

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
PARIS 7
DIDEROT

US-PC
Université Sorbonne
Paris Cité


Institut Pasteur
Unité de Microbiologie Structurale

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

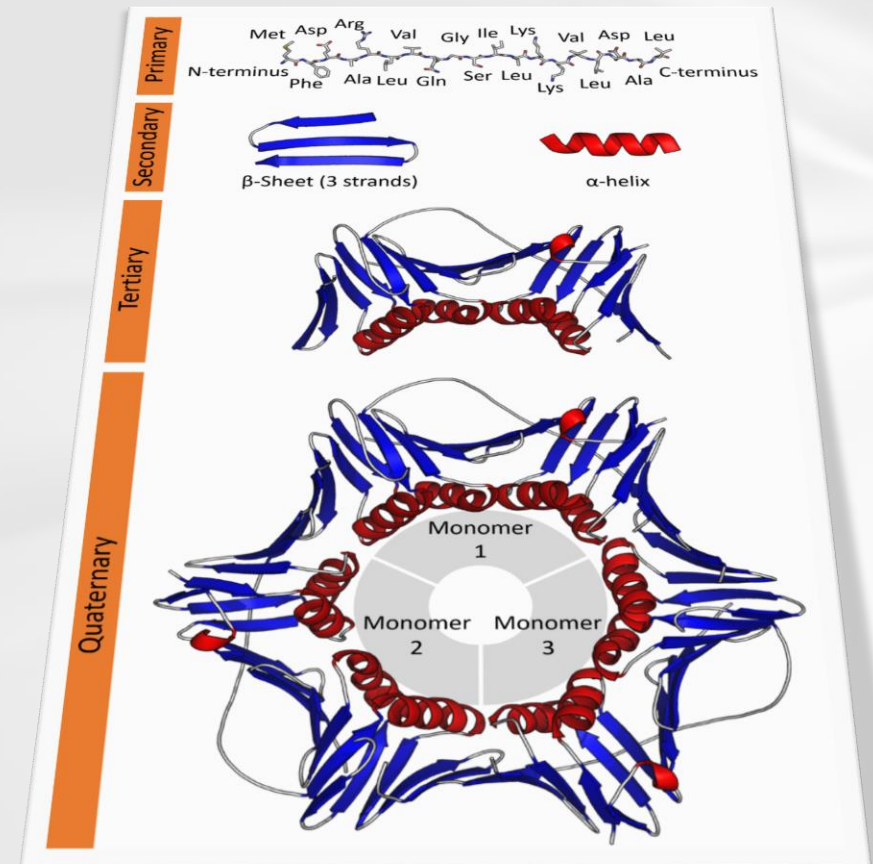
Informations données par la séquence

Prédiction de la structure

Informations données par la structure

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

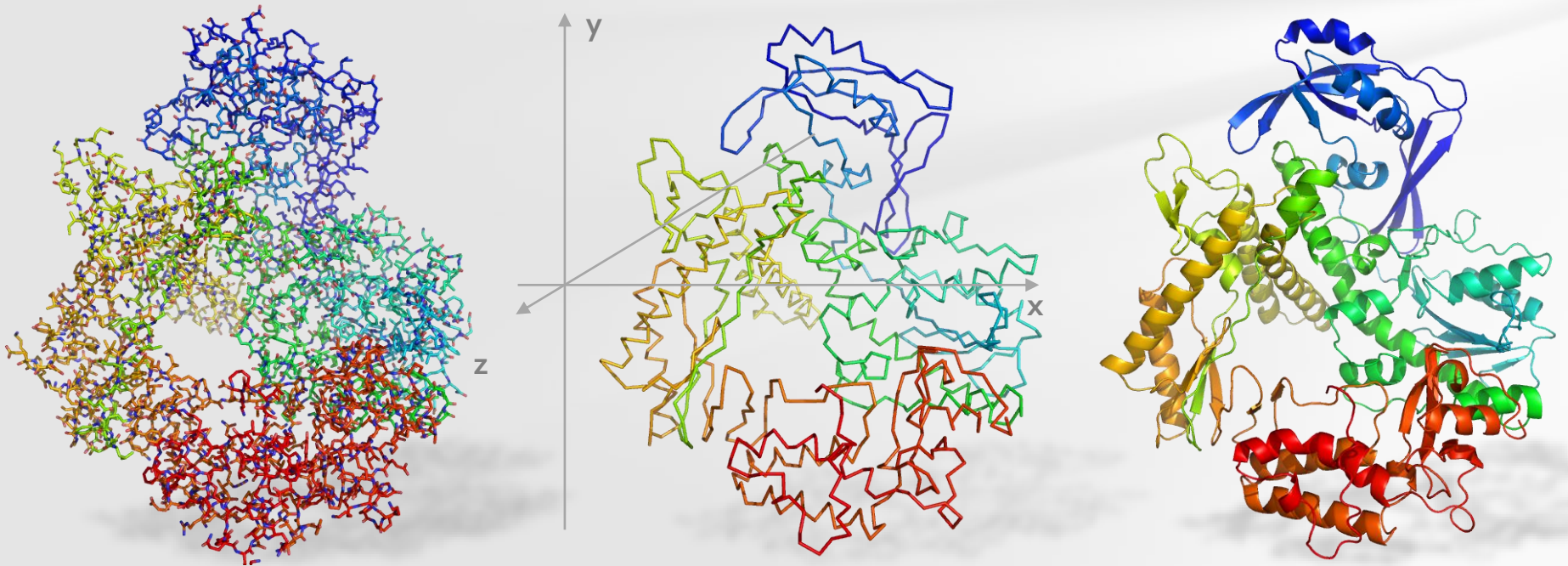
- Séquence (1D)
- Structures secondaires (2D)
- Agencement spatial (3D)
- Assemblages moléculaires (4D)
- Dynamique (5D)



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)



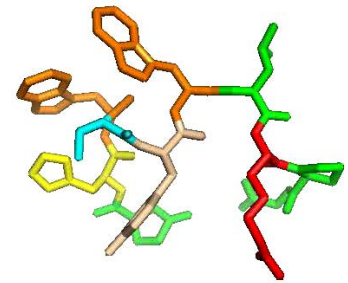
Qu'est-ce que la structure d'une macromolécule ?

