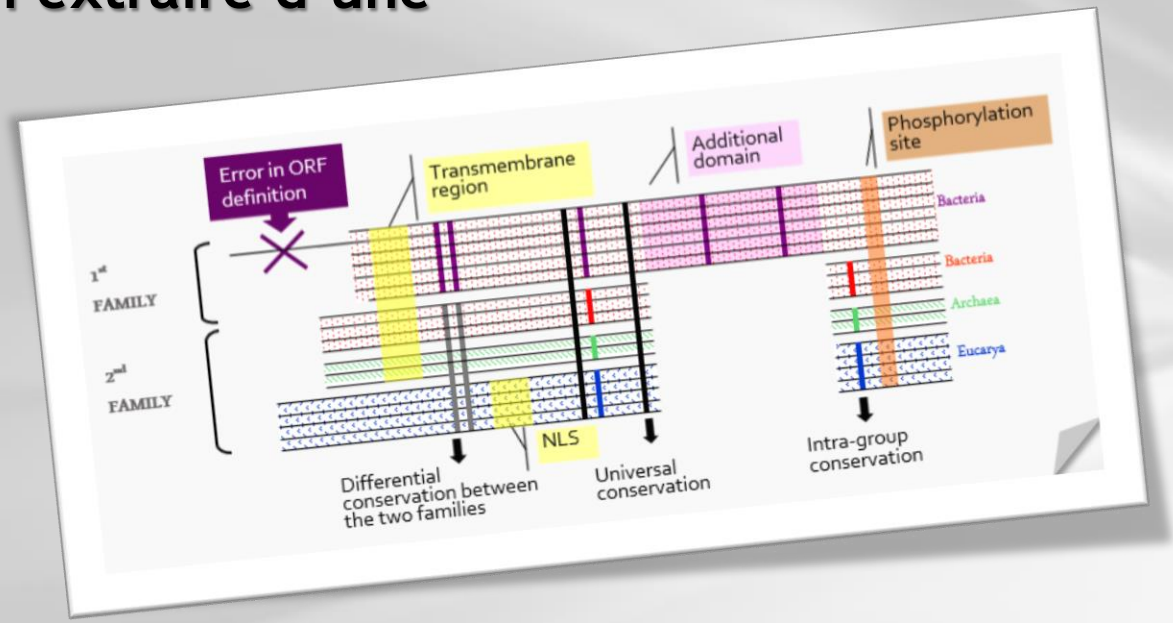
Mutagenesis experiments

Drug design



Multiple alignment of complete sequence

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



1 – Homologie de séquences

2 – Inférences fonctionnelles

3 – Organisation en modules/régions fonctionnels et/ou structurés

4 – Architecture tridimensionnelle

5 – Pressions de sélection particulières

6 – Histoire évolutive

Introduction : qu'est-ce qu'une structure ?

Informations données par la séquence

Prédiction de la structure

Informations données par la structure

Sequence retrieval

The principles

- **Recherche textuelle** *Data Mining*
  - PubMed
  - NCBI, Uniprot, RefSeq, EBI
  - Banques spécialisées

(Abstract, Keywords, etc…)

- **Recherche par séquence appât (ou consensus)**
  - Programmes de type Fasta, Blast, Psi-Blast, Ballast, Profile
    - banques nucléotidiques
    - banques peptidiques
    - banque séquence primaires des structures

- **Recherche par structure**
  - Programmes de superposition, modélisation

(type VAST, Vector Alignment Search Tool)

Sequence retrieval

Main servers

NCBI	https://www.ncbi.nlm.nih.gov/
UNIPROT	http://www.uniprot.org/
RefSeq	https://www.ncbi.nlm.nih.gov/refseq/
EBI	https://www.ebi.ac.uk/
BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
VAST	https://omictools.com/vector-alignment-search-tool-tool

An expert system

ExPASy Bioinformatics Resource Portal

https://www.expasy.org/

Basics of protein structure

Compute molecular weight and isoelectric point

http://molbiol-tools.ca/Protein_Chemistry.htm

Amino acid composition, mass & pI

Amino acid composition & Mass – ProtParam tool *(ExPASy, Switzerland)*

Isoelectric Point - Compute pI/Mw tool *(ExPASy, Switzerland)*. If you want a plot of the relationship between charge and pH use ProteinChemist *(ProteinChemist.com)* or JVirGel Proteomic Tools *(PRODORIC Net, Germany)*

Mass, pI, composition and mol% acidic, basic, aromatic, polar etc. amino acids - PEPSTATS *(EMBOSS)*

Biochemistry-online *(Vitalonic, Russia)* gives one % composition, molecular weight, pI, and charge at any desired pH

Composition/Molecular Weight Calculation *(Georgetown University Medical Center, U.S.A.)* - the only problem with this site is that when run in batch mode it does not identify the sequence by name, merely sequential number

Batch Protein Isoelectric Point determination - part of the Sequence Manipulation Suite

Batch Protein Molecular Weight determination - part of the Sequence Manipulation Suite

Protein calculator *(C. Putnam, The Scripps Research Institute, U.S.A.)* - calculates mass, pI, charge at a given pH, counts amino acid residues etc…

Computation of size of DNA and Protein Fragments from Their Electrophoretic Mobility (Raghava, G. P. S. 2001. Biotech Software and Internet Report 2:198-200)

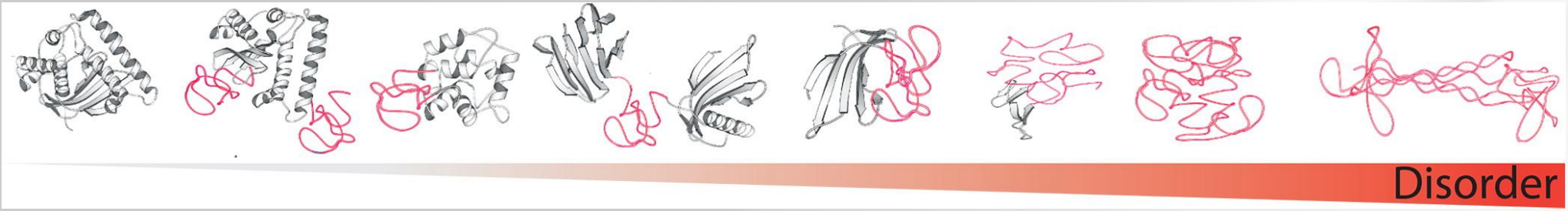
Intrinsically disordered proteins

Definition (Keith Dunker - Indiana University)

- Many proteins contain regions that lack specific 3D structure
- Some proteins lack specific 3D structure in their entireties under physiological conditions and yet carry out biological function

IDPs: Continuum of disorder

Different levels of order and disorder



no disorder	Disordered N- and C-termini	Disordered loop	Disordered linker	Disordered domain	Mostly collapsed protein	Wholly disorder	Extended protein

Functional disordered segments can be as small as only a few amino acid residues, or they can occupy rather long regions or ends

Johnny Habchi; Peter Tompa; Sonia Longhi; Vladimir N. Uversky; *Chem. Rev.* **2014,** 114, 6561-6588

Prediction of order/disorder

Prediction of intrinsic disorder

Database of protein disorder http://www.disprot.org

- these predictors were used to study whole genome data
- disorder increases in example proteomes in the order of

multicellular eukaryotes > single cellular eukaryotes > archaea > prokaryotes

- with significant associations of disorder with signaling, regulation, and posttranslational modification

Predictions of order/disorder

Algorithms based on the analysis of the sequence
Information associated to solubility often directly related to folding state, agregation or denaturation and secondary structure

Several servers and programs

http://www.disprot.org/predictors.php

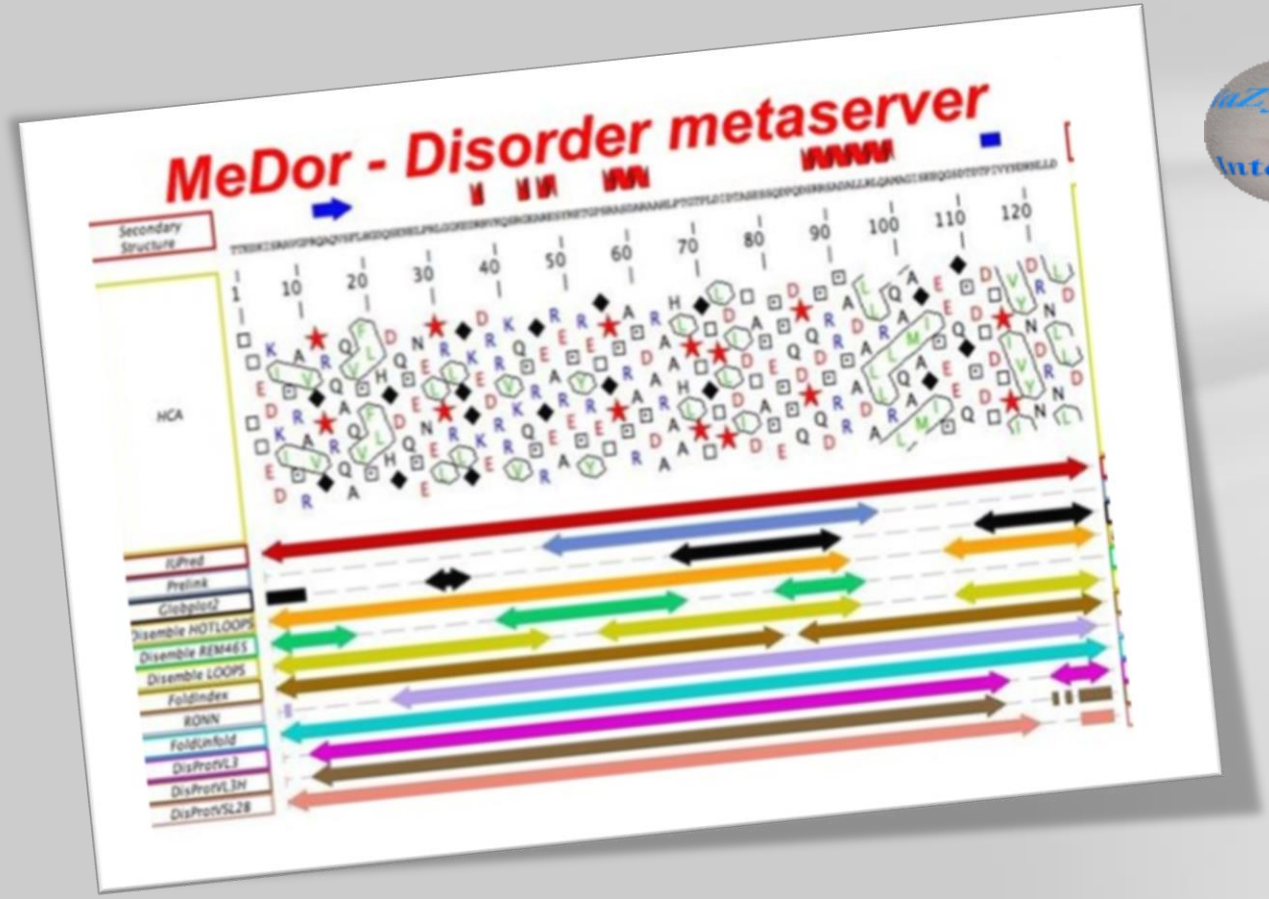
Protein Disorder Predictors

DISPROT	http://www.ist.temple.edu/disprot/Predictors.html
DisEMBL	http://dis.embl.de /
MEDOR	http://www.vazymolo.org/MeDor/
GLOBPLOT2	http://globplot.embl.de/
FoldIndex	http://bip.weizmann.ac.il/fldbin/findex/

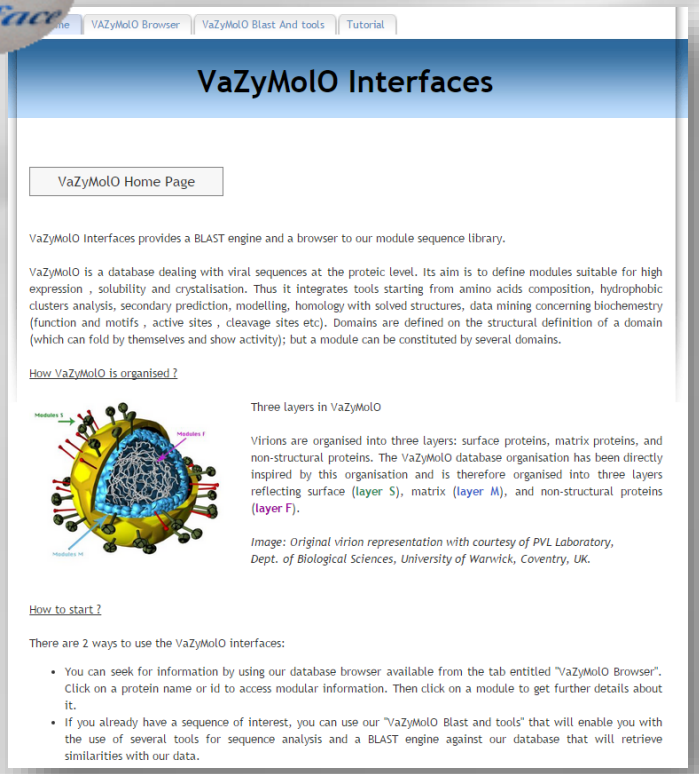
Prediction of order/disorder

MEDOR **http://www.vazymolo.org/MeDor/**

Prediction of regions sensitive to defolding, of potential interacting partners, …

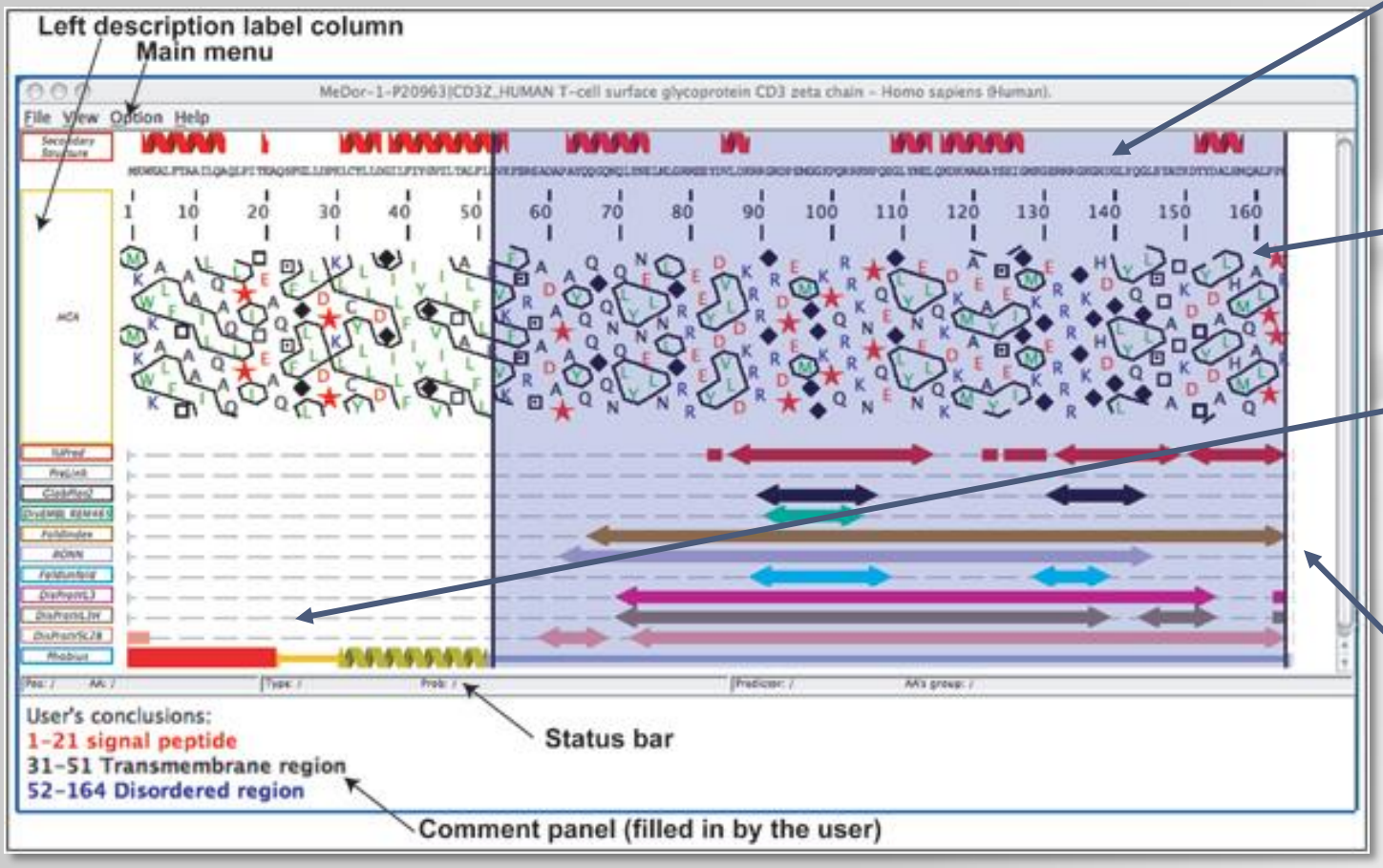


Database of viral proteins
Its aim is to define modules suitable for high expression, solubility and crystallization



Prediction of order/disorder

Example of a MeDor output



The sequence is represented below the predicted secondary structure elements (β-strands are represented by blue arrows, and α-helices are drawn in red)

HCA plot

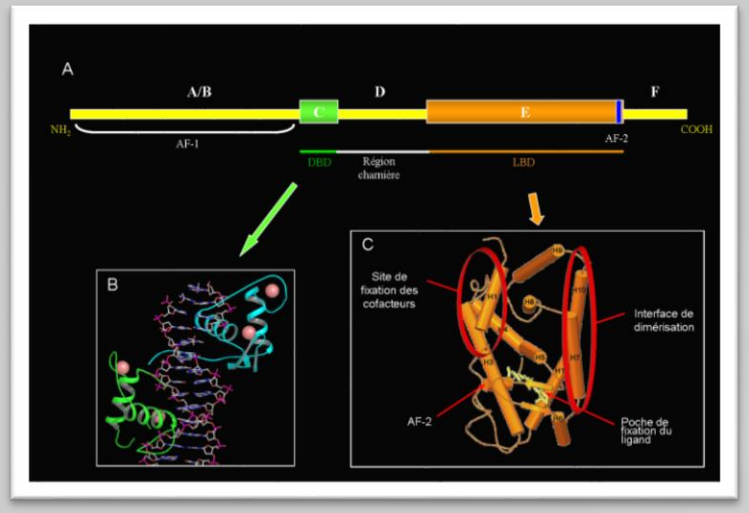
Peptide signals and TM domains predicted by Phobius are highlighted as red bars and yellow helices

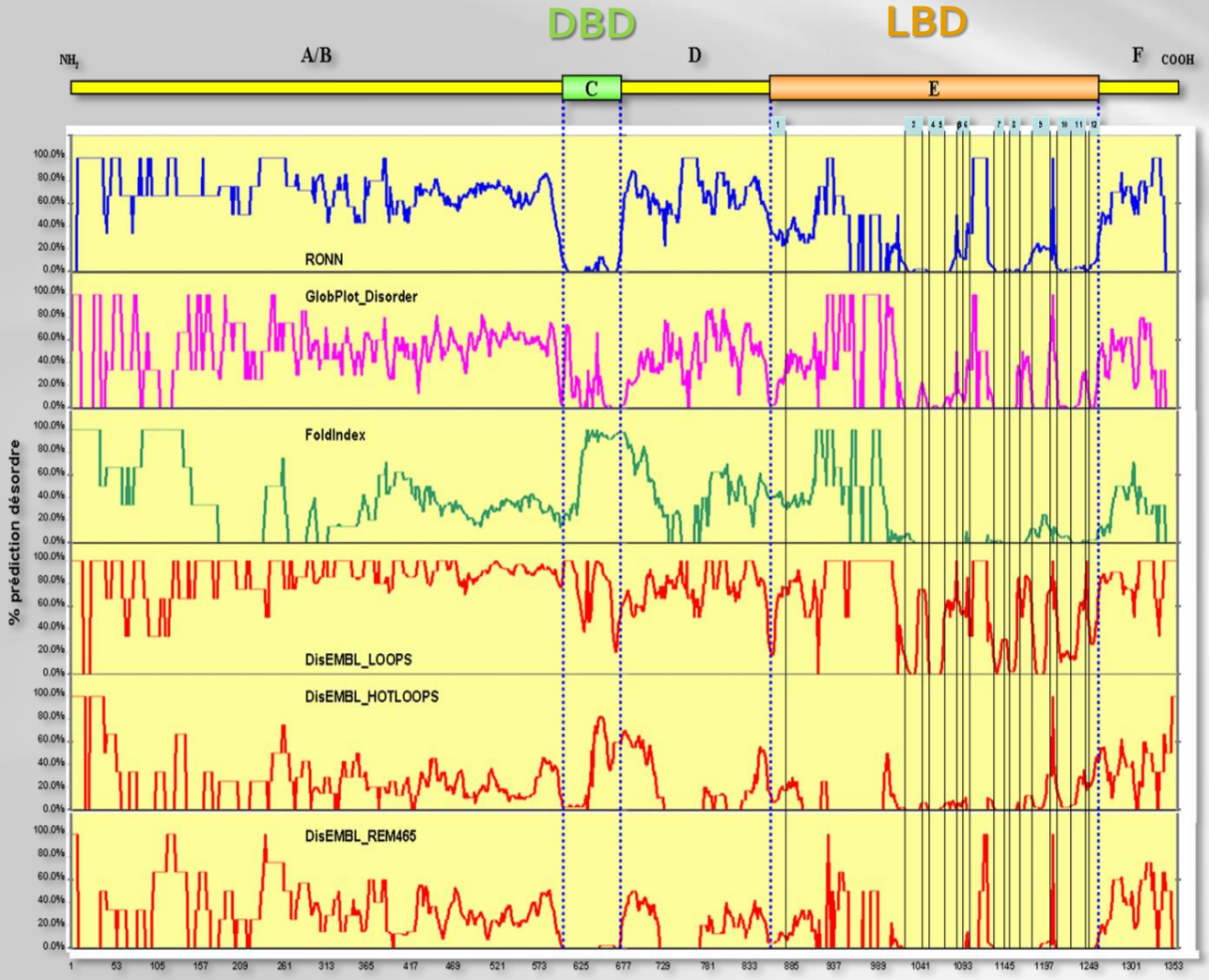
Predicted disordered regions are represented by bidirectional arrows of different colors as a function of predictors.

DisProt entry DP00200 human T cell glycoprotein CD3 Z chain (P20963)

Predictions of order/disorder

Exemple : human nuclear receptors





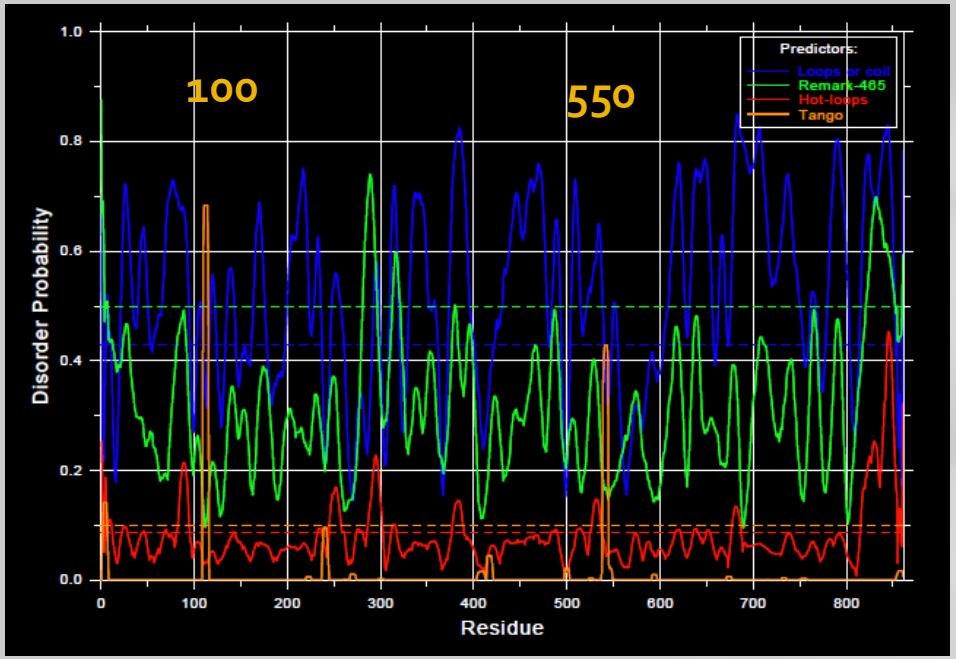
Predictions of disorder using different programs and a multiple alignment of 48 human nuclear receptors

Transcription factors sensing hydrophobic ligands (steroids, thyroid hormones, …) regulating gene expression

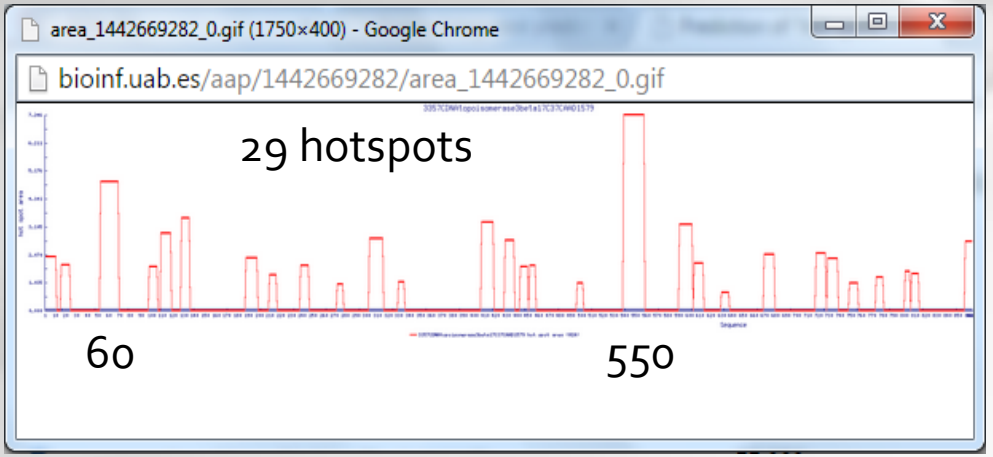
Prediction of aggregation

TANGO http://dis.embl.de/ **AGGRESCAN** http://bioinf.uab.es/aggrescan/

Prediction of domains with propensity to aggregate



Based on simple physico-chemical principles of secondary structure formation extended by the assumption that the core regions of an aggregate are fully buried



Based on an aggregation-propensity scale for natural amino acids derived from *in vivo* experiments and on the assumption that short and specific sequence stretches modulate protein aggregation

Prediction of protein solubility

http://molbiol-tools.ca/Protein_Chemistry.htm

Solubility and crystalizability

PROSO and PROSO II - are sequence-based **PRO** tein **SO**lubility evaluators which try to answer the following question: "Which of my cloned proteins have the best/worst chances to be soluble upon heterologous expression?" (Smialowski P et al. 2007. Bioinformatics **23:**2536-2542 & Smialowski P et al. 2012. FEBS J. **279:** 2192-2200)

ESPRESSO (**ES**timation of **PR**otein **E**xpre**S** sion and **SO**lubility) - is a sequence-based predictor for estimating protein expression and solubility for three different protein expression systems: *in vivo Escherichia coli*, *Brevibacillus*, and wheat germ cell-free (Hirose S, & Noguchi T. 2013. Proteomics. **13:**1444-1456)

SABLE - Accurate sequence-based prediction of relative Solvent AccessiBiLitiEs,secondary structures and transmembrane domains for proteins of unknown structure (Adamczak R et al. 2004. Proteins **56:**753-767)

SPpred (**S**oluble **P**rotein prediction) *(Bioinformatics Center, Institute of Microbial Technology, Chandigarh, India)* - is a web-server for predicting solubility of a protein on over expression in *E.coli.* The prediction is done by hybrid of SVM model trained on PSSM profile generated by PSI-BLAST search of 'nr' protein database and splitted amino acid composition

SECRET - is a **SE**quence-based **CR**ystallizability **E**valua**T** or which tries to answer the following questions:"What is the chance that my soluble protein will crystallize?" & "Which of my soluble proteins have the best/worst chances to crystallize?" (Smialowski P et al. 2006. Proteins 62: 343-355)

Surface **E**ntropy **R**eduction **p** rediction (SERp) - this exploratory tool aims to aid identification of sites that are most suitable for mutation designed to enhance crystallizability by a Surface Entropy Reduction approach (Goldschmidt L. et al. 2007. Protein Science. **16:**1569-1576)

CRYSTALP2 - for *in-silico* prediction of protein crystallization propensity (Kurgan L, et al. 2009. BMC Structural Biology **9:** 50); and PPCpred - sequence-based prediction of propensity for production of diffraction-quality crystals, production of crystals, purification and production of the protein material (M.J. Mizianty & L. Kurgan. 2011. Bioinformatics **27:** i24-i33)

Prediction of protein solubility

http://omictools.com/protein-solubility-c1306-p1.html

Protein solubility prediction tools

Recombinant protein technology is essential for conducting protein science and using proteins as materials in pharmaceutical or industrial applications. Although obtaining soluble proteins is still a major experimental obstacle, knowledge about protein expression/solubility under standard conditions may increase the efficiency and reduce the cost of proteomics studies (source text: Hirose and Noguchi, 2013).

Filter by type of tool: Program | Database | Link to literature

A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli

Authors: Habibi, N., Mohd Hashim, S.Z., Norouzi, A., and Samian, M.R.

Abstract: BACKGROUND: Over the last 20 years in biotechnology, the production of recombinant proteins has been a crucial bioprocess in both...

ccSol

A tool to predict the solubility of proteins based on their physicochemical properties.

ccSOL omics

A method to perform large-scale solubility predictions of endogenous and heterologous protein expression in E. coli.

eSOL

A database on the solubility of entire ensemble E.coli proteins.

ESPRESSO

EStimation of PRotein ExpreSsion and Solubility

A sequence-based predictor for estimating protein expression and solubility.

PROSO

A sequence-based PROtein SOlubility evaluator.

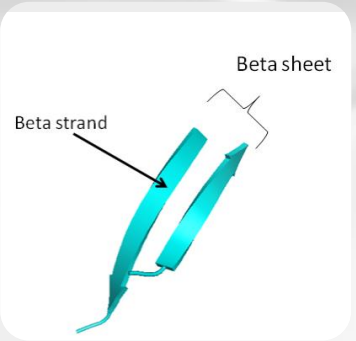
Secondary structure prediction

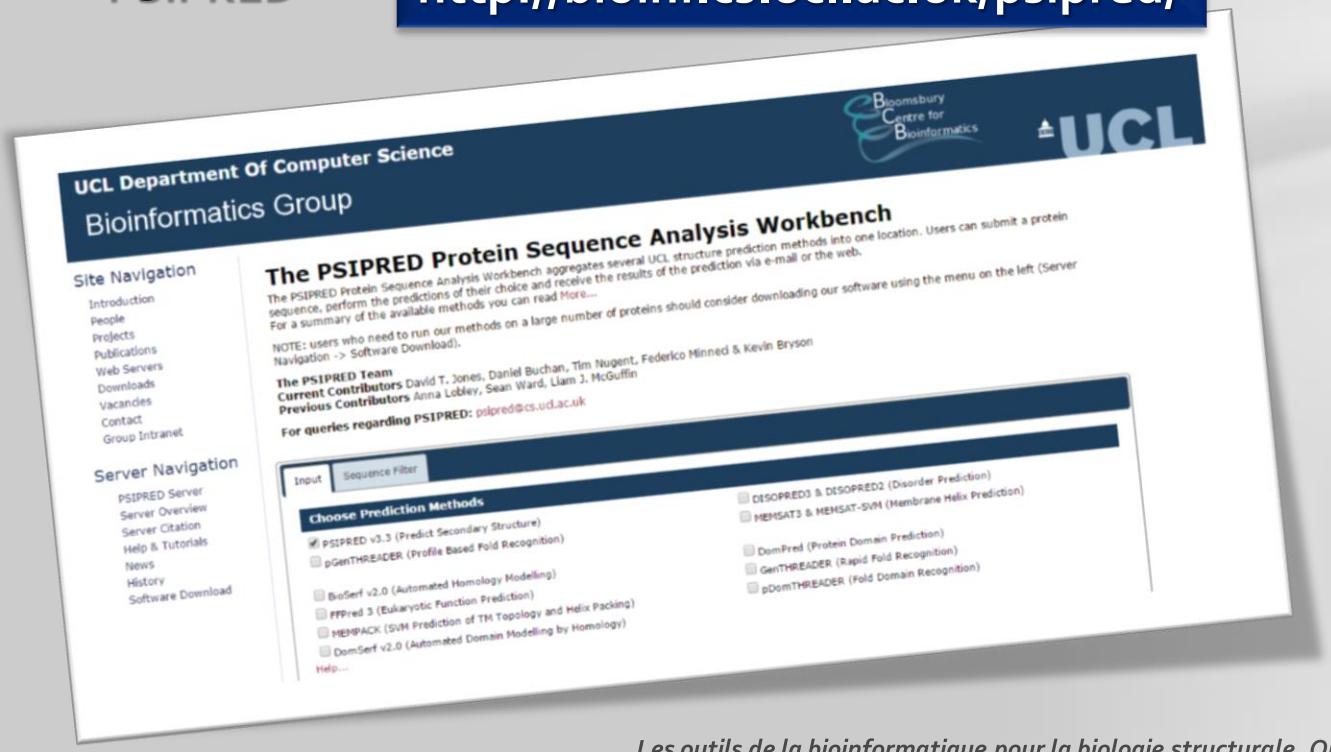
Based on the propensity of each aminoacid to form a secondary structure (helix and strand)

SOPMA https://npsa-prabi.ibcp.fr/

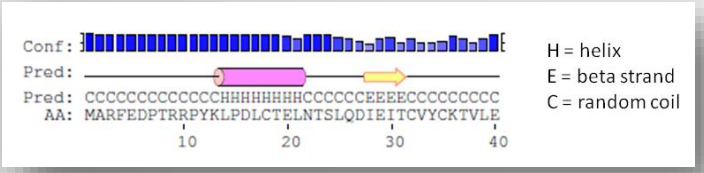
Jpred 4 http://www.compbio.dundee.ac.uk/jpred/

PSIPRED http://bioinf.cs.ucl.ac.uk/psipred/



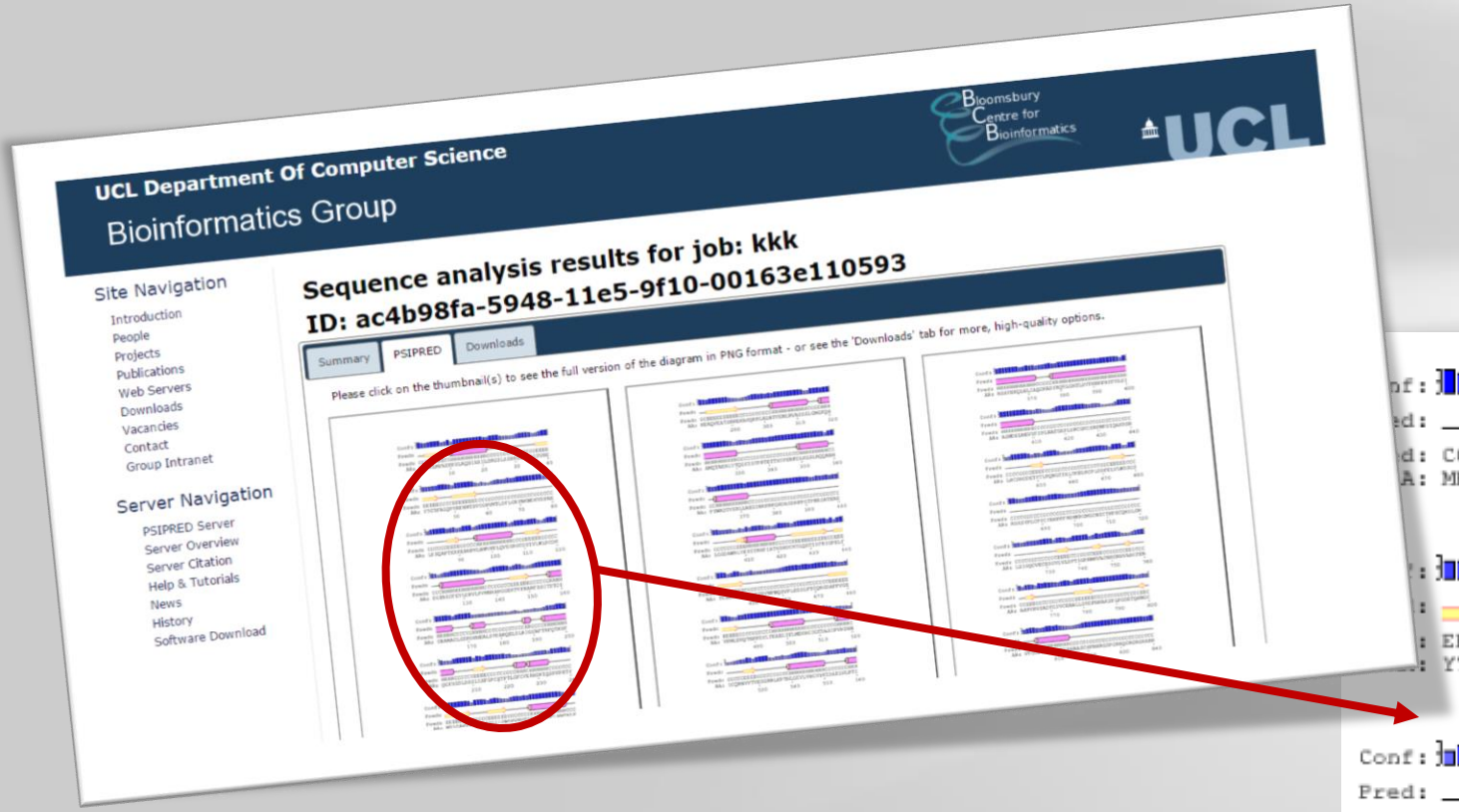


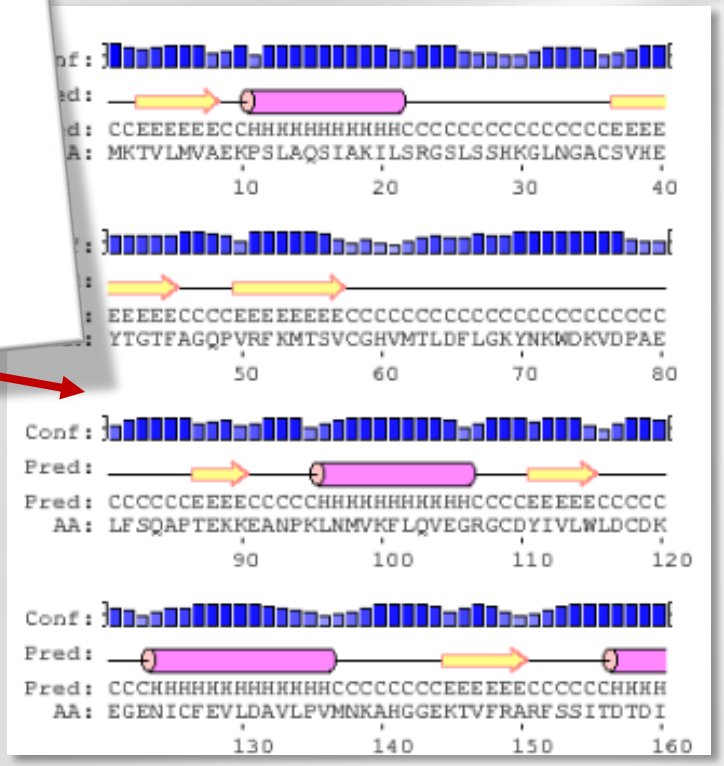
Prediction all along the sequence with a confidence index



Secondary structure prediction

Example of a PSIPRED output





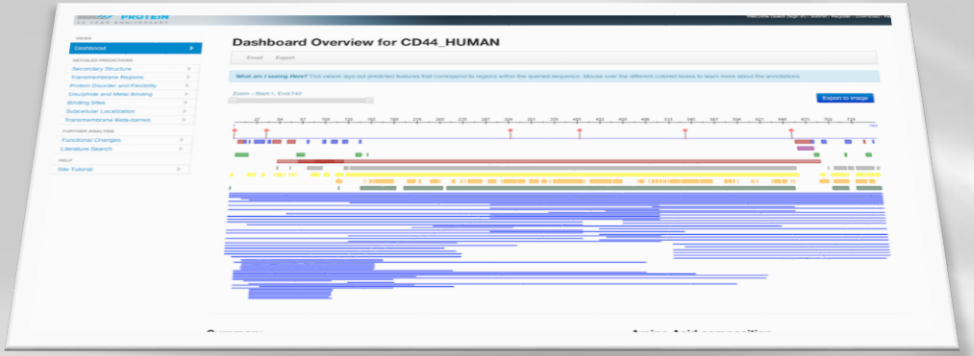
human Topo IIIα

Secondary structure prediction and more

To go further

PredictProtein Server

https://www.predictprotein.org/

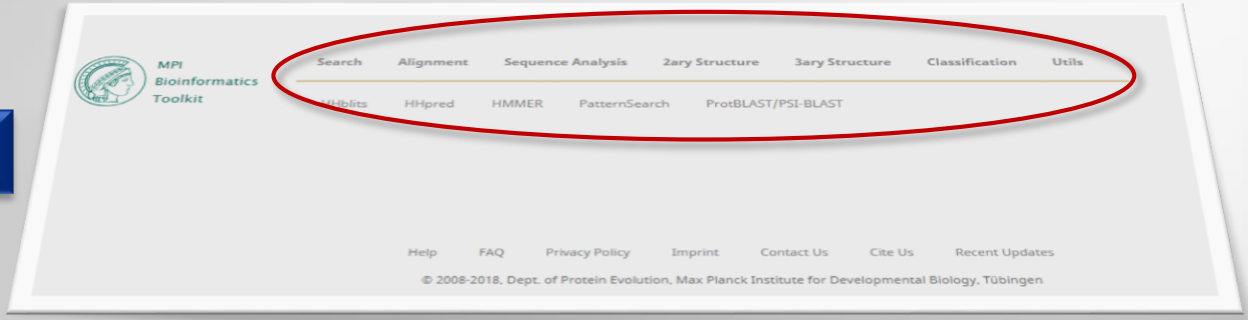


prabi
Pôle Rhône-Alpes de Bioinformatique

https://prabi.ibcp.fr/htm/site/web/home/

Toolkit

http://toolkit.tuebingen.mpg.de/

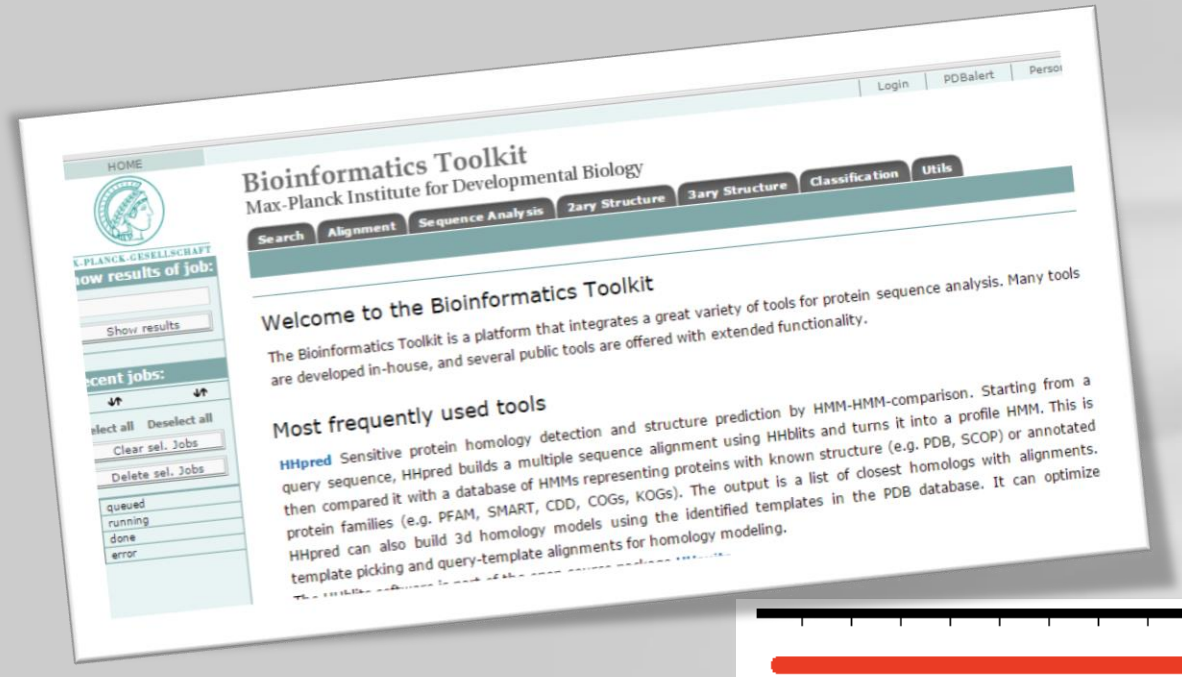


From sequence to structure prediction

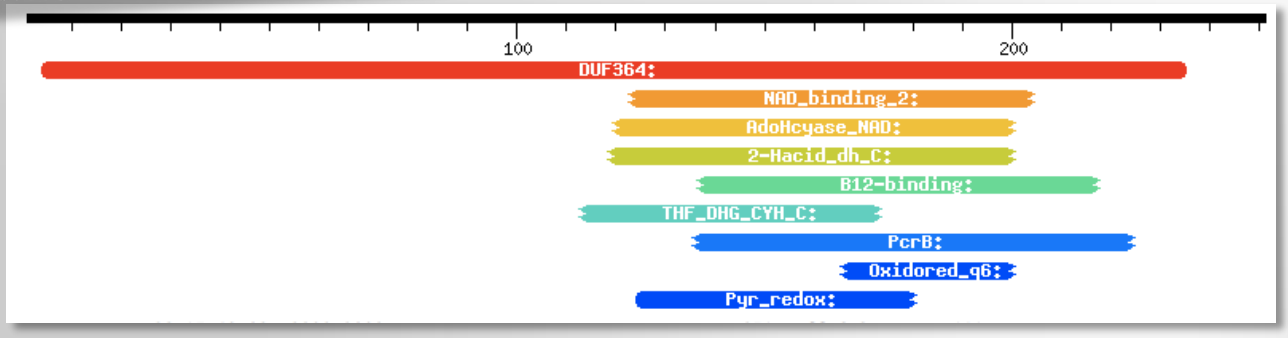
Homology detection and structure prediction by HMM-HMM comparison