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*

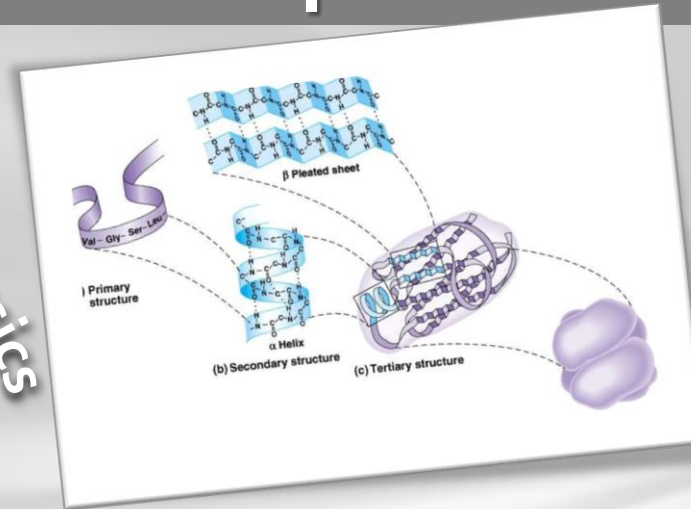
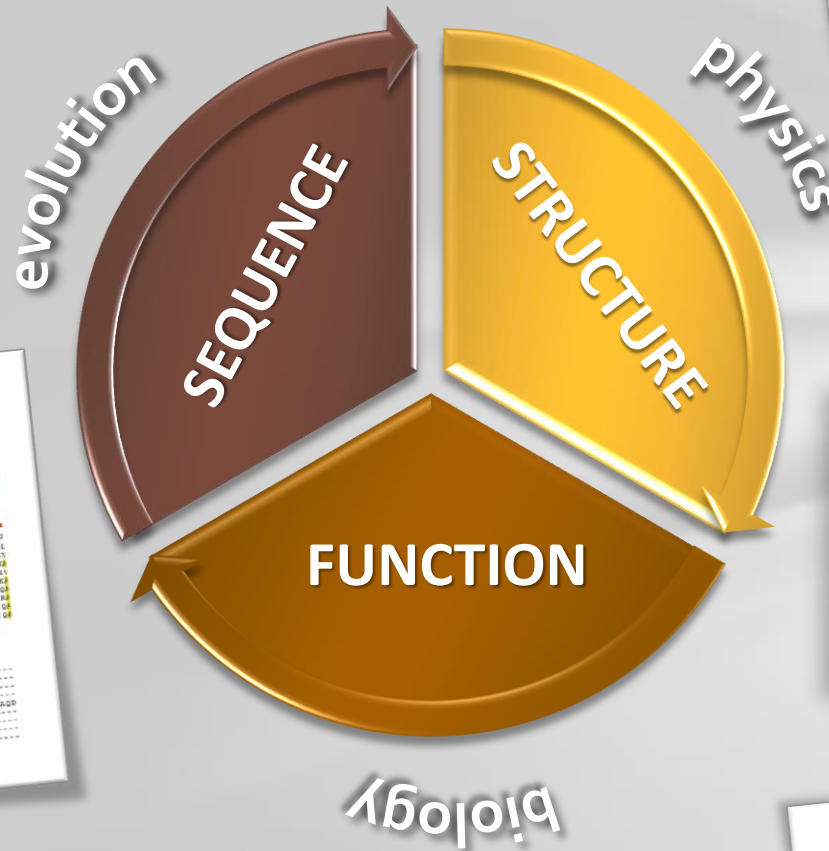


Valéry et al., Nat. Comm, 2015

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

Sequence-structure-function relationships

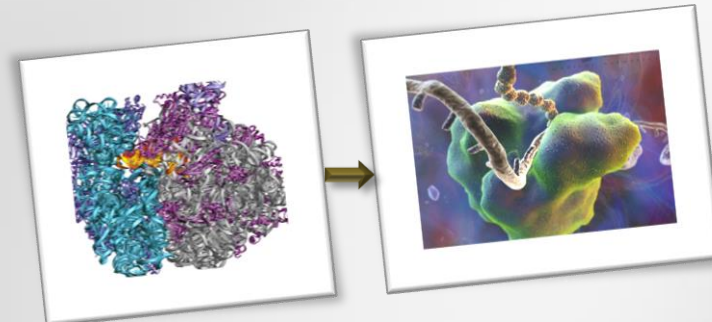
How the sequences evolve based on the phylogenetic distribution



Structure determination based on physical methods
Sequence determines structure



Understand the biological questions in a 3D perspective

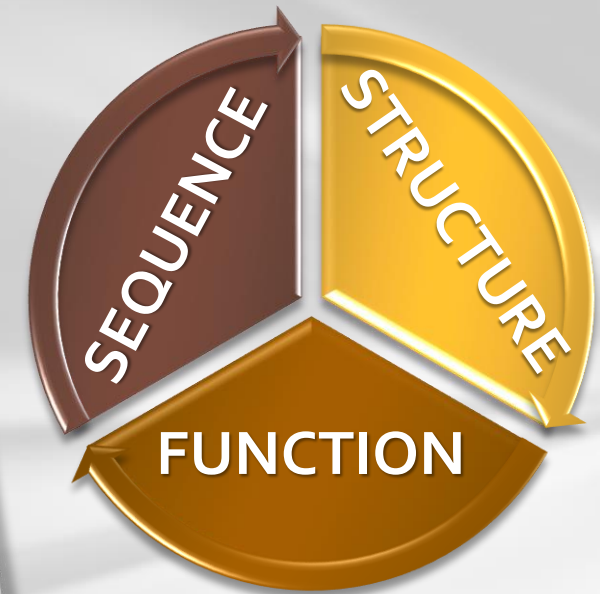


Sequence determines structure

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

Proteins that share sequence identity > 25 % are structurally similar

Homology modeling



Structure
Function

Y AAPLC...NGDPDDLIL...LA...SE VAKVVGIL...
YHTAPLC...YGDPELYA...LA...SE VAKVVGIL...
PNSLSYCQLLPTGEALDYI...ASNE...P...K...
RHE...EALTPERIV LAESAQE YGE...K...
RN...EALTPESVVALA...Y...YGE...K...
RHE...EAM...P...R...E...K...F...K...K...
RHK...KAMNSEAVVRLAESAQDRYGFKDFK...K...

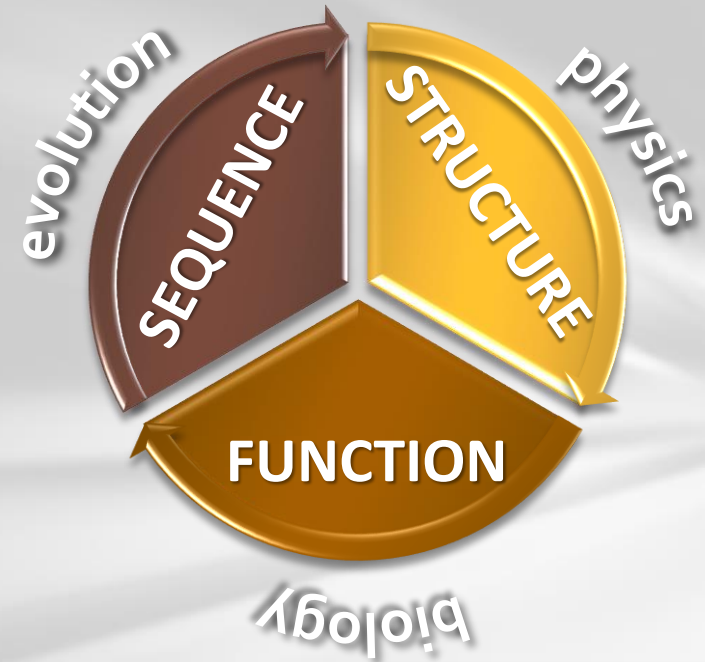
Linkage
Database

... And structure is highly correlated to function

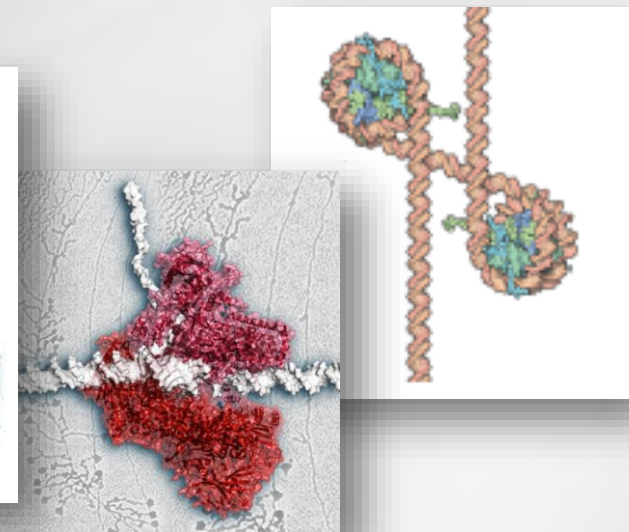
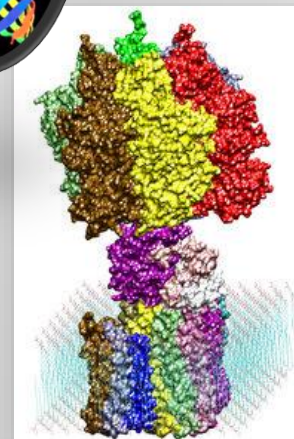
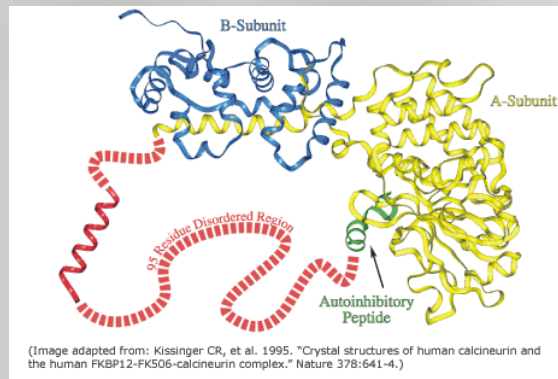
<http://sfl.d.rvbi.ucsf.edu/django/>