HHpred

http://toolkit.tuebingen.mpg.de/hhpred



HHpred is often used for remote homology detection and homology-based function prediction
It runs with the free, open-source software package **HH-suite** for fast sequence searching, protein threading and remote homology detection



From sequence to structure prediction

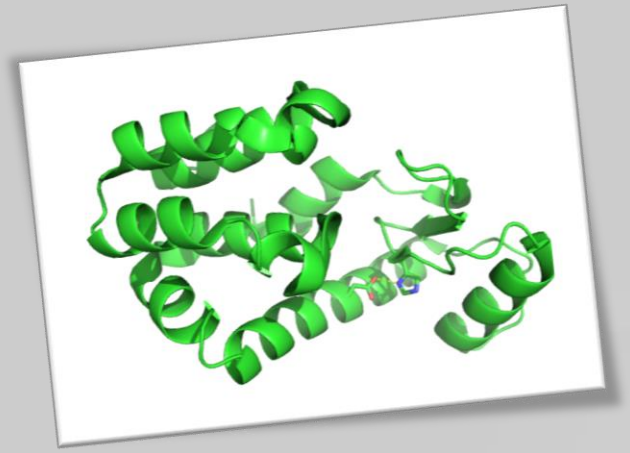
Organisation in domains

Protein organisation in structural and functional domains

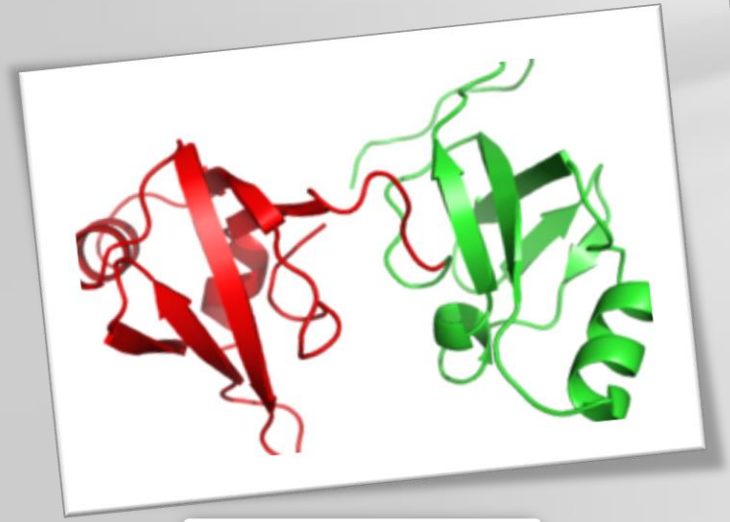
- Conserved part of a protein sequence forming an independent structure that can evolve, operate and exist independently of the rest of the protein chain
- Each domain forms a compact three-dimensional structure and is often independently stable and folded
- Many proteins consist of several structural domains
- A single domain can appear in a variety of different proteins
- Molecular evolution uses domains as building blocks, and these can be recombined in different arrangements to create proteins with different functions
- The domains length varies between about 25 to 500 amino acids
- The shorter domains like zinc fingers are stabilized by metal ions or disulfide bridges
- Molecular evolution uses domains as building blocks, and these can be recombined in different arrangements to create proteins with different functions
- Domains often form functional units

From sequence to structure prediction

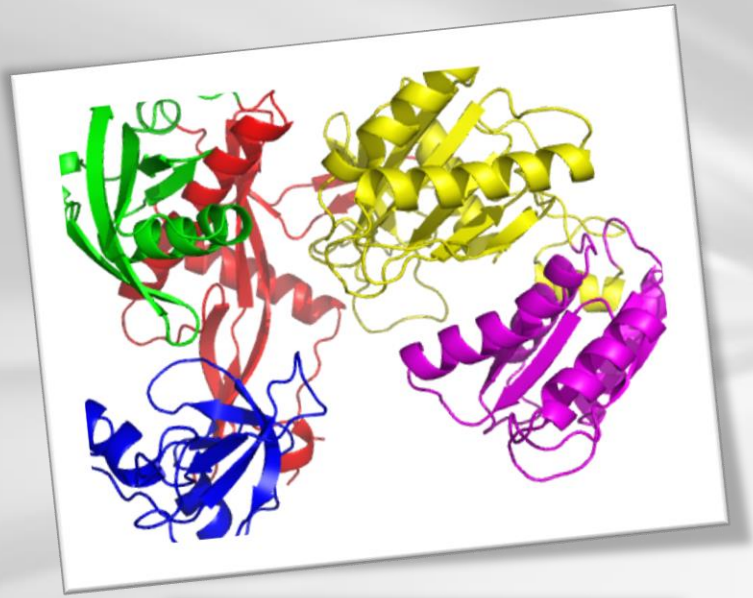
Organisation in domains



One domain



Two domains

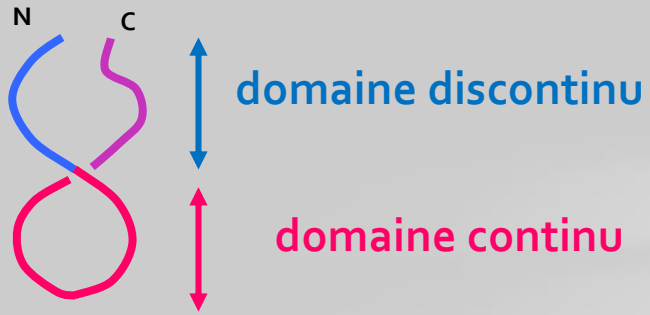


Multidomain protein

Well separated or tightly packed

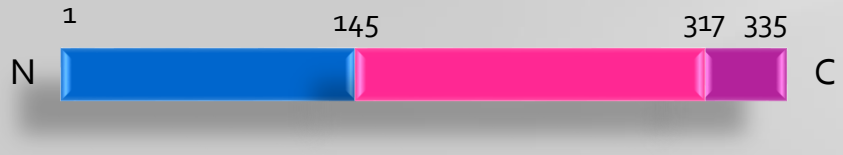
From sequence to structure prediction

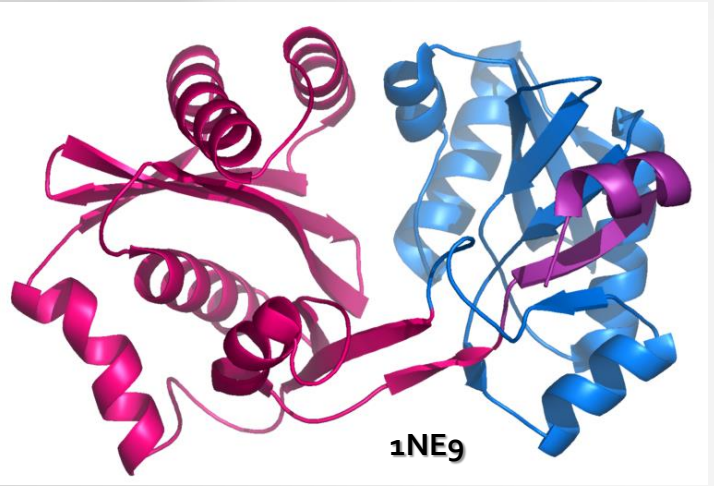
Continuous / discontinuous domains



- Very difficult to predict
- Very difficult to align when continuous/discontinuous similarity

Exemple : Fem protein family





Domain 2 Domain 1

From sequence to structure prediction

Domain prediction and assignation

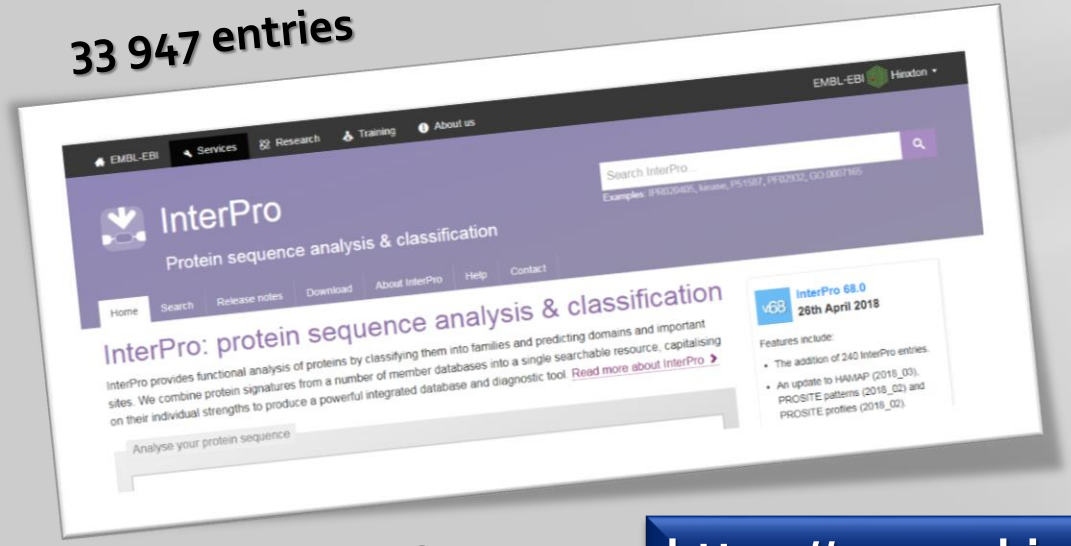
- From the sequence
- From the multiple alignment
- From the structure
- From databanks

General

UNIPROT http://www.uniprot.org/

Specialized **InterPro** https://www.ebi.ac.uk/interpro/

33 947 entries



InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.

InterProScan https://www.ebi.ac.uk/interpro/search/sequence-search

From sequence to structure prediction

Integrated databanks

Pfam

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains.

PROSITE

PROSITE is a database of protein families and domains.

Prodom

ProDom protein domain database consists of an automatic compilation of homologous domains

SMART

SMART allows the identification and annotation of genetically mobile domains

PRINTS

PRINTS is a compendium of protein fingerprints.

SuperFamily

New! SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure, based on SCOP.

PIR SuperFamily

PIR SuperFamily (PIRSF) is a classification system based on evolutionary relationship of whole proteins.

SwissProt

SWISS-PROT database consists of protein sequence entries.

TIGRFAMs

TIGRFAMs is a collection of protein families.

From sequence to structure prediction

Integrated databanks

Banque de 'patterns' [CxxC], motifs, signatures

- basé sur la présence des acides aminés
- très restrictif (nombreux faux positifs)

Banque de domaines basés sur des alignements multiples

- motifs assez longs liés à la famille alignée
- nombreuses erreurs, nombreux DUF (domains of unknown function)

Banque de 'fingerprints' susceptibles de caractériser une famille

- proche des signatures (itératifs et combinatoires)
- nombreux motifs très courts

Banque de domaines homologues définis par des recherches Psi-Blast

- basé sur Pfam, nombreux domaines (3 739 157 / 426 997 with PDB)
- très bonne présentation, facile d'accès

Banque de grands domaines (>1300) homologues et alignés

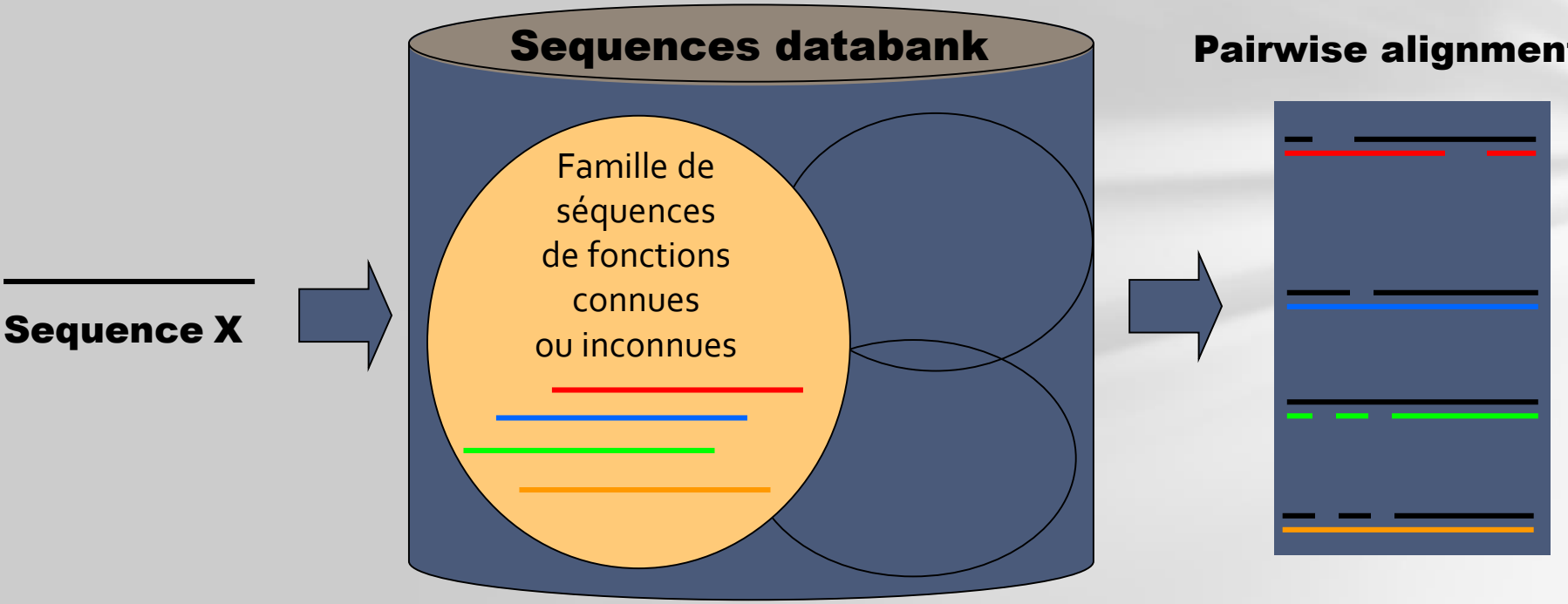
- annotation, détermination automatique par différentes méthodes (profile, HMM, Psi-Blast, alignement multiples…)
- distribution par espèce, facile d'accès

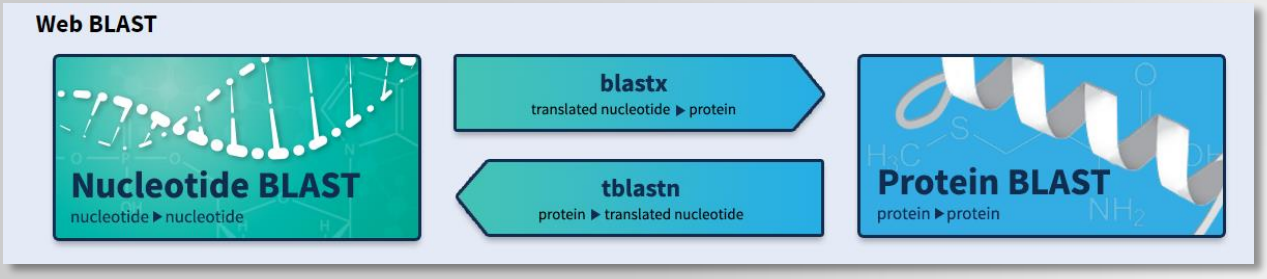
Sequence retrieval

The principles

BLAST https://blast.ncbi.nlm.nih.gov/Blast.cgi

Basic Local Alignment Search Tool

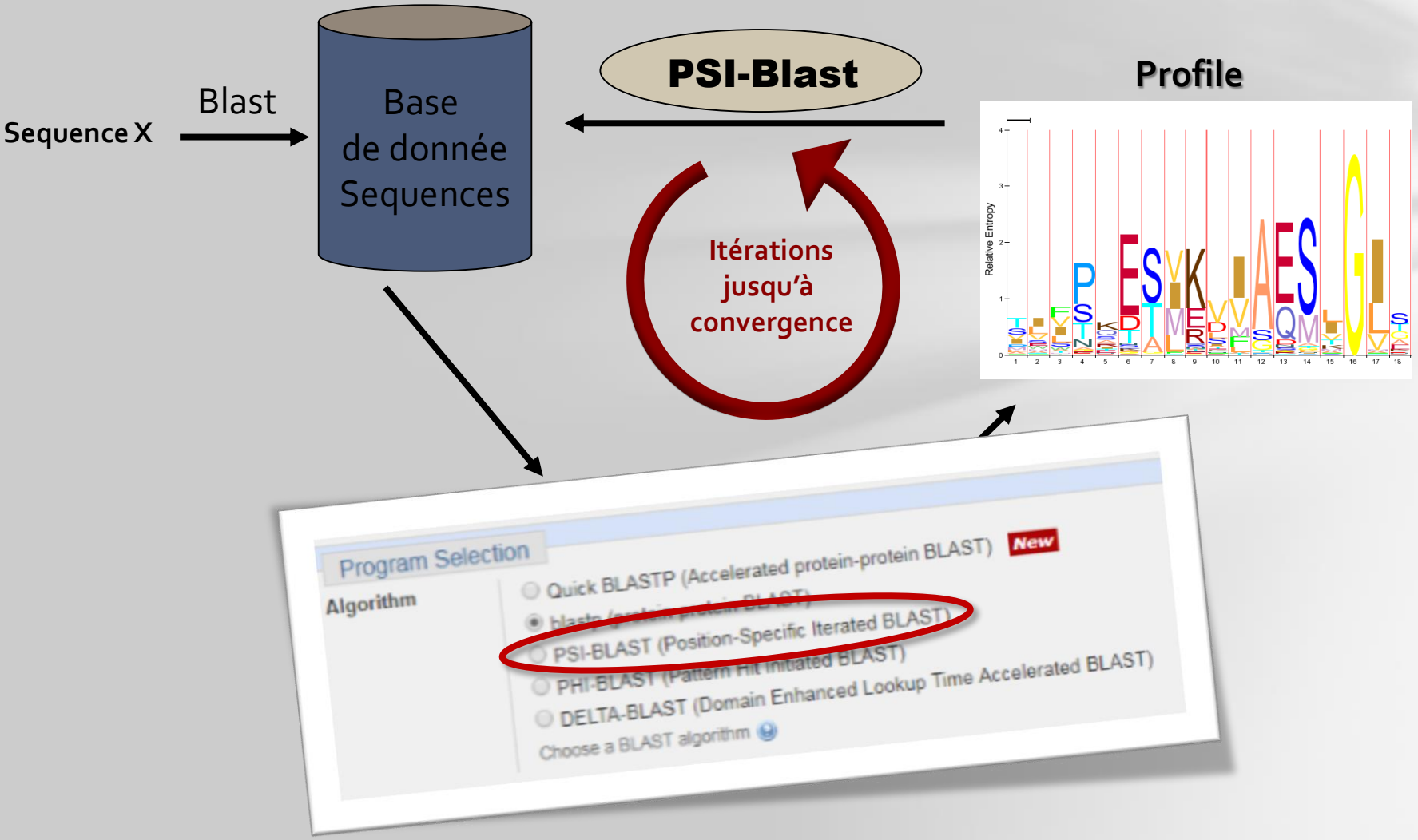




Sequence retrieval

The principles

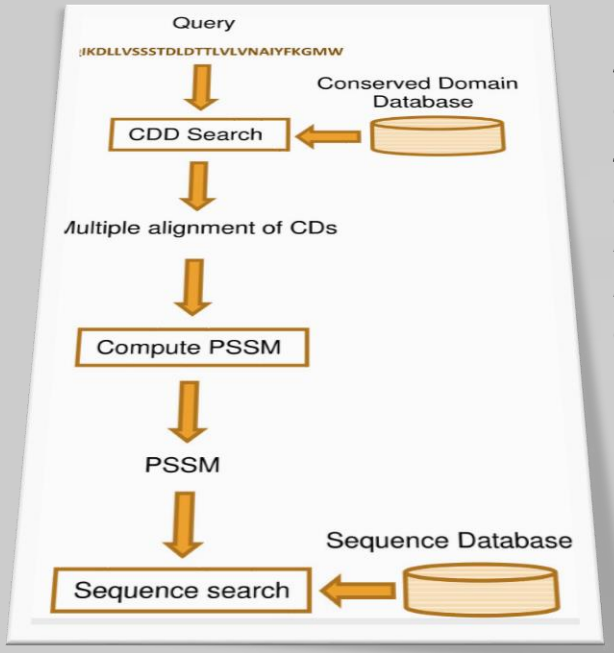
BLAST https://blast.ncbi.nlm.nih.gov/Blast.cgi