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

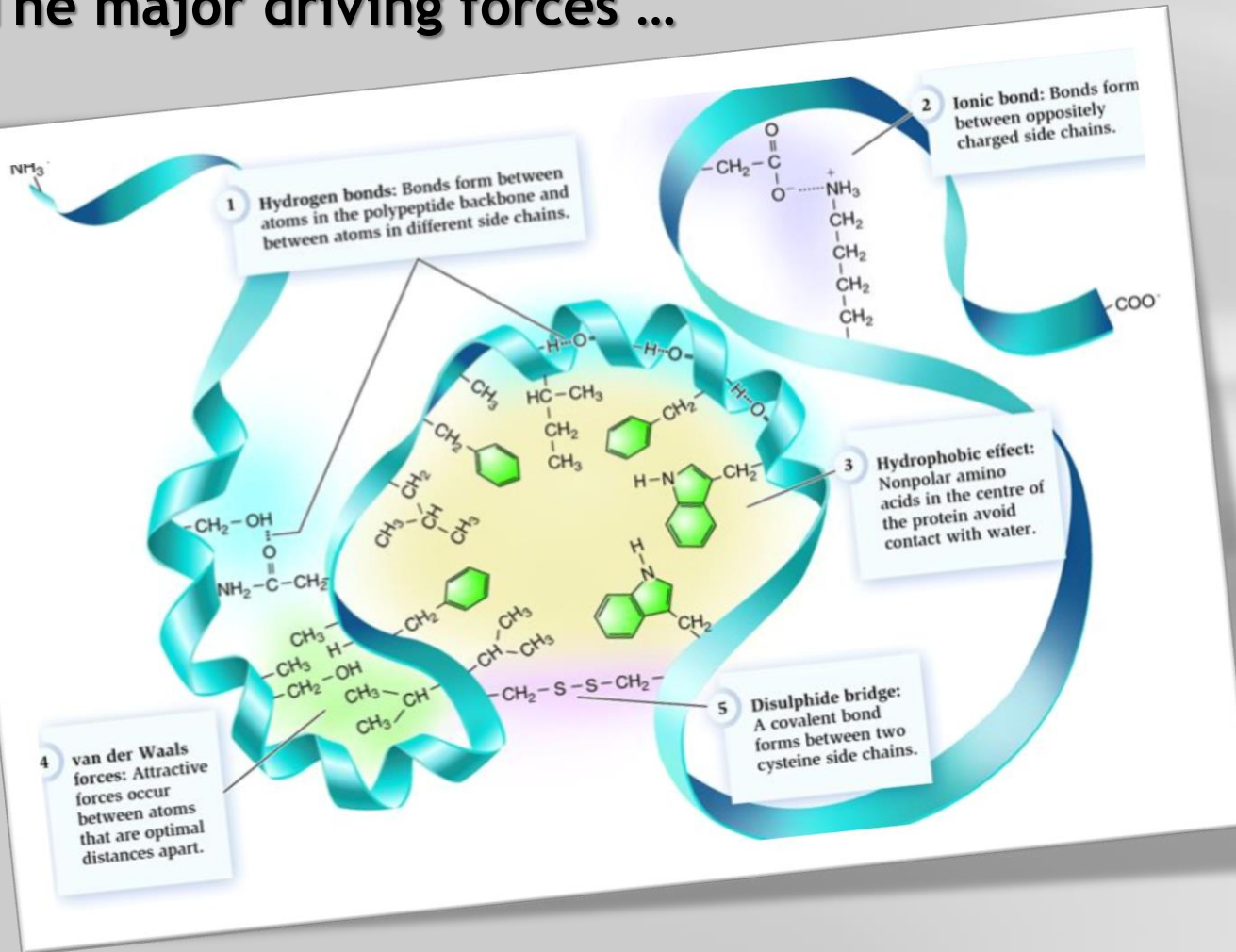


Exception
IDPs & IDRs



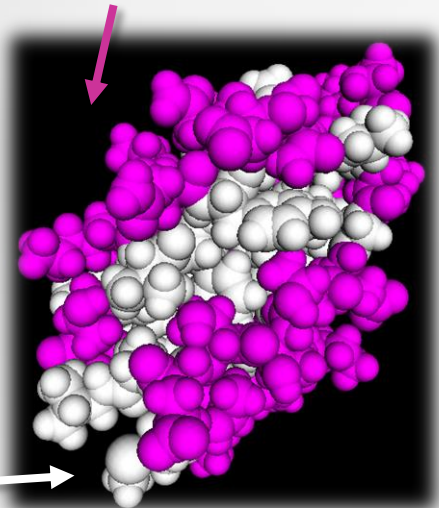
Factors determining protein folding

The major driving forces ...



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

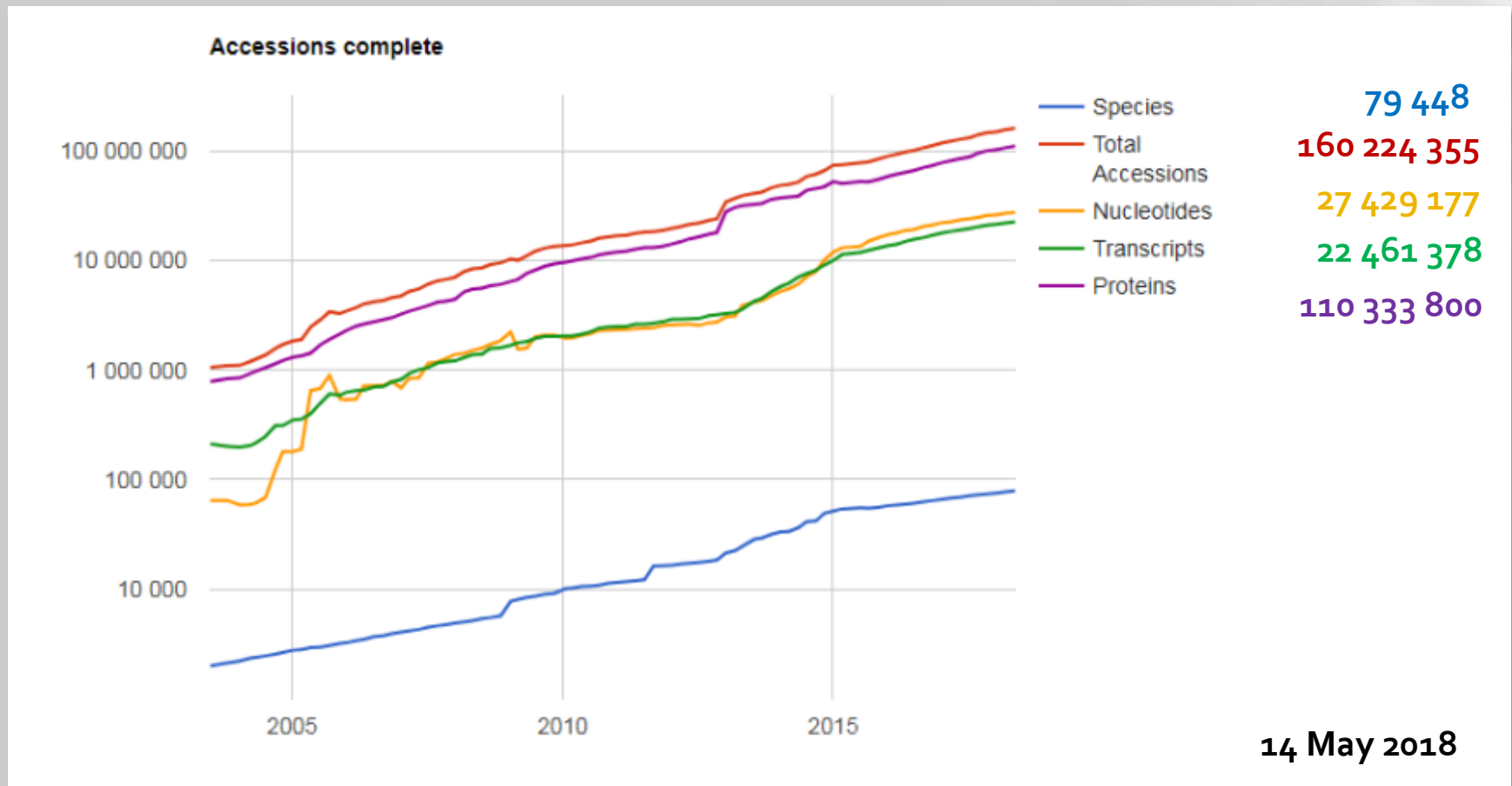
Charged and polar side chains are situated on the solvent-exposed surface where they interact with surrounding water molecules