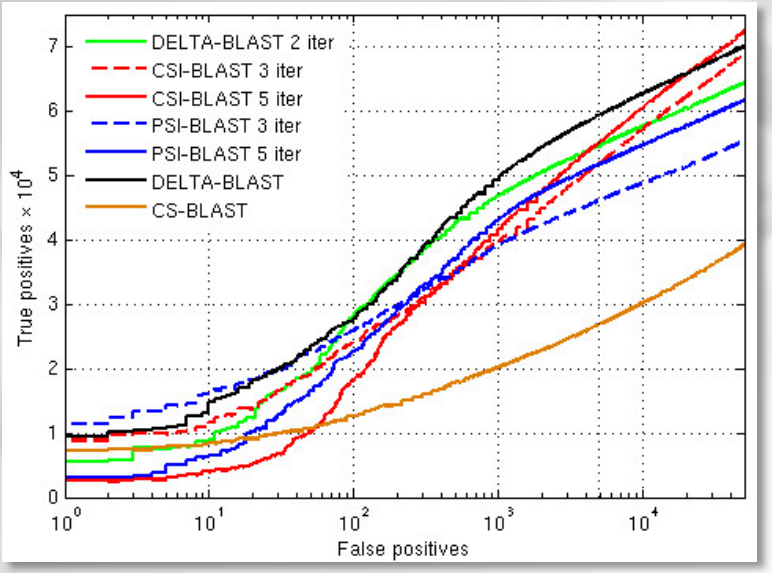
Sequence retrieval

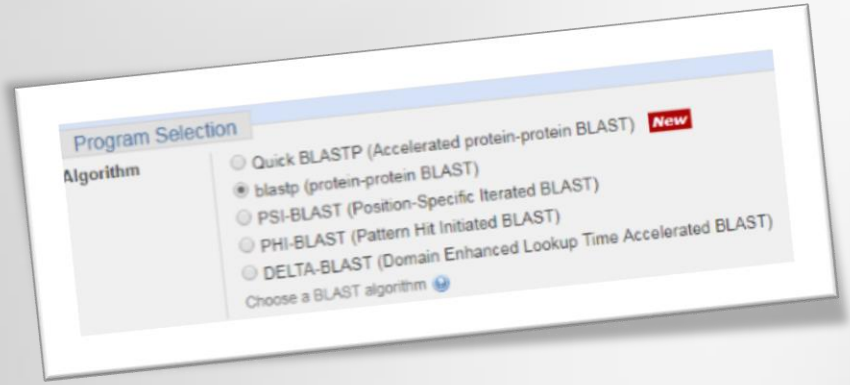
Domain enhanced lookup time accelerated BLAST

https://blast.ncbi.nlm.nih.gov/Blast.cgi



Searches a database of pre-constructed PSSMs before searching a protein-sequence database, to yield better homology detection
For its PSSMs, DELTA-BLAST employs a subset of NCBI's Conserved Domain Database (CDD)





DELTA-BLAST is a useful program for the detection of remote protein homologs

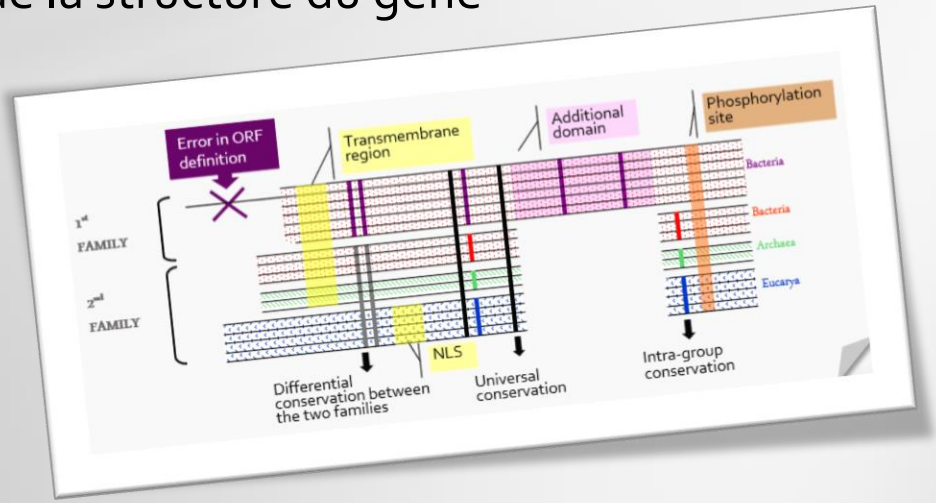
Multiple alignment of complete sequences

Information sur l'organisation en domaines de la protéine
Importance pour la prédiction de fonction, de structure

Conservation au sein de la famille
résidus conservés strictement chez toutes les séquences
importance structurale ou fonctionnelle : motif caractéristique
résidus conservés spécifiquement dans un sous-groupe de séquences
résidus discriminants

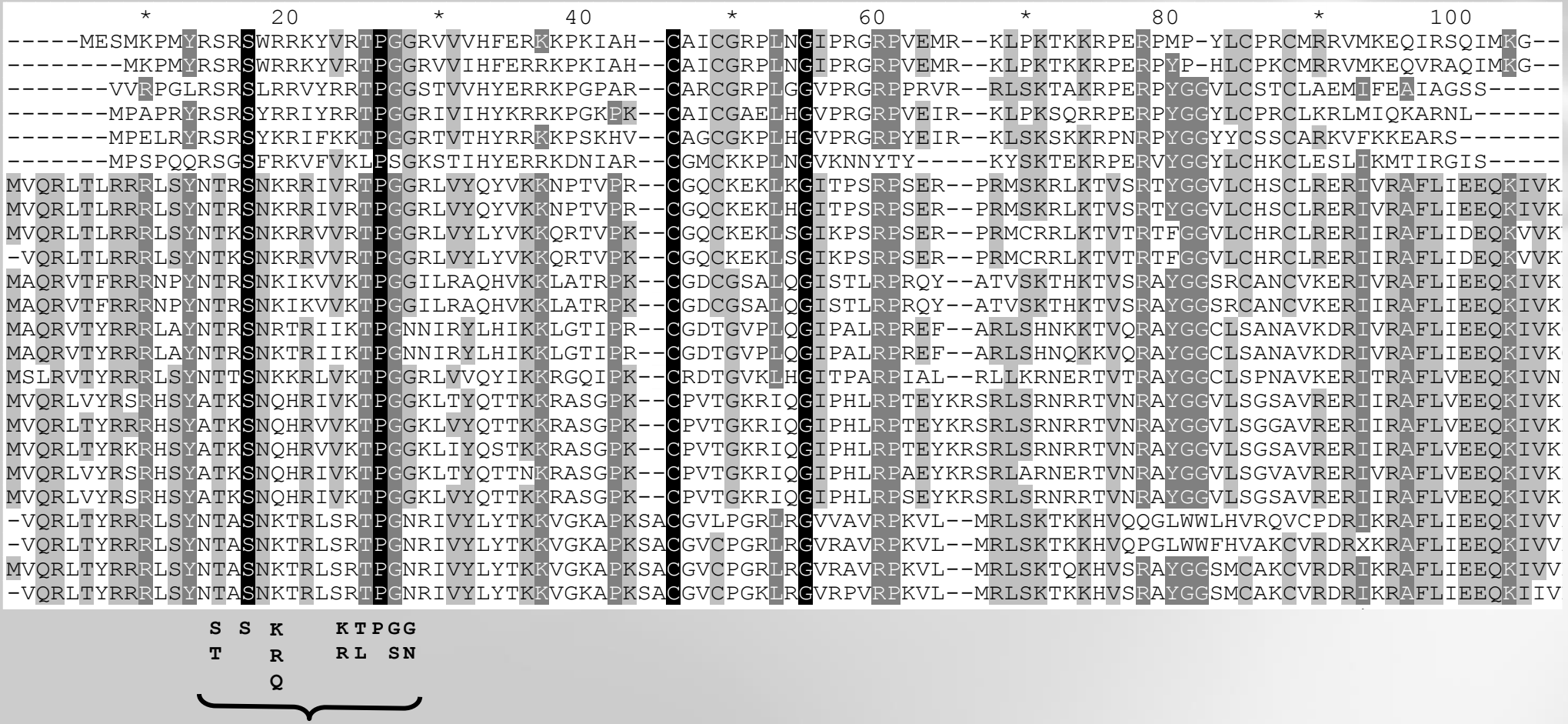
Validation de la séquence protéique
détection des erreurs de séquençage, de *frameshift*
détection des erreurs dans la prédiction de la structure du gène
codon initiateur, sites d'épissage exon/intron

Aspects évolutifs
de la famille à l'arbre de la vie
notion de transfert horizontal



Multiple alignment of complete sequences

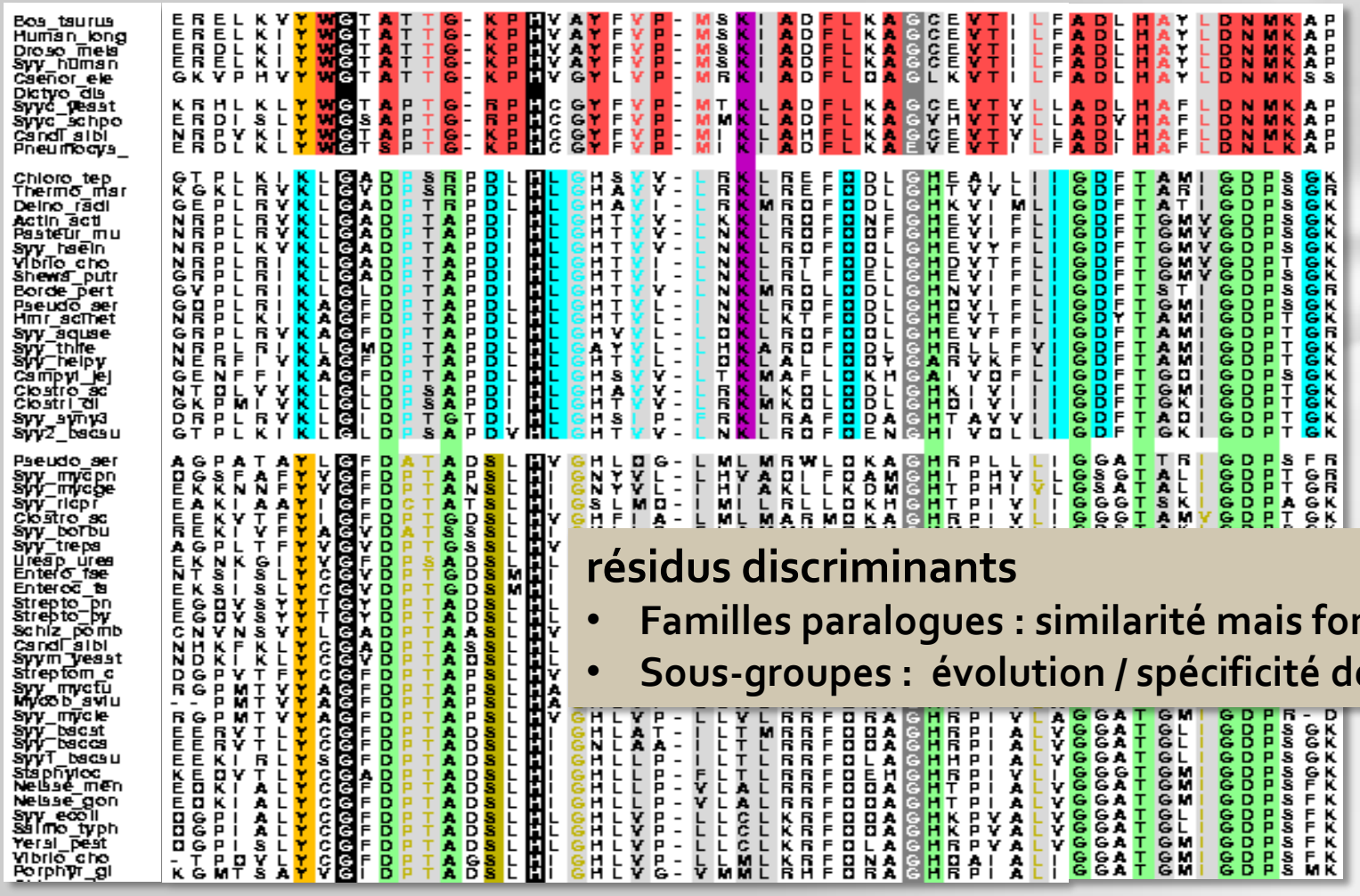
Exemple : Famille des protéines ribosomales L34e



Exemple de motif : [ST]-x-S-x[KRQ]-x-x-x-x-[KR]-[TL]-P-[GS]-[GN]

Multiple alignment of complete sequences

➡ Conservations différentielle



résidus discriminants

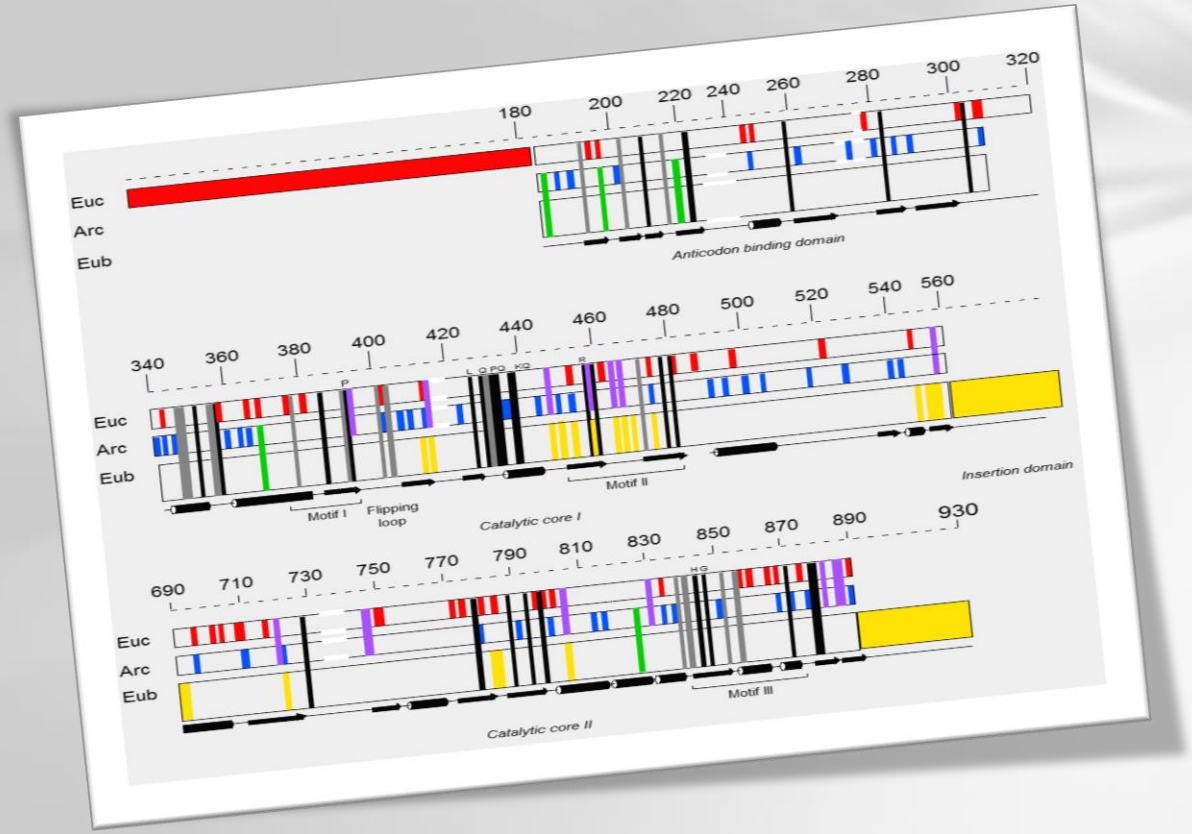
- Familles paralogues : similarité mais fonction différente
- Sous-groupes : évolution / spécificité de la fonction

Multiple alignment of complete sequences

Predictions

Intégration de l'information dans un contexte de famille

- Moyenne de prédictions des structures secondaires
- Site de modification
- Localisation cellulaire
- Zone transmembranaire
- Définition des bornes de domaines pou les études structurales
- Prédiction de la fonction biologique
- Modélisation moléculaire
- Nouveaux repliements
- …



Multiple alignment of complete sequences

How to obtain a multiple alignment?

Clustal Omega	https://www.ebi.ac.uk/Tools/msa/clustalo/
MUSCLE	http://www.drive5.com/muscle/
MAFFT	http://mafft.cbrc.jp/alignment/server/
PipeAlign2	http://lbgi.fr/pipealign

Multiple alignment of complete sequences

How to visualize a multiple alignment?

Jalview	http://www.jalview.org
Geneious	http://www.geneious.com/
Alscript	http://www.compbio.dundee.ac.uk/software.html
ESPript	http://espript.ibcp.fr/ESPript/ESPript/
WebLogo	http://weblogo.berkeley.edu/logo.cgi

Multiple alignment of complete sequences

http://lbgi.fr/pipealign

Multiple aligment starting from one sequence

Introduction : qu'est-ce qu'une structure ?

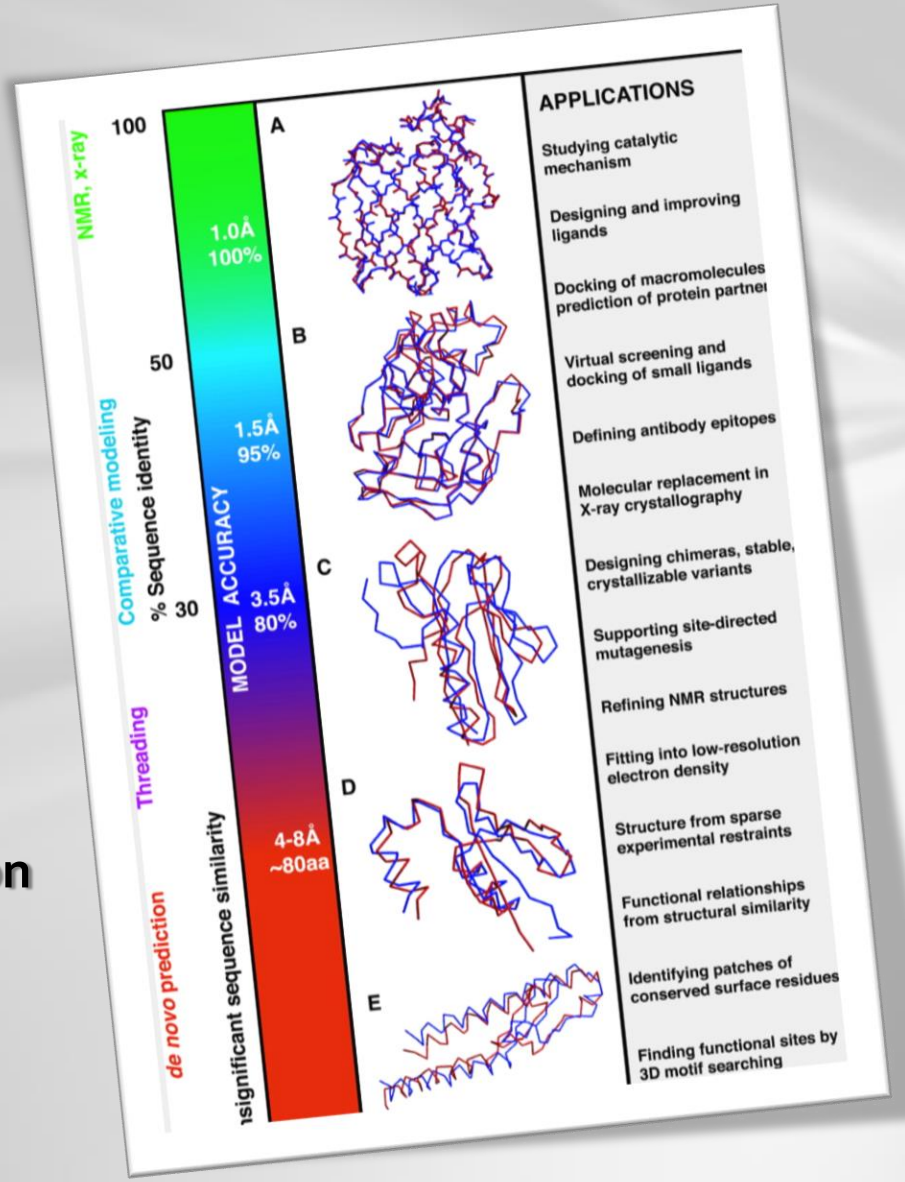
Informations données par la séquence

Prédiction de la structure

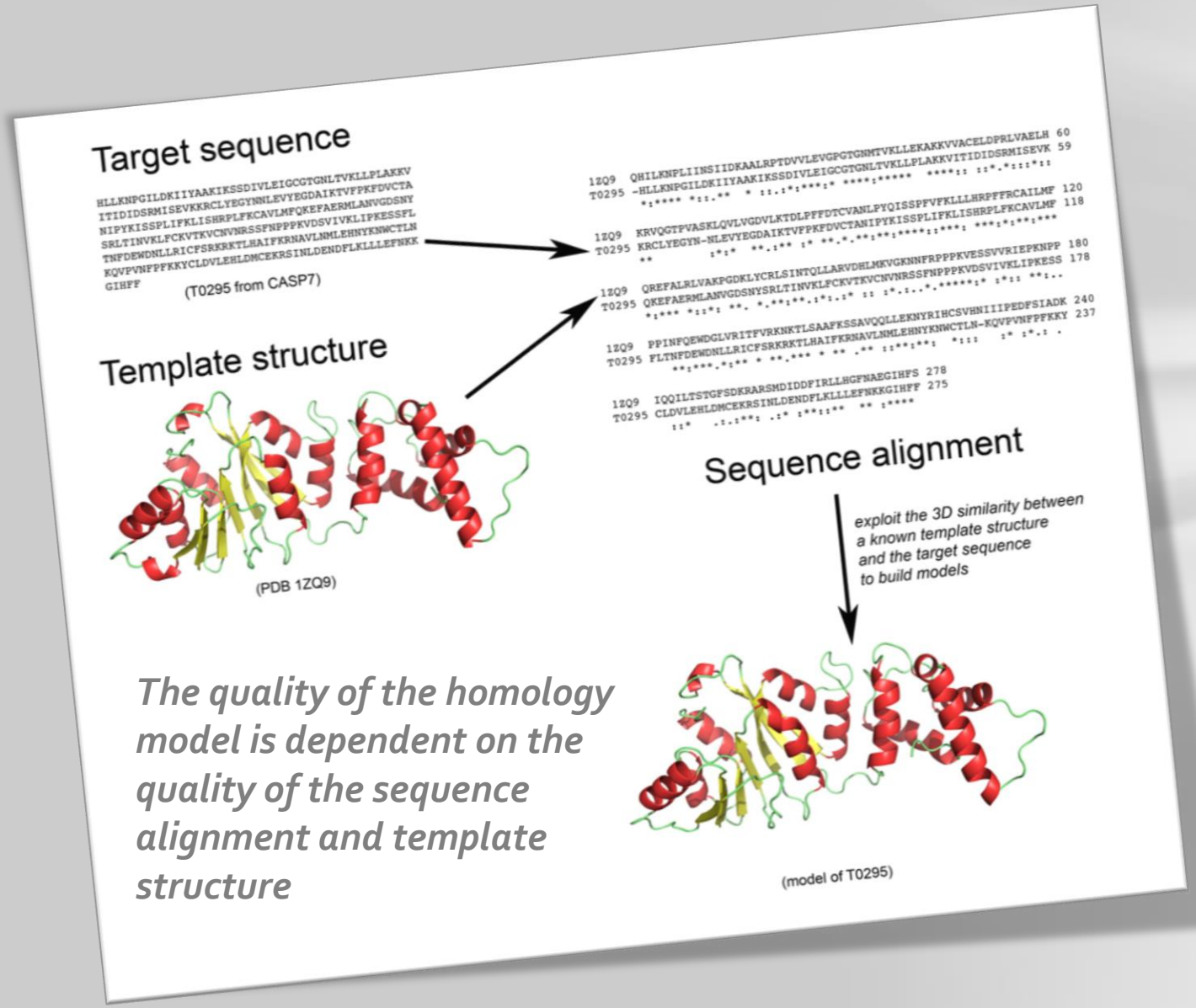
Informations données par la structure

Structure prediction methods

- Comparative modeling
- Secondary structure prediction
- Fold recognition
- *Ab initio* prediction
- Transmembrane segment prediction



Homology modeling



The quality of the homology model is dependent on the quality of the sequence alignment and template structure

Evolutionarily related proteins have similar sequences and naturally occurring homologous proteins have similar protein structure

3D protein structure is evolutionarily more conserved than would be expected on the basis of sequence conservation alone

The sequence alignment and template structure are then used to produce a structural model of the target

Homology modeling

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 *cristallographique 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 supérieures à 40 % d'identité donne des résultats équivalents

MAIS

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

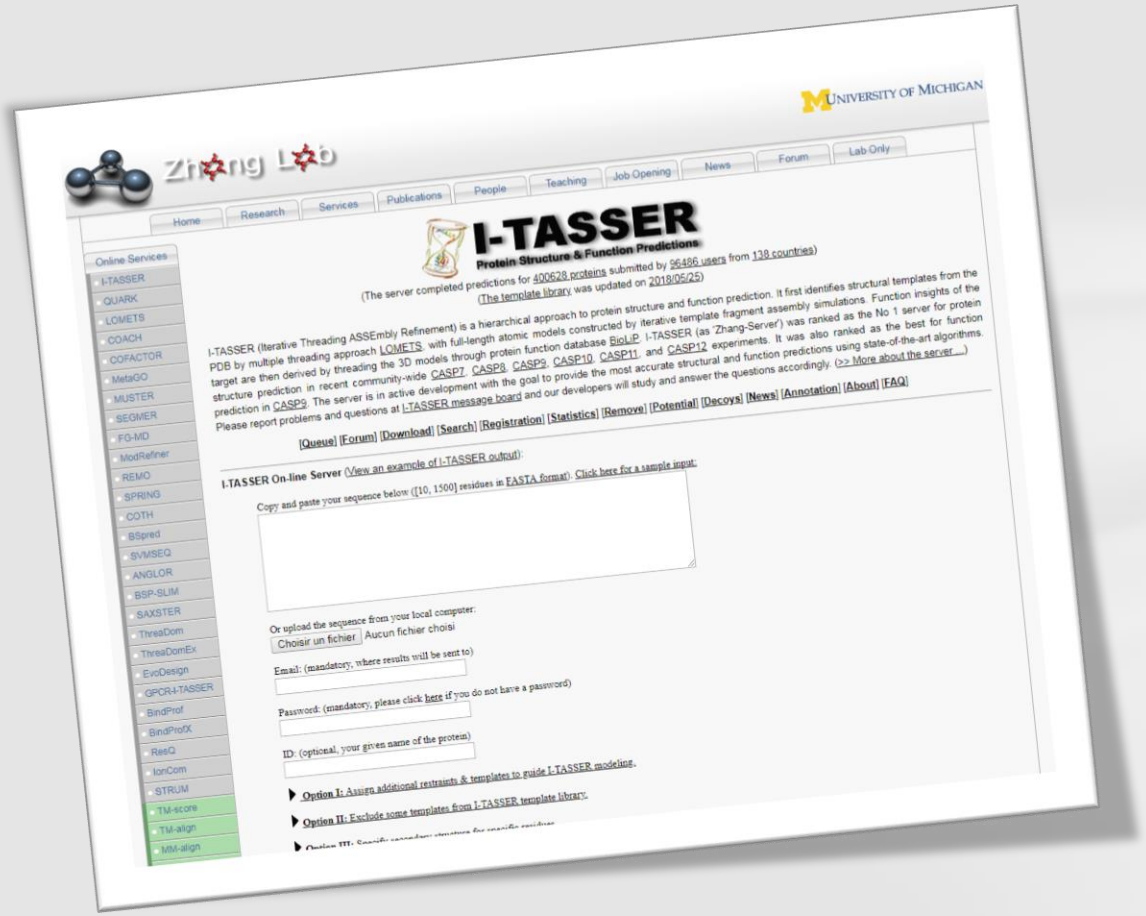
- L'emploi de *templates* multiples améliore grandement la qualité des alignements, donc des modèles obtenus

Homology modeling

I-TASSER	is the best server for protein structure prediction according to the 2006-2012 CASP experiments
RaptorX	excels at aligning hard targets according to the 2010 CASP9 experiments RaptorX generates the significantly better alignments for the hardest 50 CASP9 template-based modeling targets than other servers
MODELLER	is a popular software tool for producing homology models by satisfaction of spatial restraints using methodology derived from NMR data processing The ModWeb comparative protein structure modeling web-server uses primarily MODELLER for automatic comparative modeling
SWISS-MODEL	provides an automated web server for protein structure homology modeling
Robetta	widely used servers for protein structure prediction
SPARKSx	is one of the top performing servers in the CASP focused on the remote fold recognition
PEP-FOLD	is a *de novo* approach aimed at predicting peptide structures from amino acid sequences, based on a HMM structural alphabet
QUARK	is an algorithm developed for *ab initio* protein structure modeling

Structural similarity

I-TASSER (Iterative Threading ASSEmbly Refinement) is a hierarchical approach to protein structure and function prediction



It first identifies structural templates from the PDB by multiple threading approach **LOMETS**, with full-length atomic models constructed by iterative template fragment assembly simulations
Function insights of the target are then derived by threading the 3D models through protein function database **BioLiP**

I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent community-wide CASP7, CASP8, CASP9, CASP10, CASP11, and CASP12 experiments

Homology modeling

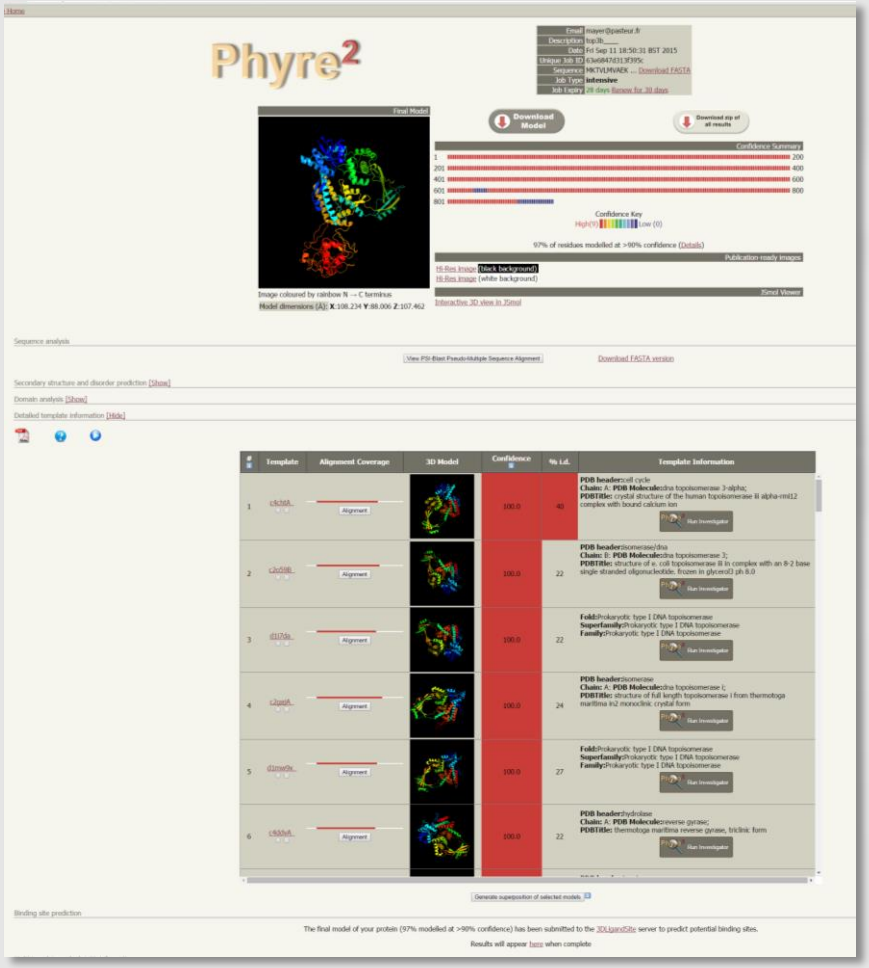
PHYRE2	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
3D-JIGSAW	http://www.bmm.icnet.uk/servers/3djisaw/
SWISS-MODEL	https://swissmodel.expasy.org/
MODELLER	https://toolkit.tuebingen.mpg.de/#/tools/modeller

Threading approach

I-TASSER	https://zhanglab.ccmb.med.umich.edu/I-TASSER/

Homology modeling

Phyre2 is amongst the top performing server in the CASP international blind trials of structure prediction in homology modelling and remote fold recognition, and are designed with an **emphasis on ease of use for non-experts**

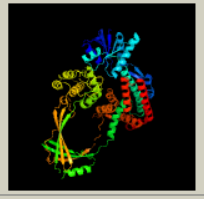
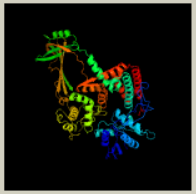
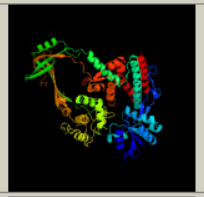
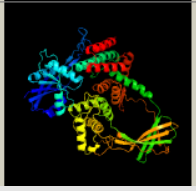


Homology modeling

Phyre

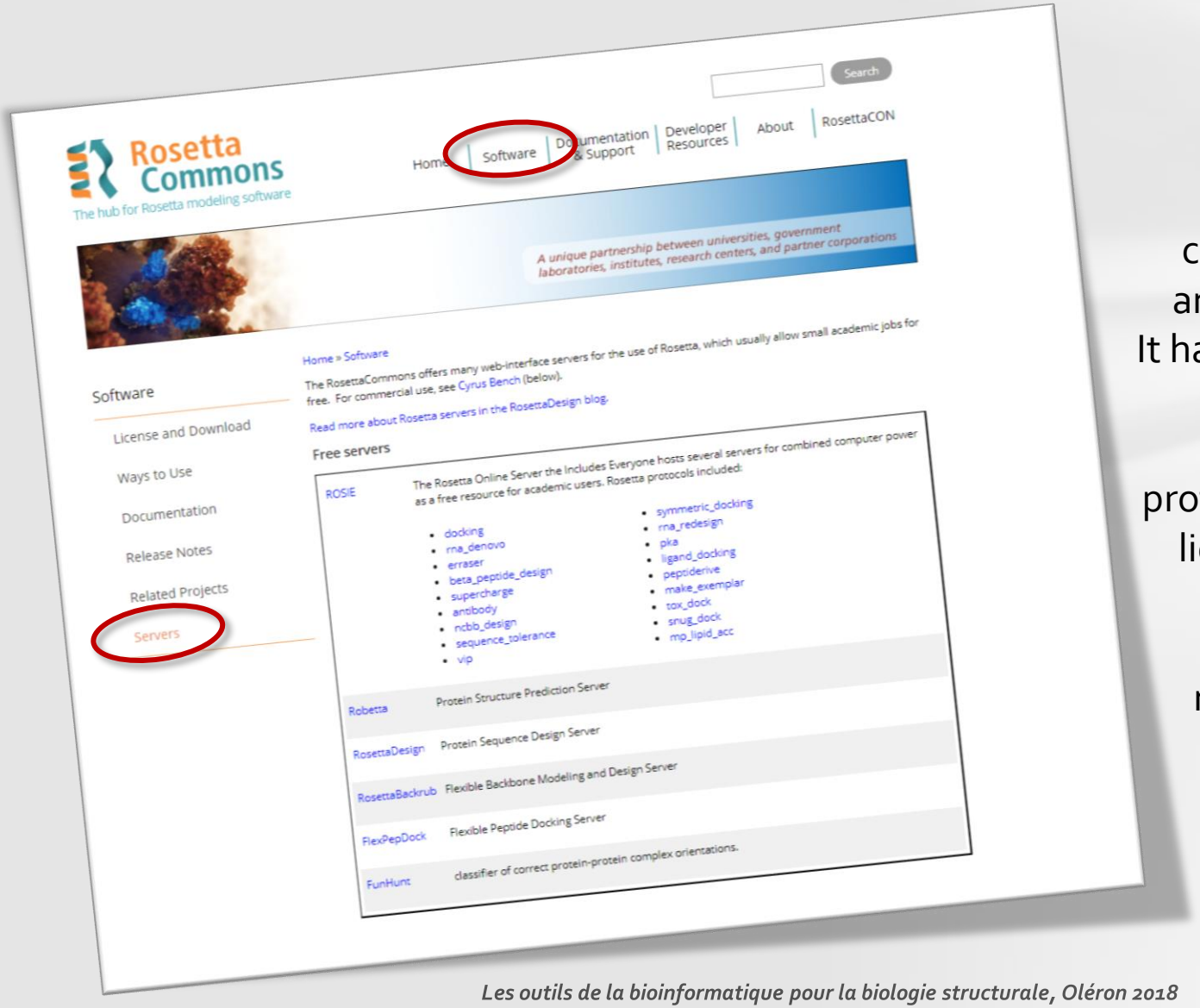
ls of

, and are

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c4chtA	Alignment		100.0	40	**PDB header:**cell cycle **Chain:** A: **PDB Molecule:**dna topoisomerase 3-alpha; **PDBTitle:** crystal structure of the human topoisomerase iii alpha-rmi12 complex with bound calcium ion Run Investigator
2	c2o59B	Alignment		100.0	22	**PDB header:**isomerase/dna **Chain:** B: **PDB Molecule:**dna topoisomerase 3; **PDBTitle:** structure of e. coli topoisomerase iii in complex with an 8-2 base single stranded oligonucleotide. frozen in glycerol3 ph 8.0 Run Investigator
3	d1i7da	Alignment		100.0	22	**Fold:**Prokaryotic type I DNA topoisomerase **Superfamily:**Prokaryotic type I DNA topoisomerase **Family:**Prokaryotic type I DNA topoisomerase Run Investigator
4	c2gajA	Alignment		100.0	24	**PDB header:**isomerase **Chain:** A: **PDB Molecule:**dna topoisomerase i; **PDBTitle:** structure of full length topoisomerase i from thermotoga maritima in2 monoclinic crystal form Run Investigator
5	d1mw9x	Alignment		100.0	27	**Fold:**Prokaryotic type I DNA topoisomerase **Superfamily:**Prokaryotic type I DNA topoisomerase **Family:**Prokaryotic type I DNA topoisomerase Run Investigator
6	c4ddvA	Alignment		100.0	22	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**reverse gyrase; **PDBTitle:** thermotoga maritima reverse gyrase, triclinic form Run Investigator

Homology modeling

Rosetta Software: The premier suite for macromolecular modeling



The Rosetta software suite includes algorithms for computational modeling and analysis of protein structures. It has enabled notable scientific advances in computational biology, including de novo protein design, enzyme design, ligand docking, and structure prediction of biological macromolecules and macromolecular complexes.

Model validation

SAVES	http://nihserver.mbi.ucla.edu/SAVES/
ERRAT	http://nihserver.mbi.ucla.edu/ERRAT/
VERIFY3D	http://servicesn.mbi.ucla.edu/Verify3D/
ProSA	https://prosa.services.came.sbg.ac.at/prosa.php/
RAMPAGE	http://mordred.bioc.cam.ac.uk/~rapper/rampage.php
ANOLEA	http://melolab.org/anolea/

ERRAT plot gives which residue is creating problem in the model

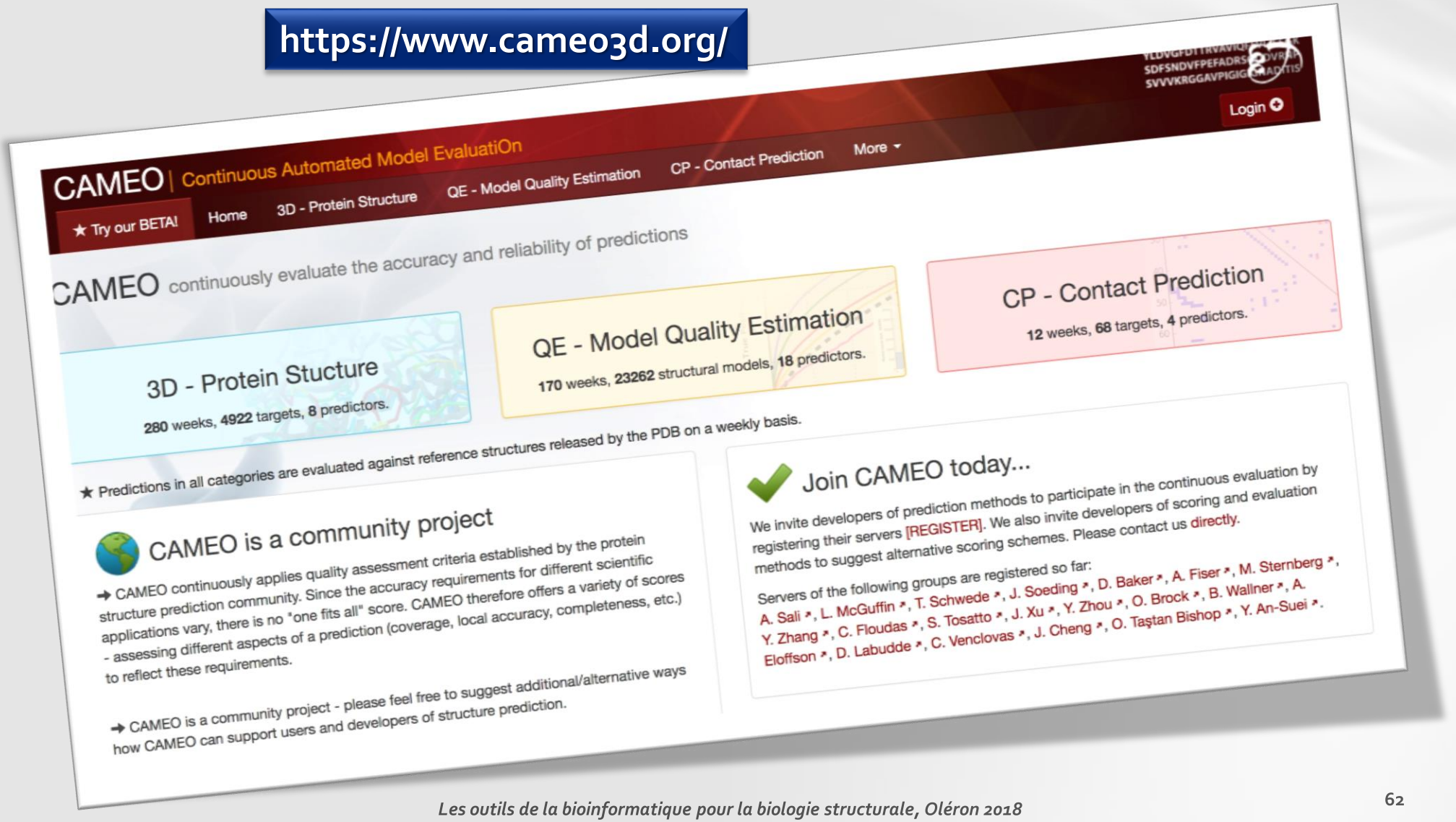
PROSA2 really helps in validating the protein structure by comparing the global energy profile of model to energy profiles of a non redundant set of good quality models

Loop modeling using ModLoop module (Modeller)

Model validation

Continuous Automated Model EvaluatiOn

https://www.cameo3d.org/



Ab initio modeling

Full-chain protein structure prediction server

http://robetta.bakerlab.org/

Robetta provides both *ab initio* and comparative models of protein domains

- Domains without a detectable PDB homolog are modeled with the Rosetta de novo protocol

Simons et al. (1997) J Mol Biol. 268:209-225, Bradley et al. (2005) Science 309, 1868-71

- Comparative models are built from template PDBs detected and aligned using locally installed versions of **HHSEARCH/HHpred**, **RaptorX**, and **Sparks-X**. Alignments are clustered and comparative models are generated using the **RosettaCM** protocol
- The procedure is fully automated

Robetta uses the Rosetta software package

Introduction : qu'est-ce qu'une structure ?