*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*

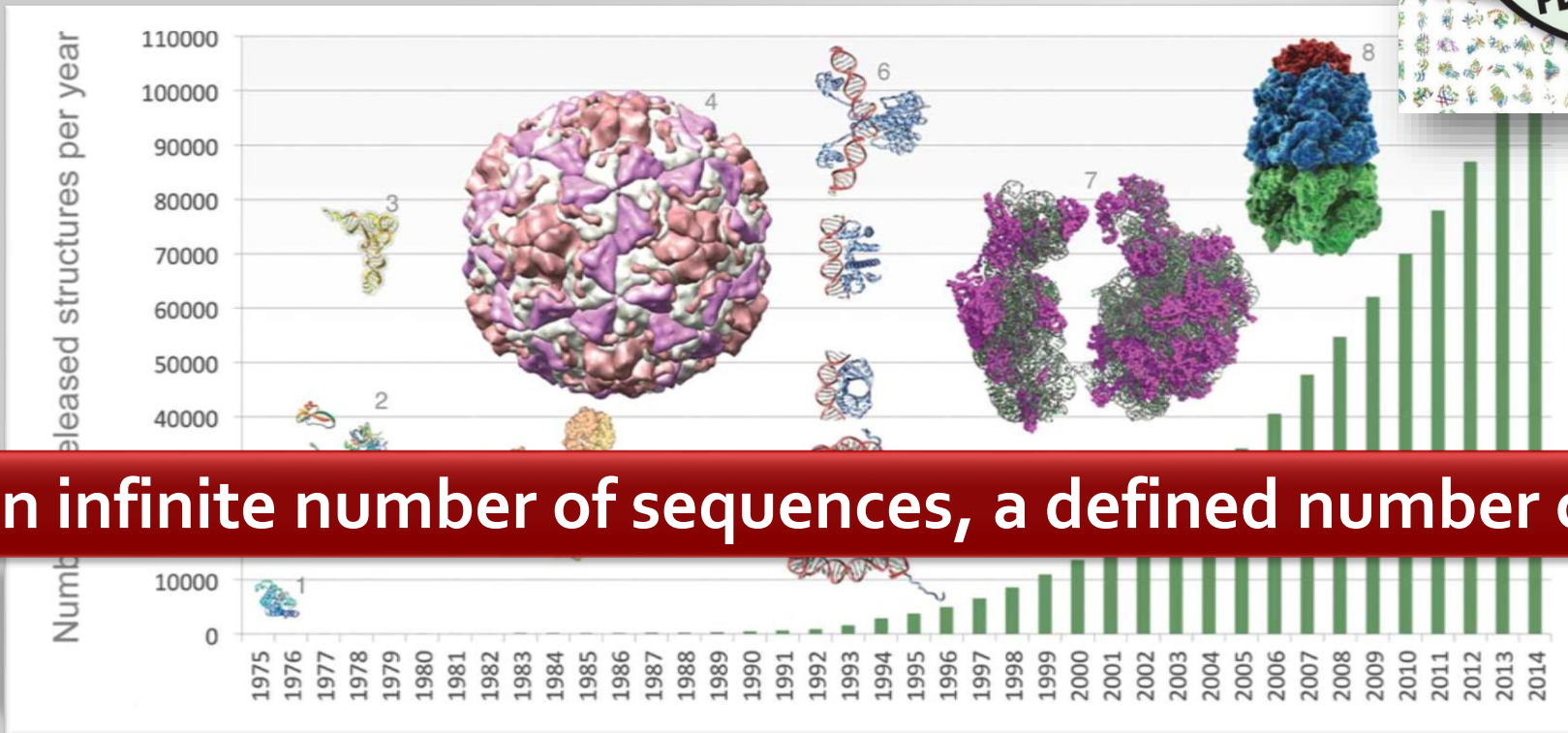
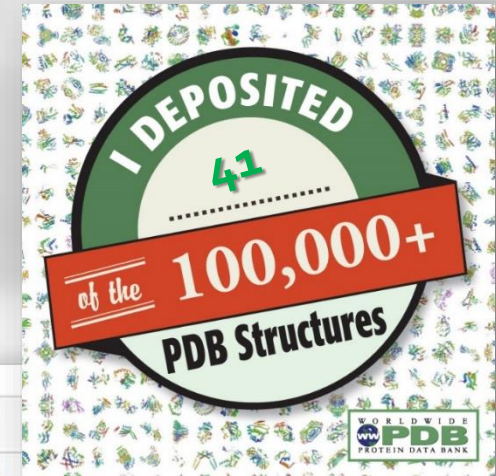
Available sequence data in RefSeq