Informations données par la séquence

Prédiction de la structure

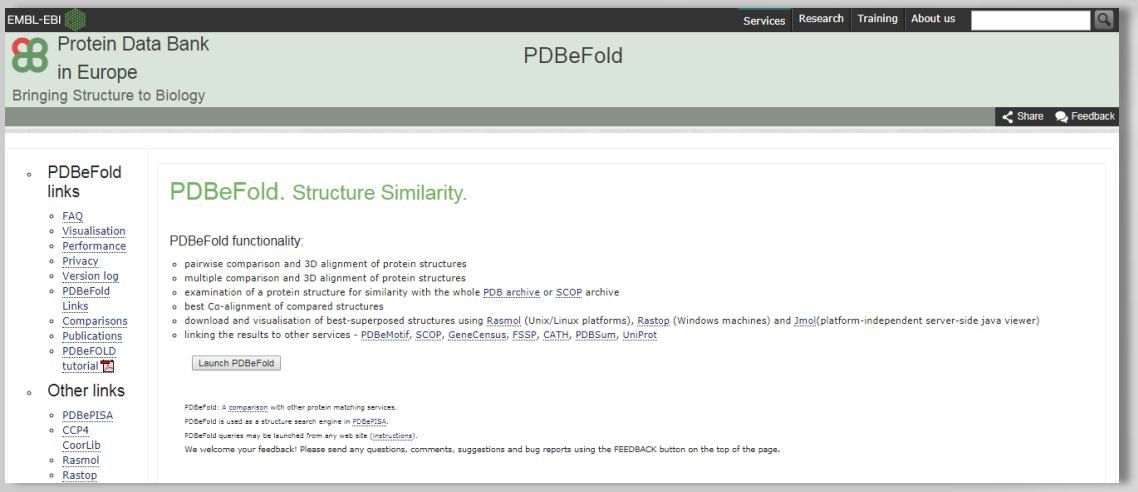
Informations données par la structure

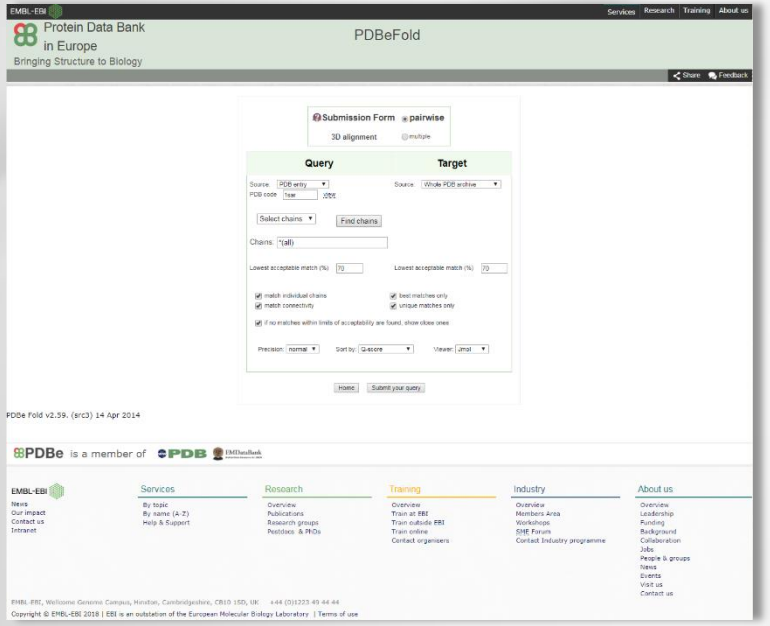
Structural similarity

Though all structural similarity algorithms have a similar goal at their core, there are several different particular applications in biology today that call for somewhat different approaches

- ➡ **Matching of protein structures** (typically measured in RMSD)
- ➡ **Protein structure classification**

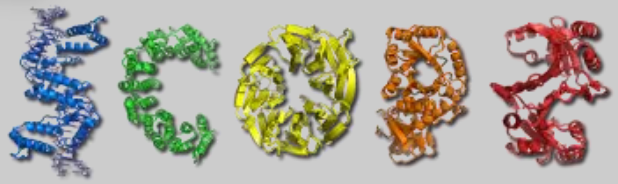
http://www.ebi.ac.uk/msd-srv/ssm/

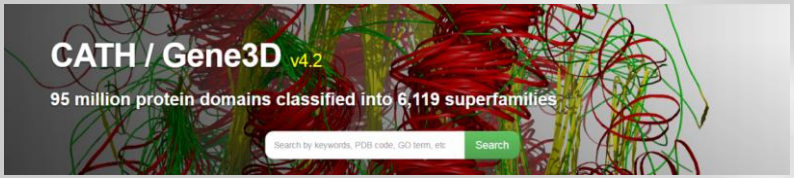




Structural similarity

➡ Protein structure classification

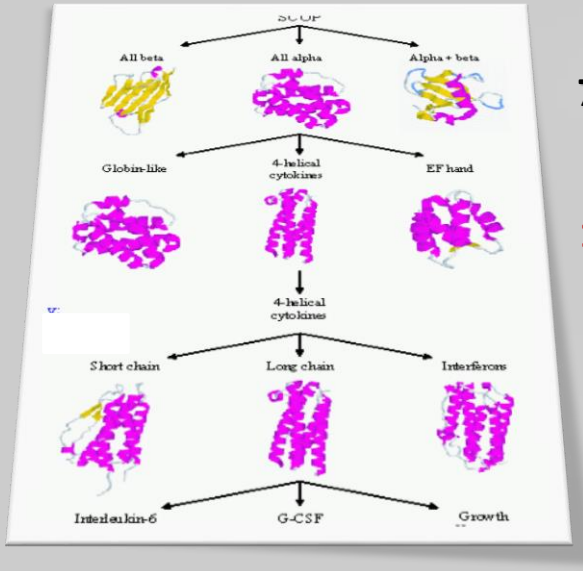




http://scop2.mrc-lmb.cam.ac.uk/

http://www.cathdb.info/

Phylogenetic tree of all the existing folds

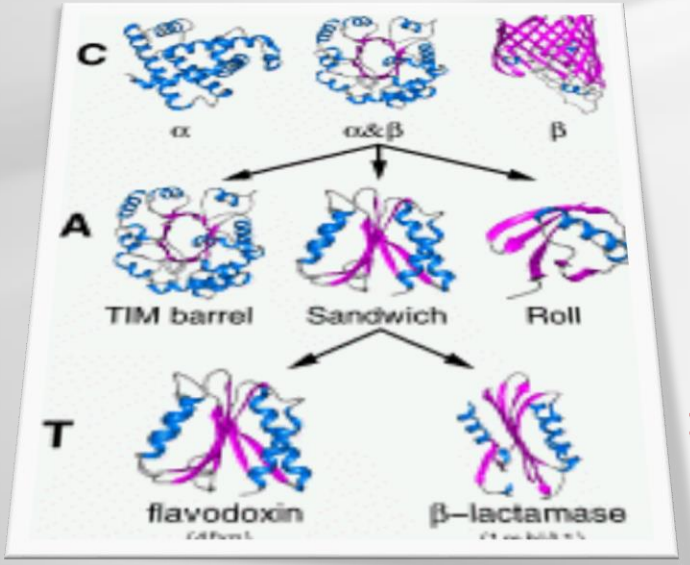


7 Classes

1395 Plis

1962 Superfamilles

3902 Familles



3 Classes

40 Architectures

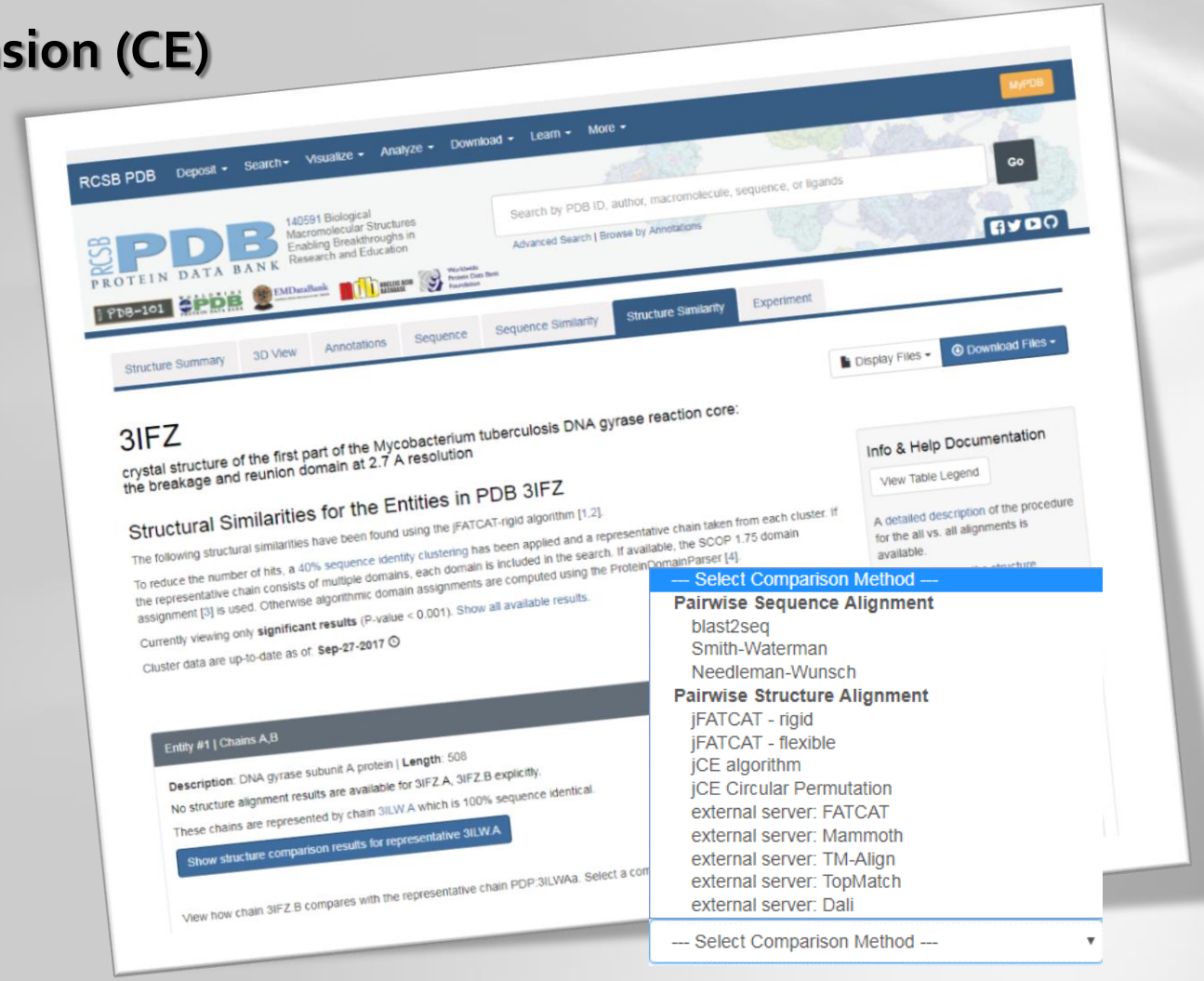
1400 Topologies

2700 superfamilles

Structural similarity

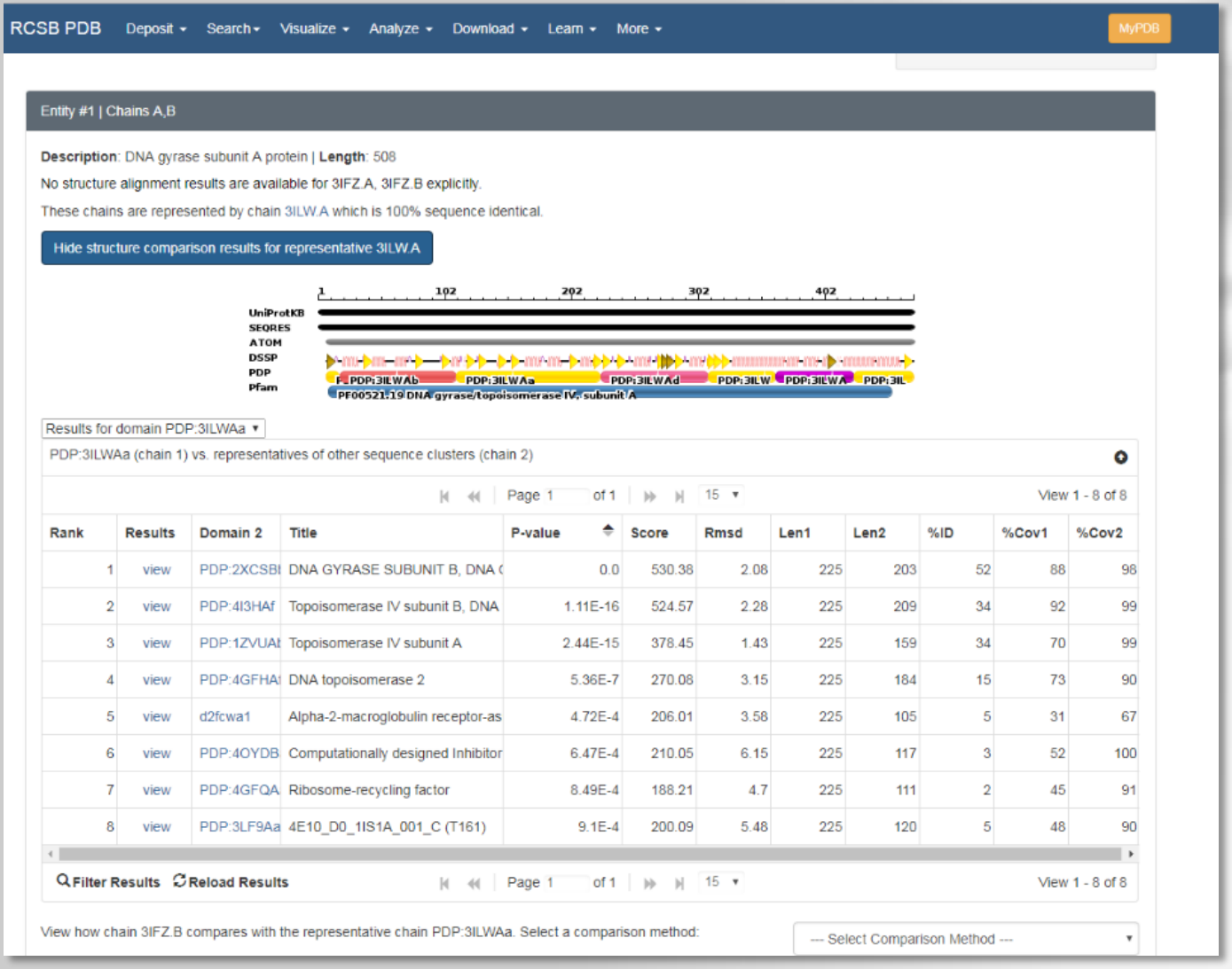
➡ Matching of protein structures

- DALI
- Combinatorial extension (CE)
- GANGSTA+
- MAMMOTH
- FATCAT
- ProBiS
- RAPIDO
- SABERTOOTH
- SSAP
- Spalign
- TOPOFIT
- SSM
- TM-Align
- TopMatch



Structural similarity

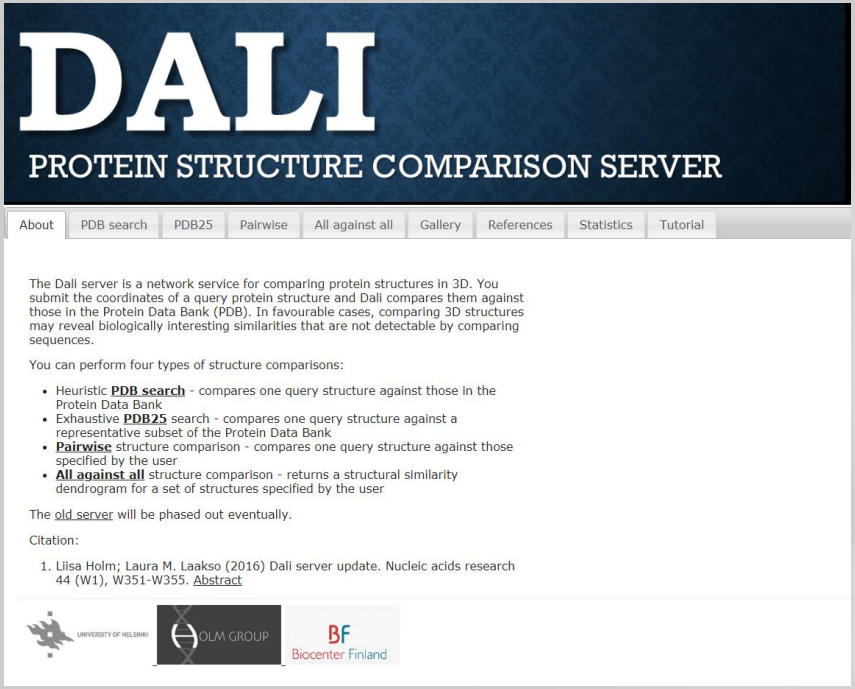
➡ Matching of protein structures



Structural similarity

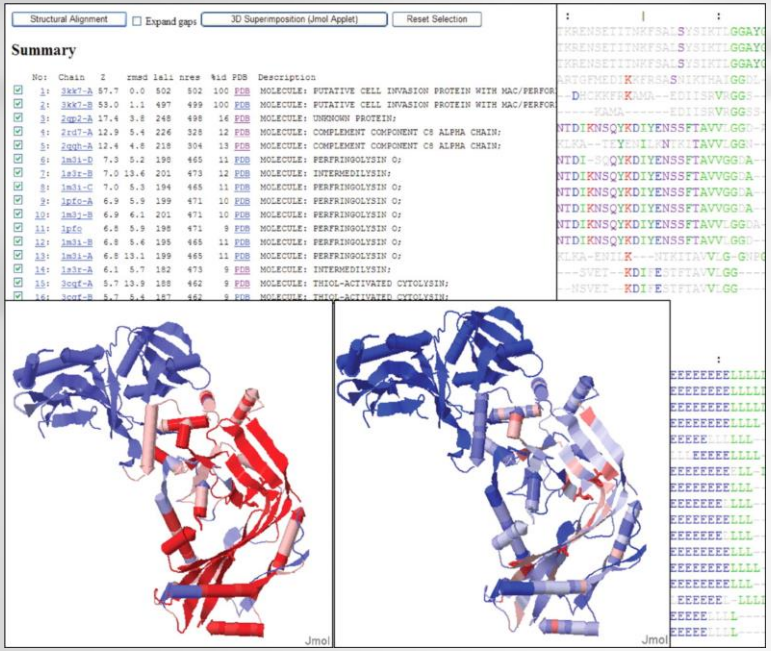
3D comparison - Dali server

http://ekhidna2.biocenter.helsinki.fi/dali/



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

Structural analyses

Calculations or estimation of structural parameters that contributes to protein stability

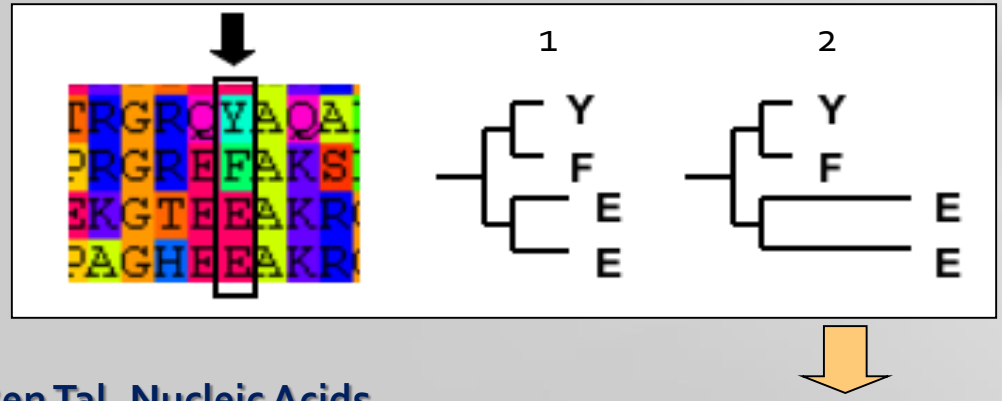
PROPKA	estimation of the pKa values of ionisable aa
VADAR	structure validation server that allows to calculate volumes, accessible surfaces, contact surface, …
MarkUs	analysis and comparison of the structural and functional properties of proteins
HotSpot Wizard	a tool for automatic identification of hot spot sites for engineering of substrate specificity, activity or enantioselectivity of enzymes
FoldX	a protein design algorithm that uses an empirical force field. It can determine the energetic effect of point mutations as well as the interaction energy of protein complexes (including Protein-DNA) ***Stability***: prediction of free energy changes between alternative structures

Analysis of the conservations

ConSurf : Analyse automatique des conservations

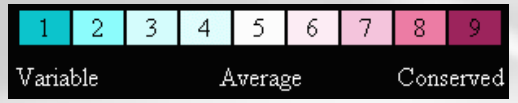
http://consurf.tau.ac.il/

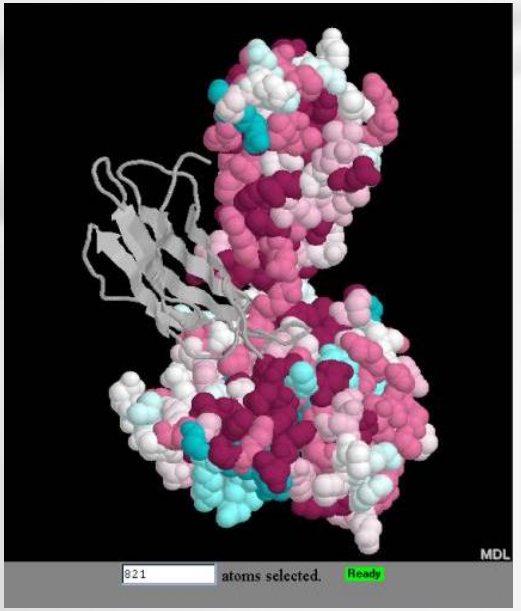
Méthode dont la sensibilité est accrue par une prise en compte des vitesses d'évolution à chaque position
(Pupko et al Bioinfo 2002)



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

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



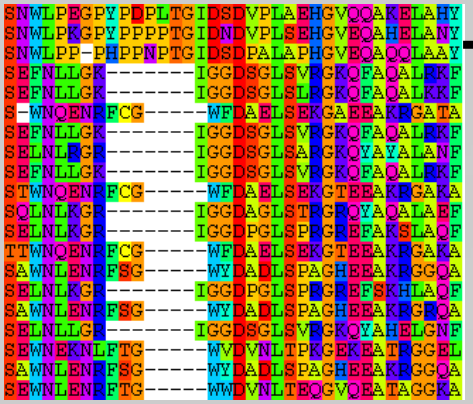


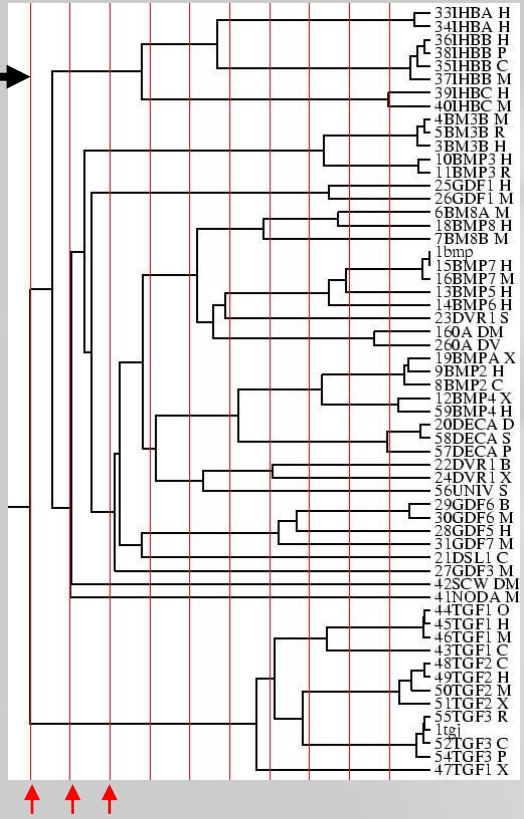
Score de conservation dans la colonne des facteurs B du fichier pdb

Structure and evolution

Analyse des traces évolutives

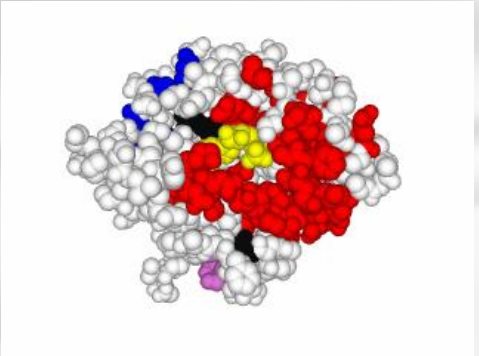
http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html





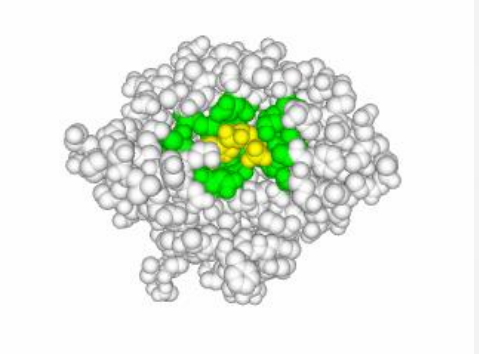
Analyse de l'arbre phylogénétique par partition de classes

« Evolutionnary Traces »



Patchs colorés = classes d'aa

Epitope fonctionnel



http://lichtargelab.org/

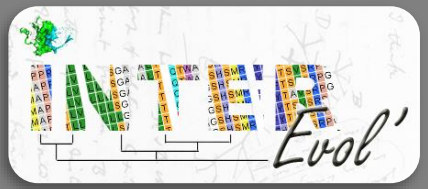
Yao, H., …and O. Lichtarge (2003) J Mol Biol. 326:255-261

Structure and evolution

Principe de la CoEvolution d'Interface

Andreani et al , Plos Comp Biol (2012) & Bioinformatics (2013)

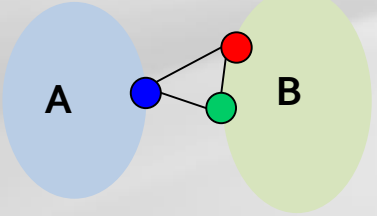
Analyse Statistique → Extraction des caractéristiques → Score de Docking

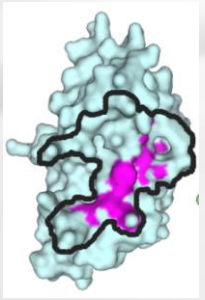


Interfaces with Evolutionary information

InterEvScore : Discriminate co-evolved interfaces

Contacts multi-domaines





Patches apolaires

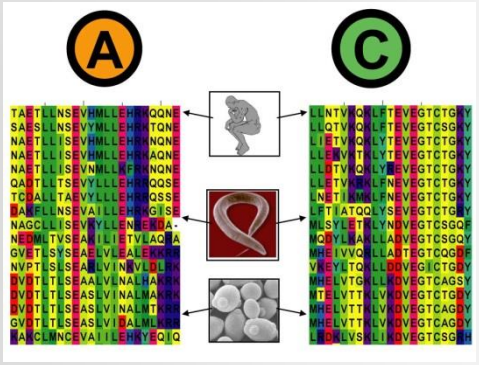
InterEvol:

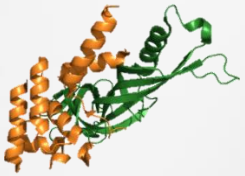
~18,000 non-redundant interfaces
among which
~4,000 heteromeric interfaces

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

http://biodev.cea.fr/interevol

Information d'évolution



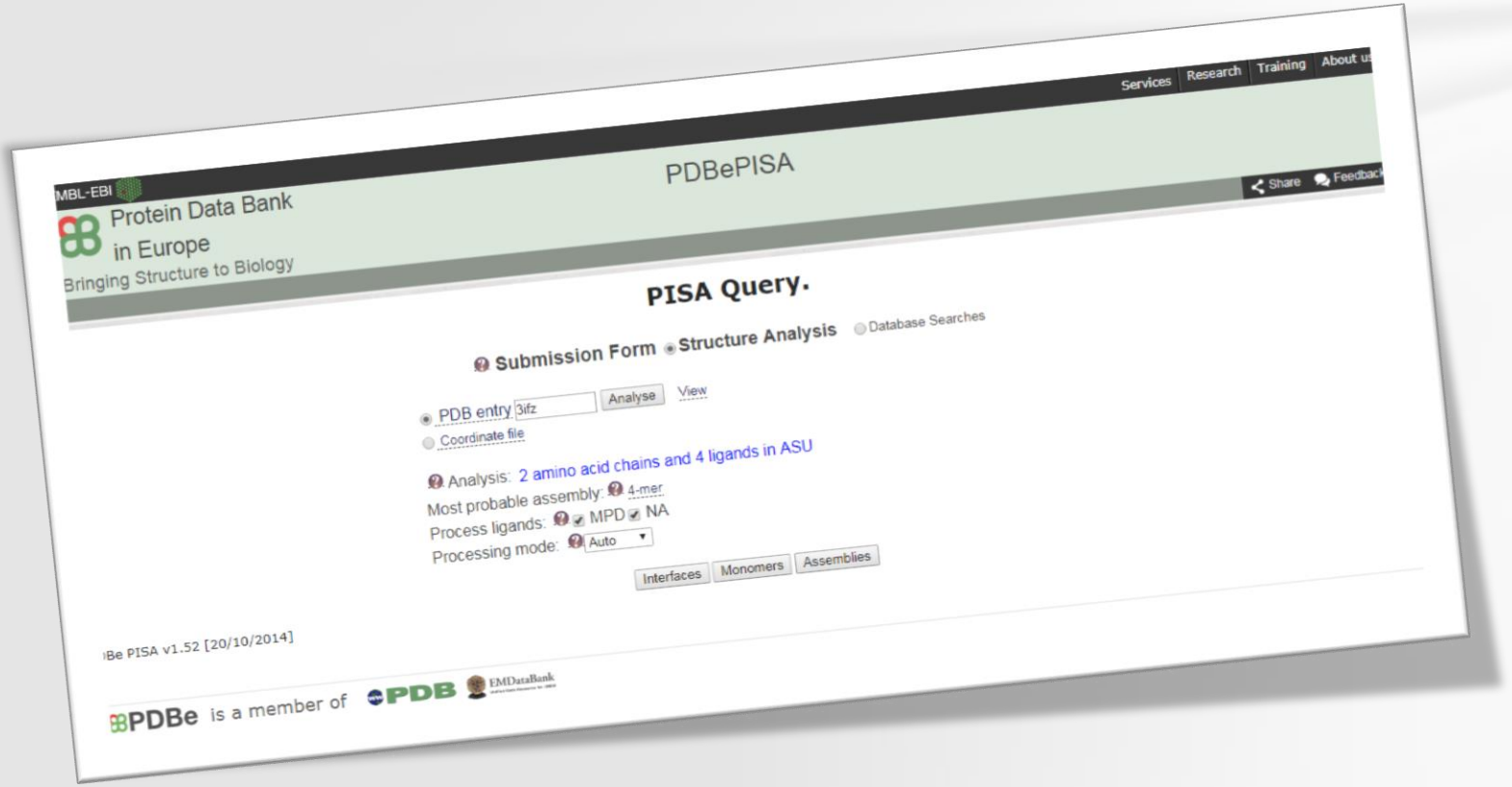


Structure and interactions

PDBePISA is an interactive tool for the exploration of macromolecular interfaces

http://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver/

Proteins, Interfaces, Structures and Assemblies



Structure and interactions

With PDBePISA, you can:

- Retrieve pre-calculated results for the whole PDB archive
- Calculate results interactively for structures uploaded as PDB or mmCIF files that include
  - structural and chemical properties of macromolecular surfaces and interfaces
  - probable quaternary structures (assemblies), their structural and chemical properties and probable dissociation pattern
- search the PDB archive for particular interfaces formed by structural homologs
- search the PISA database of pre-calculated results using a wide range of options, such as
  - multimeric state
  - symmetry number
  - space group
  - accessible/buried surface area
  - free energy of dissociation
  - presence/absence of salt bridges and disulphide bonds
  - homomeric type
  - ligands
  - keywords
- assess the significance (biological role) of macromolecular interfaces
- download and visualise structures, interfaces and assemblies using Rasmol (Unix/Linux platforms), Rastop (MS Windows machines) and Jmol (platform-independent server-side java viewer)

Structure and interactions

Example: 3IFZ
Crystal structure of the breakage-reunion domain of the *Mycobacterium tuberculosis* DNA gyrase

Session Map (id=275-G8-JGN)

Start | Interfaces | Interface Search | Monomers | Assemblies

Interfaces in PDB 3ifz crystal.

Space symmetry group: C 1 2 1. Resolution: 2.70 Å

CRYSTAL STRUCTURE OF THE FIRST PART OF THE MYCOBACTERIUM TUBERCULOSIS DNA GYRASE REACTION CORE: THE BREAKAGE AND REUNION DOMAIN AT 2.7 A RESOLUTION

Interfaces XML | View | Details | Download | Search

##			Structure 1			×	Structure 2					interface	Δ^iG	Δ^iG	N_{HB}	N_{SB}	N_{DS}	CSS
Id	NN	«»	Range	$^iN_{at}$	$^iN_{res}$	Surface Å^2		Range	Symmetry op-n	Sym.ID	$^iN_{at}$	$^iN_{res}$	Surface Å^2	area, Å^2	kcal/mol	P-value				
1	1	◉	B	188	50	22775	◊	A	x,y,z	1_555	194	49	25080	1941.4	-16.5	0.061	25	12	0	0.419
2	2	○	A	128	37	25080	◊	A	-x,y,-z	2_555	128	37	25080	1227.7	-19.8	0.006	12	4	0	0.394
3	3	○	B	66	20	22775	x	B	-x-1/2,y-1/2,-z	4_445	74	23	22775	646.1	3.1	0.761	5	6	0	0.000
4	4	○	B	37	11	22775	◊	B	-x,y,-z+1	2_556	37	11	22775	293.3	-5.0	0.131	0	0	0	0.000
5	5	○	B	30	12	22775	◊	A	-x,y,-z	2_555	29	8	25080	290.4	-1.9	0.354	3	3	0	0.042
6	6	○	[MPD]B:509	7	1	259	f	B	x,y,z	1_555	26	14	22775	186.3	-12.1	0.603	1	0	0	0.179
7	7	○	[MPD]A:509	7	1	258	f	A	x,y,z	1_555	30	13	25080	183.1	-11.6	0.595	0	0	0	0.355
8	8	○	B	21	9	22775	◊	A	x-1/2,y-1/2,z	3_445	27	10	25080	176.5	1.1	0.687	2	0	0	0.000
9	9	○	A	21	4	25080	x	A	-x+1/2,y-1/2,-z+1	4_546	23	9	25080	163.4	-0.3	0.537	1	4	0	0.000
10	10	○	B	8	2	22775	◊	A	-x-1/2,y-1/2,-z	4_445	6	2	25080	68.9	1.7	0.868	1	5	0	0.000
11	11	○	[NA]A:510	1	1	125	f	A	x,y,z	1_555	12	6	25080	66.0	-8.2	0.000	0	0	0	0.251
	12	○	[NA]B:510	1	1	125	f	B	x,y,z	1_555	12	6	22775	65.6	-8.4	0.000	0	0	0	0.251
													Average:	65.8	-8.3	0.000	0	0	0	0.251
12	13	○	B	9	5	22775	◊	A	x-1/2,y+1/2,z	3_455	9	4	25080	55.7	1.6	0.734	1	2	0	0.000
13	14	○	B	6	3	22775	◊	A	-x,y,-z+1	2_556	5	3	25080	30.0	0.2	0.661	0	0	0	0.000

View | Details | Download | Search

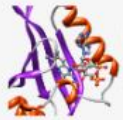
Predicition of interactions

https://zhanglab.ccmb.med.umich.edu/services/

Introduction: COACH is a meta-server approach to protein-ligand binding site prediction. Starting from given structure of target proteins, COACH will generate complementray ligand binding site predictions using two comparative methods, TM-SITE and S-SITE, which recognize ligand-binding templates from the BioLiP database by substructure and binding-specific sequence-profile comparisons. These predictions will be combined with results from other methods (including COFACTOR, FINDSITE and ConCavity to generate final ligand binding site predictions. Users are also allowed to input primary sequence, where I-TASSER will be used to generate 3D models first which are then fed into the COACH pipeline for ligand-binding site prediction.

References:

- Jianyi Yang, Ambrish Roy, and Yang Zhang. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment, Bioinformatics, 29:2588-2595 (2013). [PDF] [Support Information] [Server]



COFACTOR

Structure-based function predictions

Enzyme Commission Gene Ontology Ligand Binding Site

Introduction: COFACTOR is an automated method for biological function annotation of protein molecules, based on protein 3D structures. When user provides a structure model of the target protein, COFACTOR will match the target proteins to the known proteins (templates) in three comprehensive protein function libraries by global and local structure comparisons. Functional insights, including ligand-binding site, gene-ontology term, and enzyme classification, are then derived from the best template proteins of the highest confidence score (C-score). The COFACTOR algorithm was ranked as the best method for ligand-binding site predictions in the community-wide CASP9 experiments.

References:

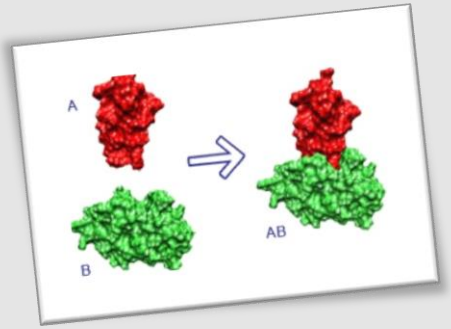
- Ambrish Roy, Jianyi Yang, and Yang Zhang. *COFACTOR: An accurate comparative algorithm for structure-based protein function annotation.* Nucleic Acids Research, 40:W471-W477 (2012). (download the PDF file)
- Ambrish Roy, Yang Zhang. *Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement.* Structure, 20: 987-997 (2012) (download the PDF file and Support Information)
- Chengxin Zhang, Peter L. Freddolino, Yang Zhang *COFACTOR: improved protein function prediction by combining structure, sequence, and protein-protein interaction information.* Nucleic Acids Research, 45: W291-299 (2017). (download the PDF file and Support Information)

Prediction of protein-protein interactions

"***Here one should remember that any protein fails to execute its function unless it interacts with other biomolecules***"

Ito *et al*. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 4569

A comprehensive two-hybrid analysis to explore the yeast protein interactome



Webservers

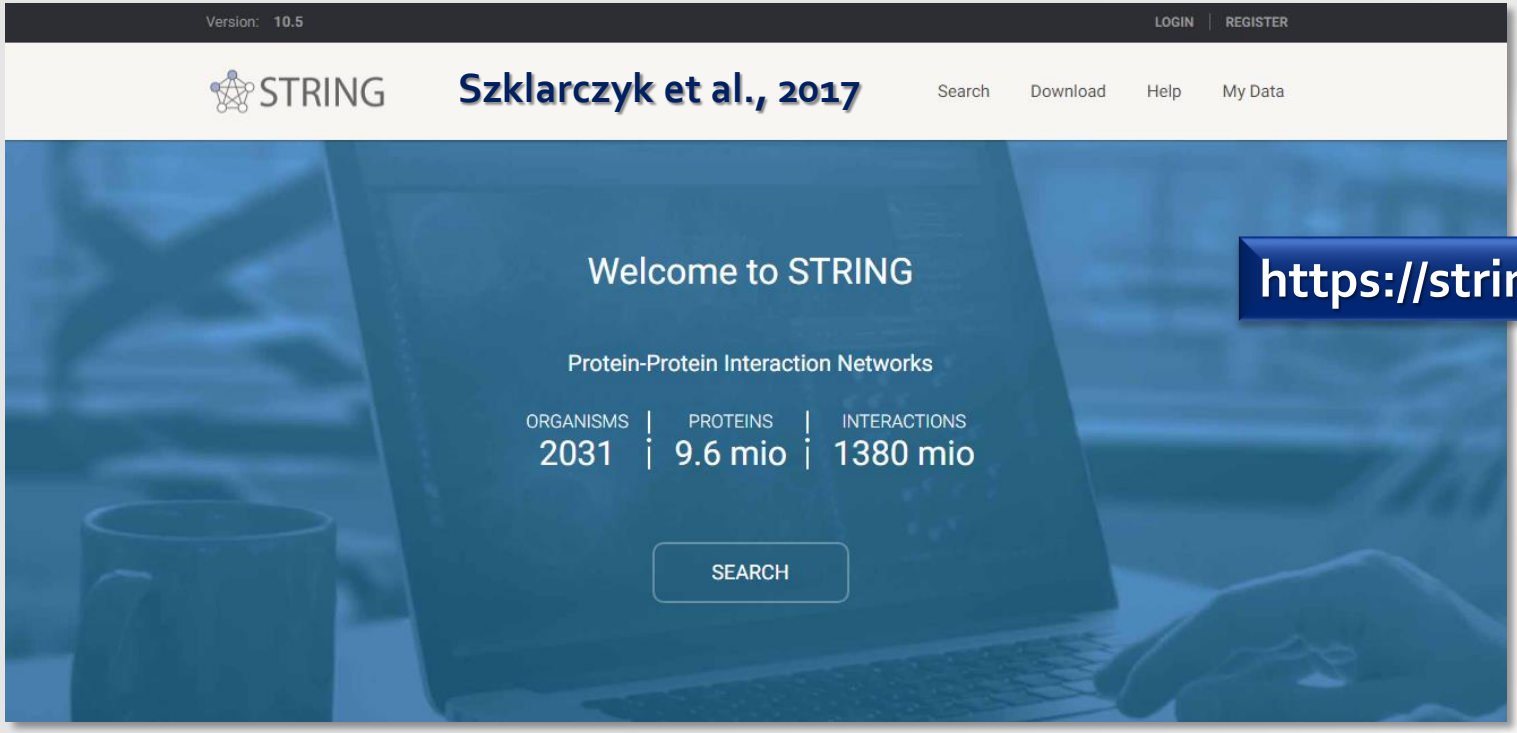
Rosettadock	Prediction of interaction from 2 structural models
Patchdock	Docking based on surface complementarity , easy to use
Firedock	docking protein-protein, easy to use
ClusPro	tops the competition in the latest rounds of CAPRI experiment
Zdock	rigid-body search of docking orientations