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

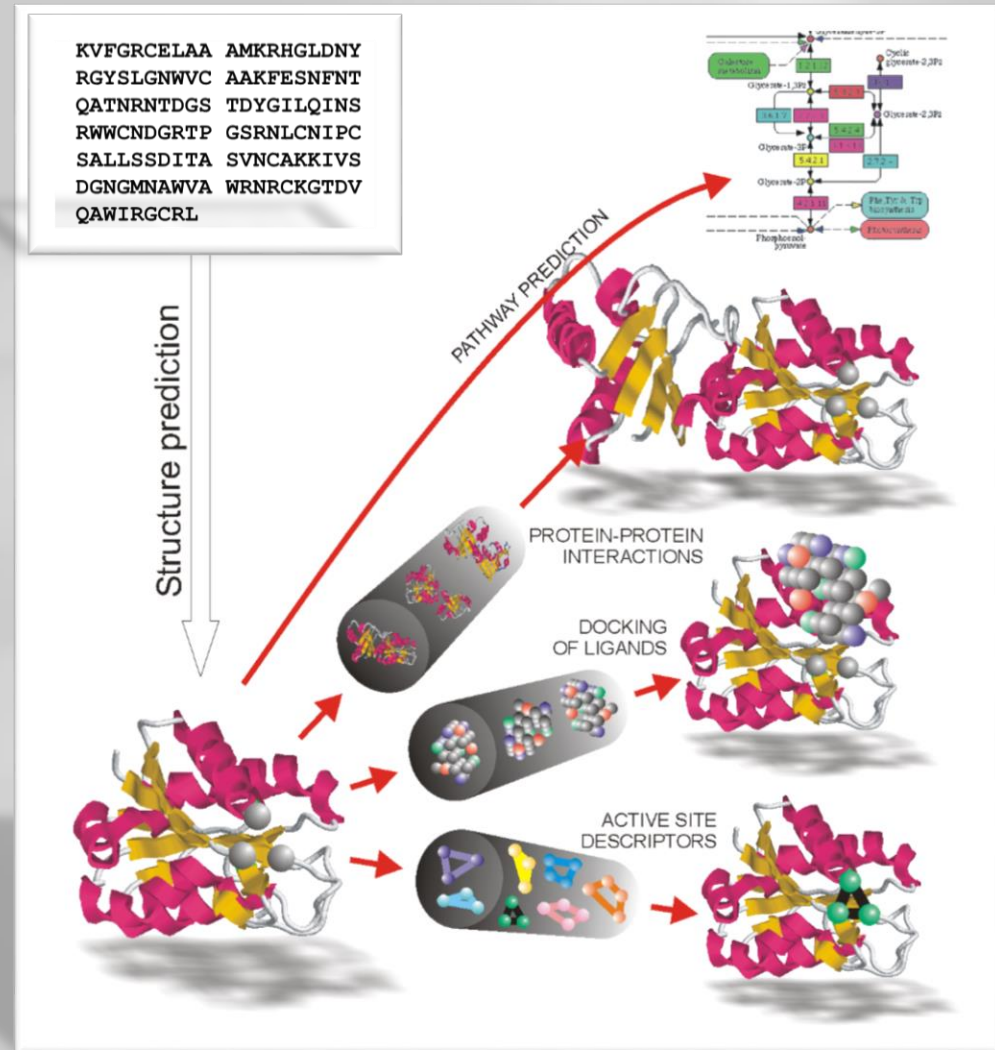


An infinite number of sequences, a defined number of folds

How to predict a structure

Why predicting a structure?

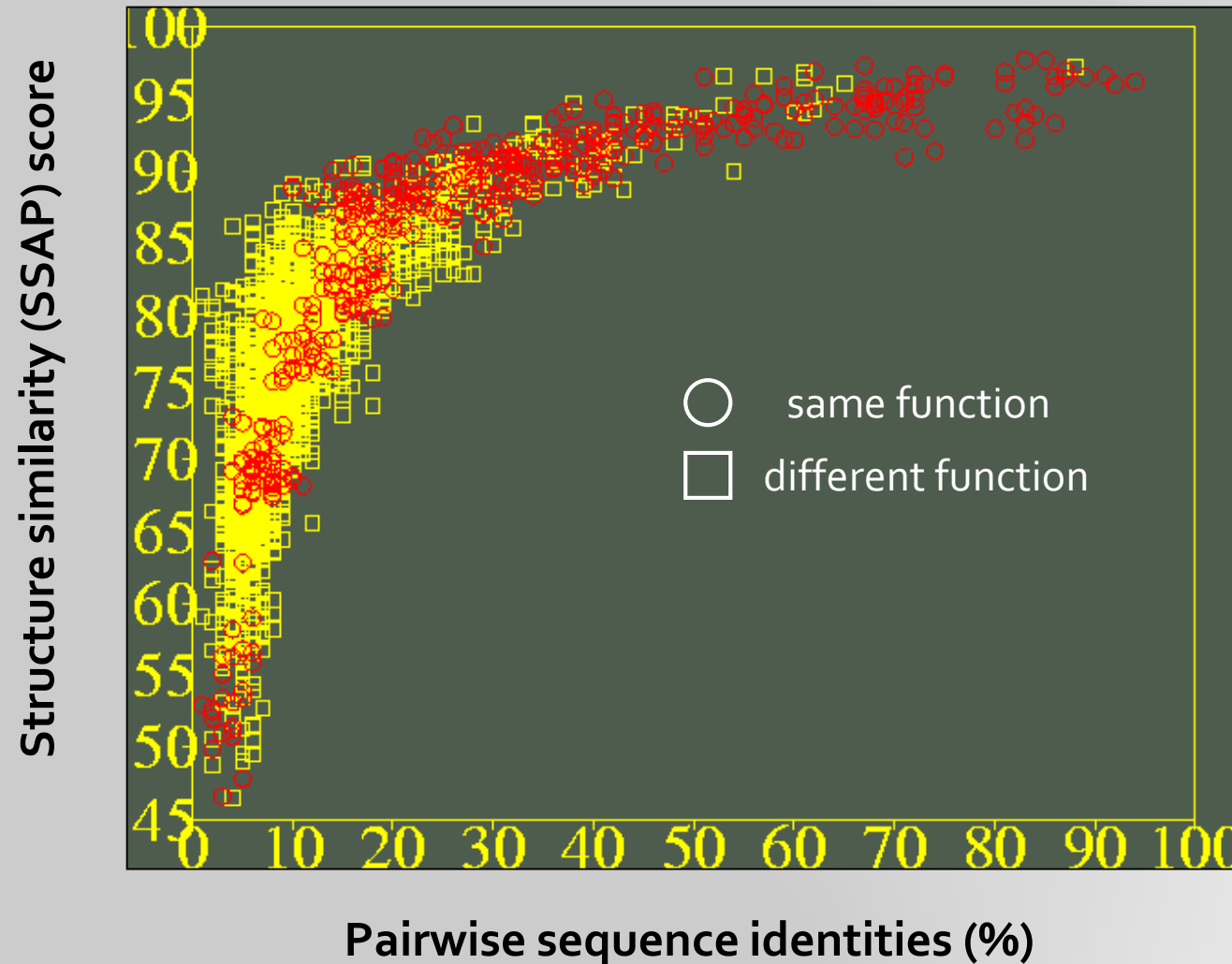
In a first approach, depending on the biological question, numerous sequences can be structurally modelled