Stand-alone

HADDOCK	Docking with possibility to implement experimental data (mutagenesis, cross-linking, NMR chemical shift, …)
HEX	protein-protein docking, webserver also exists

Prediction of protein-protein interactions

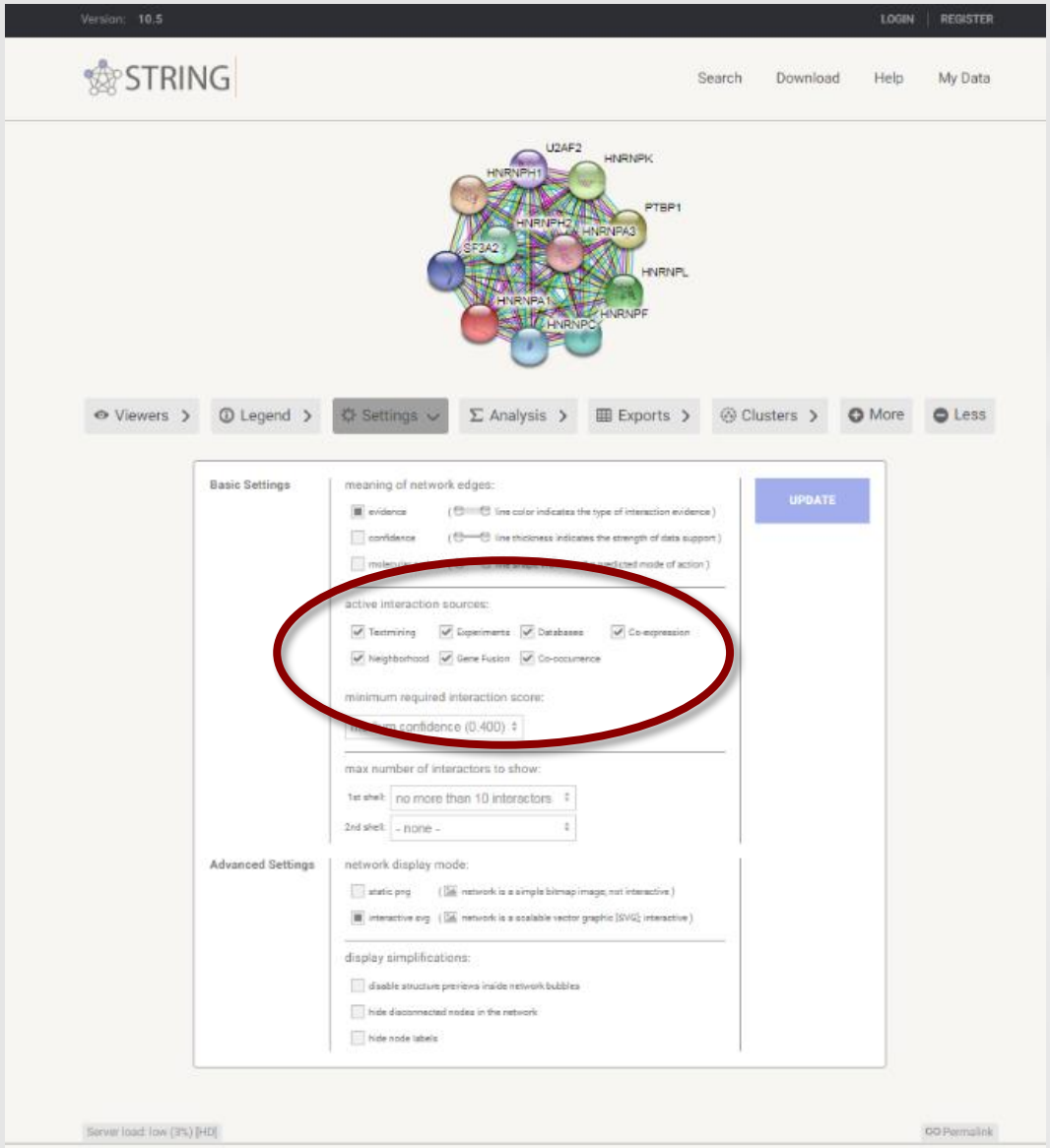
STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a biological database and web resource of known and predicted protein-protein interactions



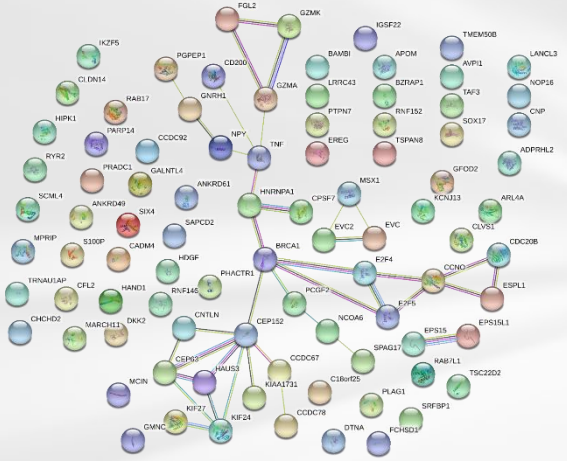
https://string-db.org/

The STRING database contains information from numerous sources, including experimental data, computational prediction methods and public text collections. The resource also serves to highlight functional enrichments in user-provided lists of proteins, using a number of functional classification systems such as GO, Pfam and KEGG. The latest version 10.0 contains information on about 9.6 million proteins from more than 2000 organisms.

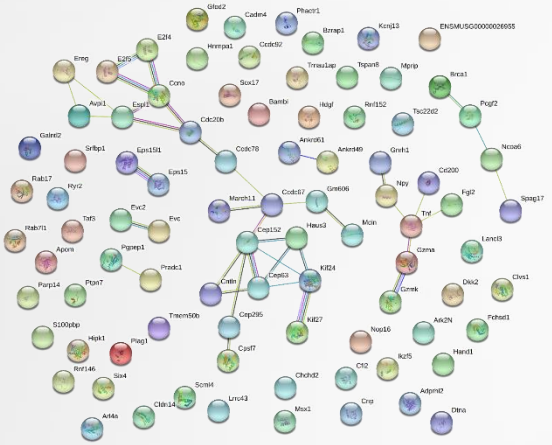
Prediction of protein-protein interactions



Human network



Mouse network

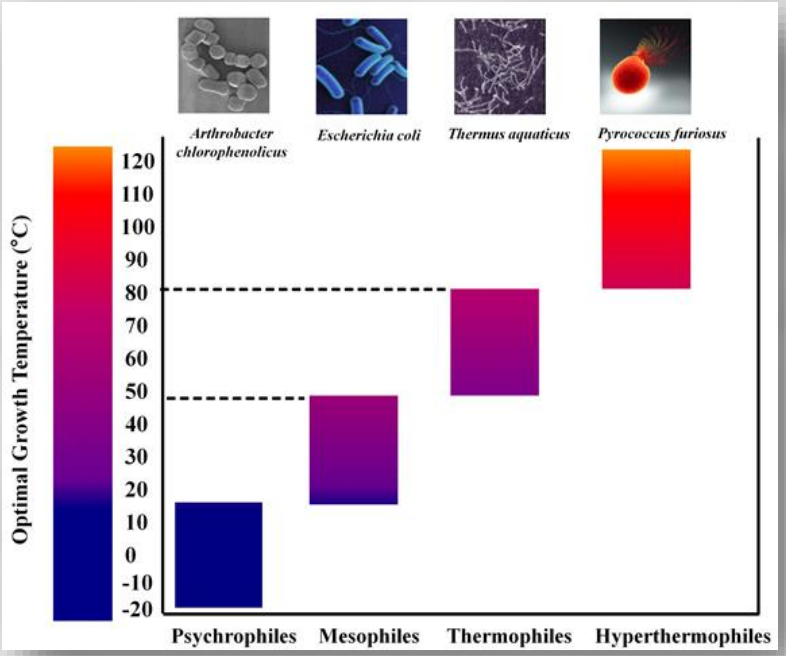


Protein thermostability

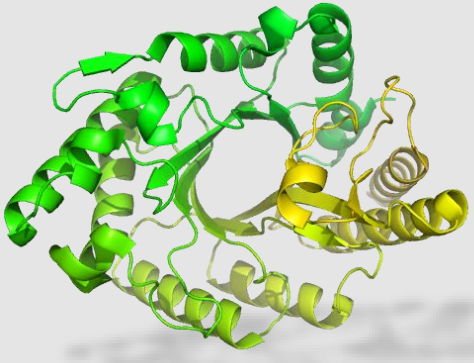
Effect of the temperature on protein stability

A powerful method is the comparison of mesophilic and thermostable homologous proteins

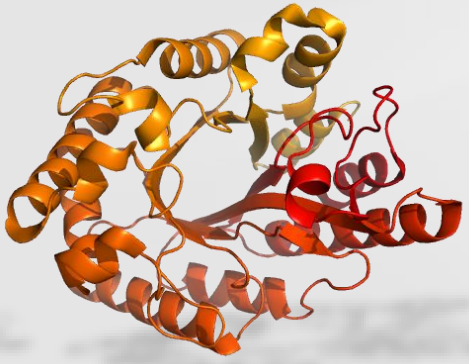
- Presence of **extra hydrogen bonds** and **salt bridges** in thermostable proteins
  the protein structure is more resistant to unfolding
- Other factors are **compactness** of protein structure, **oligomerization** and **interaction strength** between subunits



How to increase the thermostability of target proteins ?



mesophile



thermophile

Design stabilizing mutations

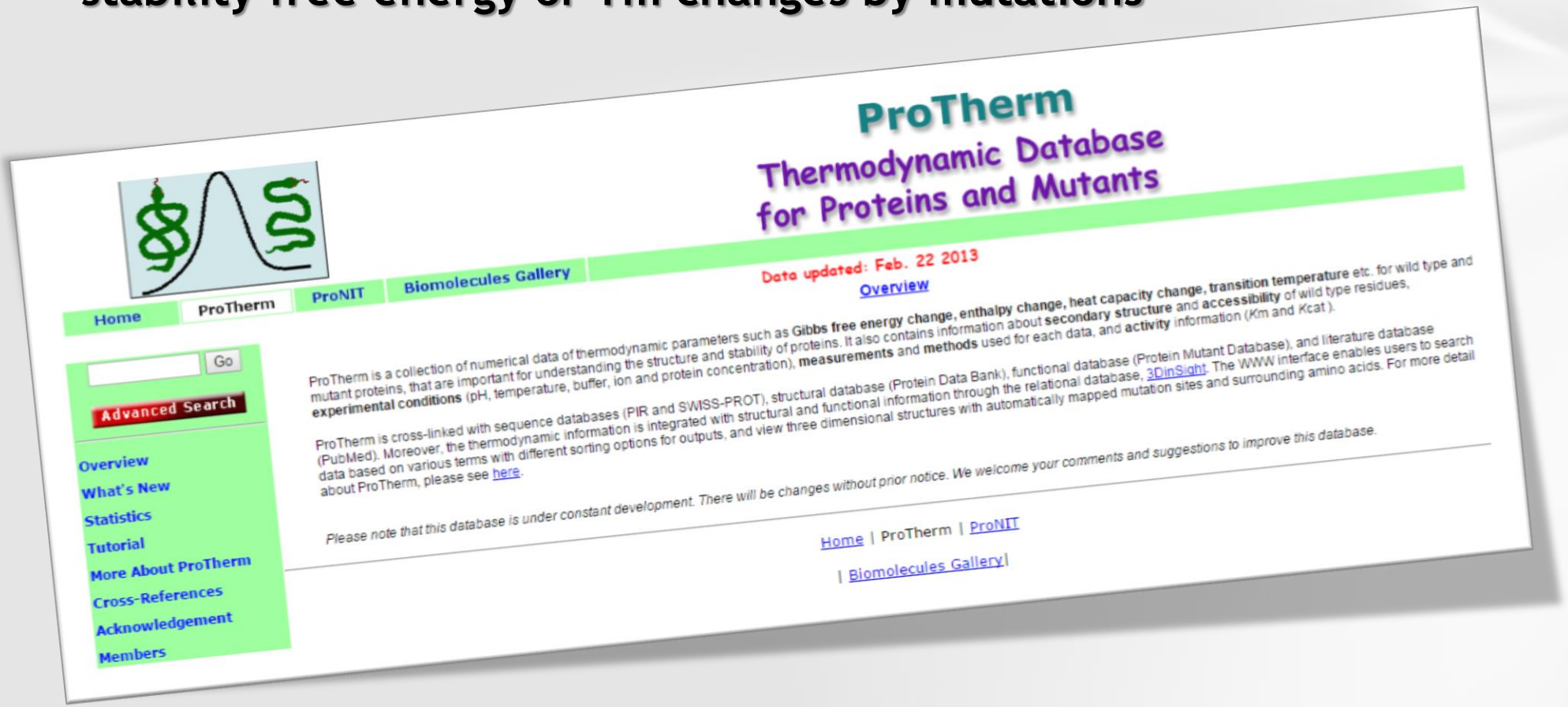
- mutations which truncate loops
- increase salt bridges or hydrogen bonds
- introduced disulfide bonds

Ligand binding can increase the stability

Effect of mutations on protein stability

ProTherm **http://www.abren.net/protherm**

The reference database for experimentally determined protein stability free energy or Tm changes by mutations

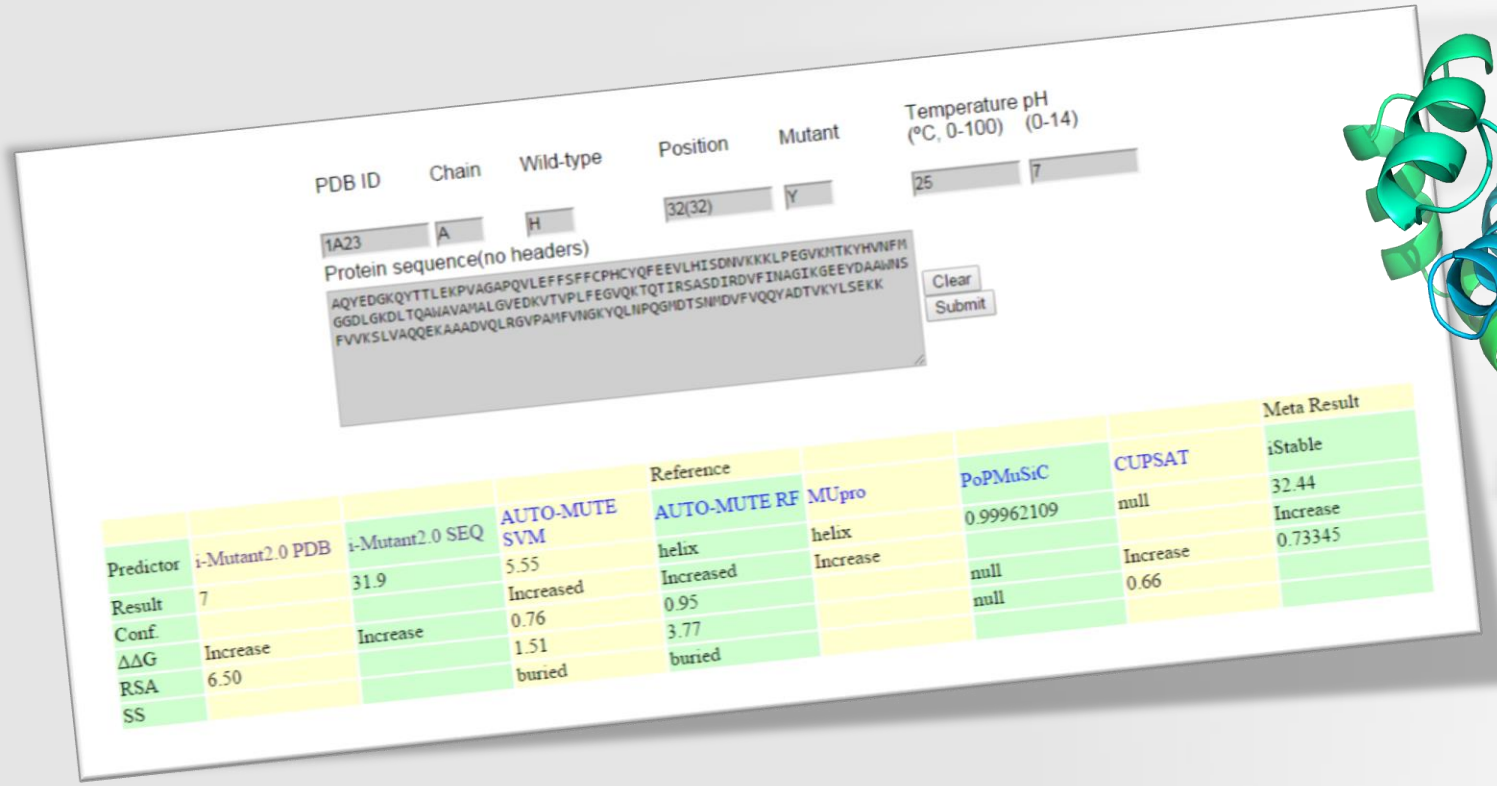


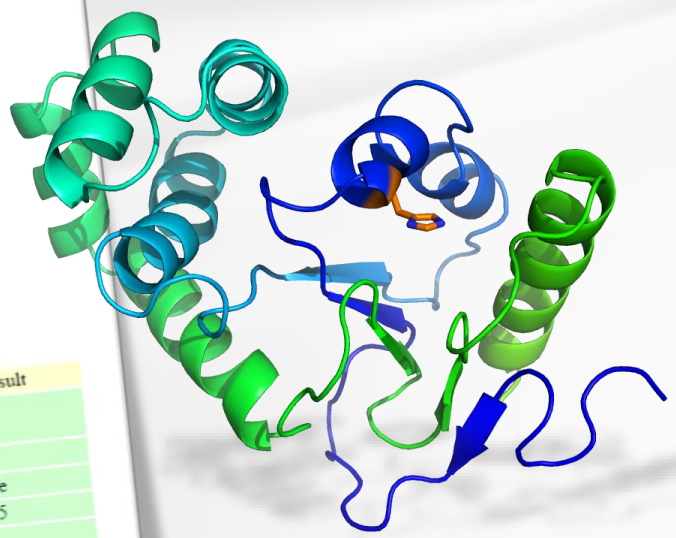
Effect of mutations on protein stability

iStable

http://predictor.nchu.edu.tw/istable/

An integrated predictor constructed by using sequence information and prediction results from different element predictors. In the learning model, iStable adopted the support vector machine as an integrator, while not just choosing the majority answer given by element predictors



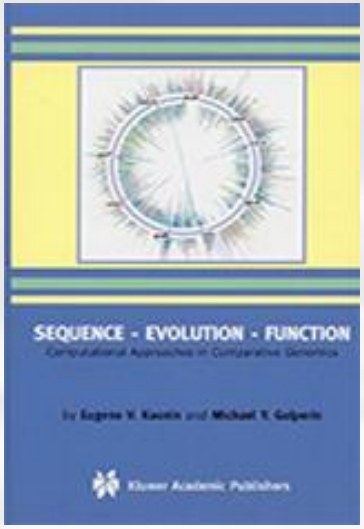


***E. coli* DsbA**

Conclusion

- Always perform a multiple alignment of complete sequence before structural studies
- Try to extract the maximum of information from the sequence
- Doing a model by homology always helps before structural studies
- This course is clearly non-exhaustive…

Bibliography

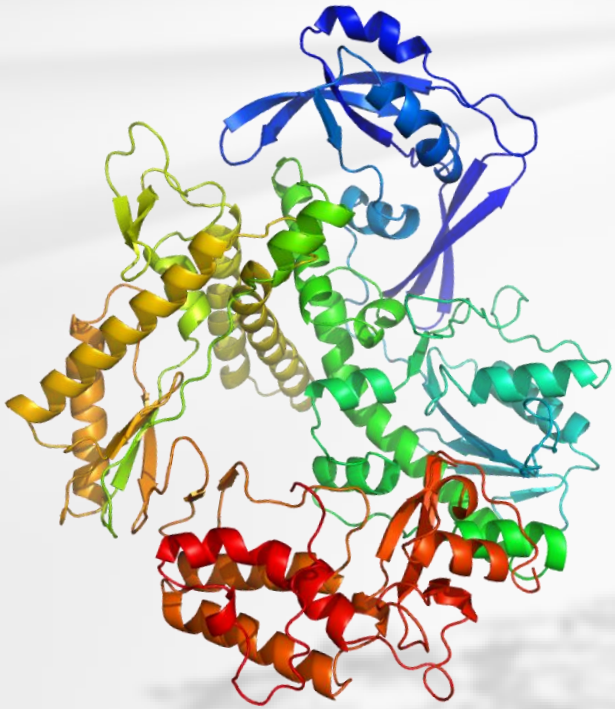


- Koonin EV, Galperin MY. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics

- Protein sequence comparison and Protein evolution Tutorial

  http://www.people.virginia.edu/~wrp/papers/ismb2000.pdf

It’s finished !



Marie-Hélène Le Du
Jessica Andreani
Raphaël Guérois

CEA Saclay, Institut Joliot, Gif sur Yvette