How to predict a structure

Before predicting the structure

Proteins that share **sequence identity** are structurally similar



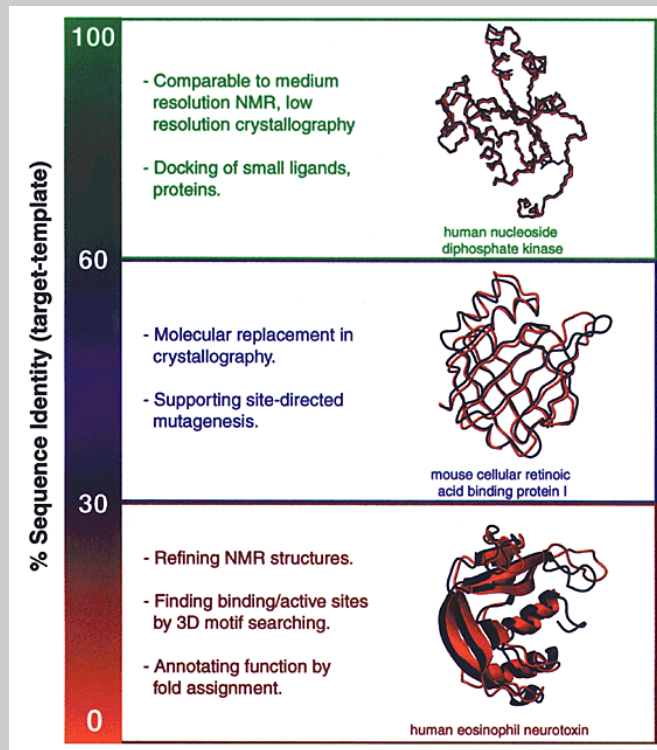
How to predict a structure

Twilight zone

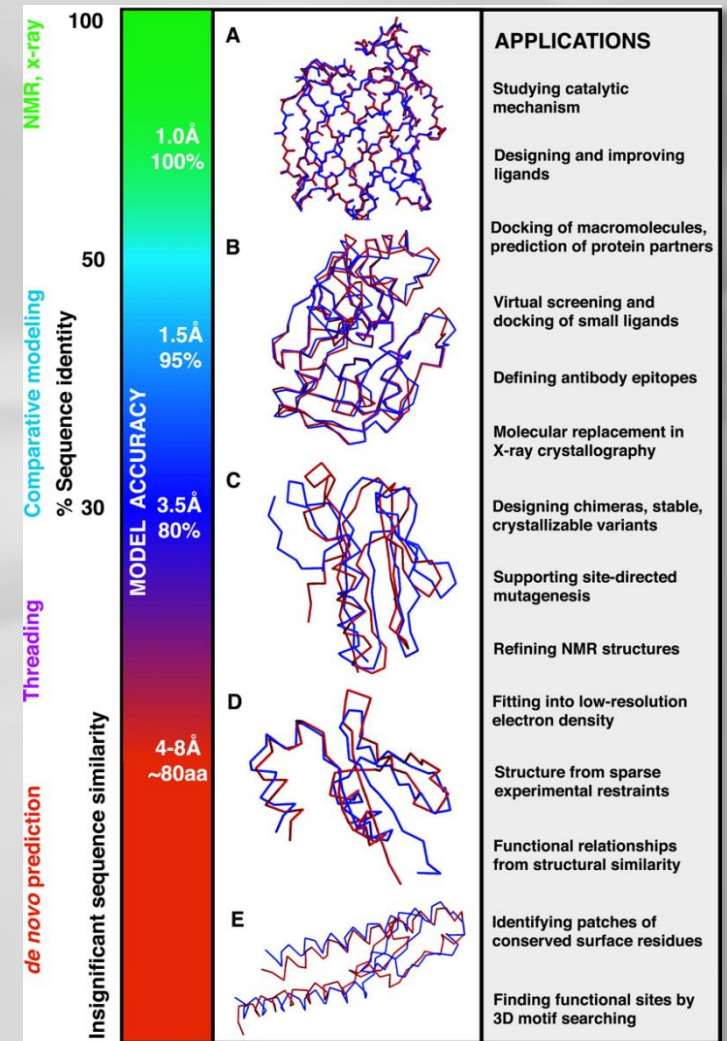
18%

25%

identité



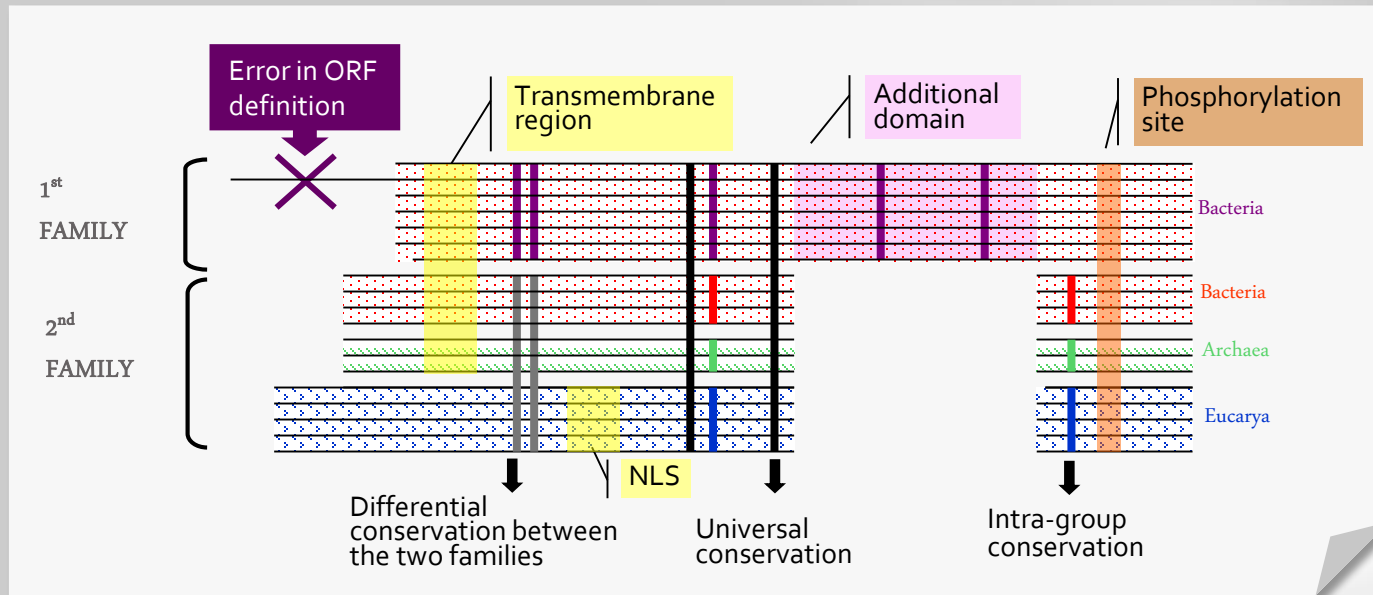
Sali, A. & Kuriyan, J. *Trends Biochem. Sci.* 22, M20–M24 (1999)



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

Multiple alignment of complete sequences

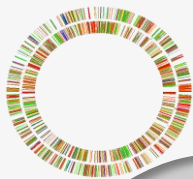
MACS



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

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

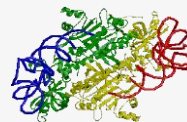
Functional
genomics



Evolutionary
studies



Structure
modeling



Mutagenesis
experiments

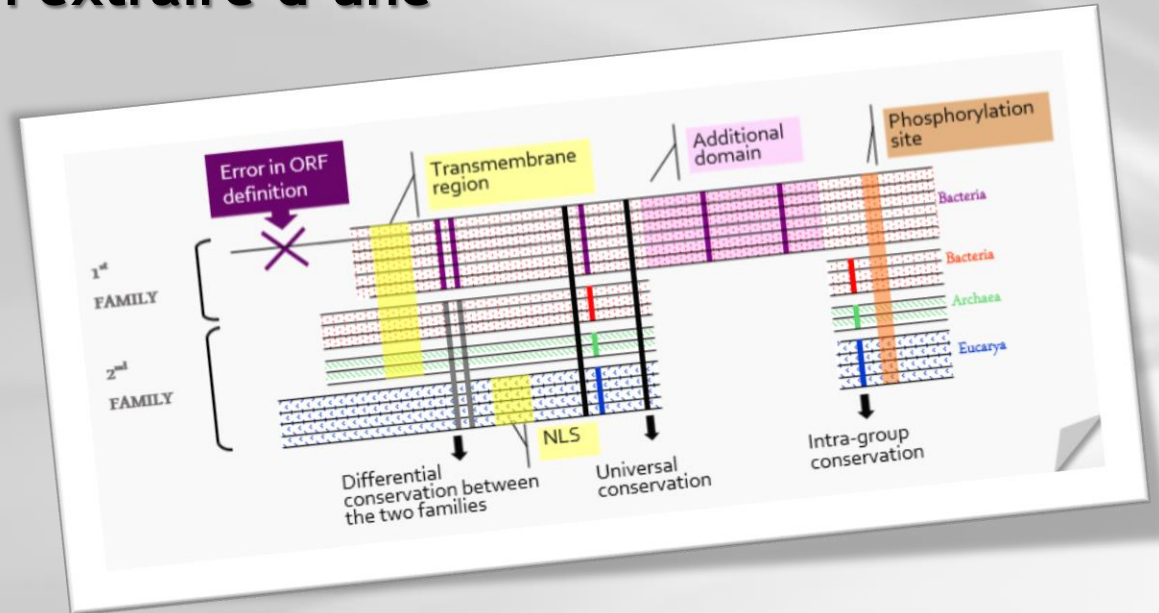


Drug design



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/>

Categories

- proteomics
 - protein sequences and identification
 - proteomics experiment
 - function analysis
 - sequence sites, features and motifs
 - protein modifications
 - protein structure
 - protein interactions
 - similarity search/alignment
- 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]
- EMBLnet services • bioinformatics tools, databases and courses • [more]
- ENZYME • enzyme nomenclature • [more]
- GlyTouCan • international glycan structure repository • [more]
- GPSDB • gene and protein synonyms • [more]
- HAMAP • UniProtKB family classification and annotation • [more]
- MatrixDB • protein-glycosaminoglycan interactions • [more]
- MetaNetX • Metabolic Network Repository & Analysis • [more]
- MIAPEGelDB • MIAPE document edition • [more]
- MyHits • protein domains database and tools • [more]
- PaxDb • protein abundance database • [more]
- Prolune • Popular science articles (in French) • [more]
- Protein Model Portal • structural information for a protein • [more]
- Protein Spotlight • Informally written reviews on proteins • [more]
- Rhea • expert curated resource of biochemical reactions • [more]
- SugarBind • pathogen sugar-binding • [more]
- SWISS-2DPAGE • proteins on 2-D and SDS PAGE maps • [more]
- SwissBiolsistere • biolsisteres for small molecules • [more]
- SwissLipids • knowledge resource for lipid biology • [more]
- SwissPalm • database of S-palmitoylation events • [more]
- SwissSidechain • non-natural amino-acid sidechains • [more]
- SwissVar • variants in UniProtKB entries • [more]
- TCS • interaction specificity in two-component systems • [more]
- UniCarb-DB • glycan mass and structural data • [more]
- UniCarbKB • curated glycan database • [more]
- UniParc (UniProt sequence archive) • UniProt sequence archive • [more]
- UniPathway • metabolic pathways for the UniProtKB • [more]
- UniRef (UniProt sequence clusters) • UniProtKB sequence clusters • [more]
- UniRef (UniProt sequence clusters) • UniProtKB sequence clusters • [more]
- UniRef (UniProt sequence clusters) • UniProtKB sequence clusters • [more]
- VenomZone • portal to venom protein UniProtKB entries • [more]
- World-2DPAGE Constellation • set of 2DPAGE resources • [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]
- 3c15 • find user-defined patterns in protein sequences • [more]
- AACompIdent • protein identification by aa composition • [more]
- AACompSim • amino acid composition comparison • [more]
- Agadir • Prediction of the helical content of peptides • [more]
- ALF • simulation of genome evolution • [more]
- Alignment tools • Four tools for multiple alignments • [more]
- APSSP • Advanced Protein Secondary Structure Prediction • [more]
- Ascalaph • Molecular modeling software • [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]
- BLAST - PBIL • BLAST search on protein sequence databases • [more]
- BLAST - PBIL • BLAST search on protein sequence databases • [more]
- Blast2Fasta • Blast to Fasta conversion • [more]
- boxshade • MSA pretty printer • [more]
- CFSSP • Protein secondary structure prediction • [more]
- ChloroP • chloroplast transit peptides & cleavage sites • [more]
- Click2Drug • Directory of computational drug design tools • [more]
- Click2Drug • Directory of computational drug design tools • [more]
- Click2Drug • Directory of computational drug design tools • [more]
- ClustalO (UniProt) • Align two or more protein sequences • [more]
- ClustalW • Multiple sequence alignment • [more]
- ClustalW • Multiple sequence alignment program • [more]
- ClustalW2 • Multiple sequence alignment program • [more]
- Coiled-Coils prediction • Prediction of coiled coils regions • [more]
- COILS • Prediction of Coiled Coil Regions in Proteins • [more]
- ColorSeq • Color Protein Sequence • [more]
- ColorSeq • Color Protein Sequence • [more]
- Compute pI/MW • theoretical pI and Mw computation • [more]
- CPHmodels • Protein homology modeling • [more]
- CPHmodels • Protein homology modeling • [more]
- CSS-Palm • Prediction of palmitoylation sites in proteins • [more]
- CSS-Palm • Prediction of palmitoylation sites in proteins • [more]
- DAS-TMfilter • Prediction of transmembrane regions • [more]
- DAS-TMfilter • Prediction of transmembrane regions • [more]
- Decrease redundancy • Sequence redundancy reduction • [more]
- DIALIGN • Local multiple sequence alignment • [more]
- DIALIGN • Local multiple sequence alignment • [more]
- DictyOGlyc • GlcNAc O-glycosylation sites in D.discoideum • [more]
- DisEMBL • Prediction of disordered protein regions • [more]
- DLP-SVM • Domain linker predictor • [more]

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)

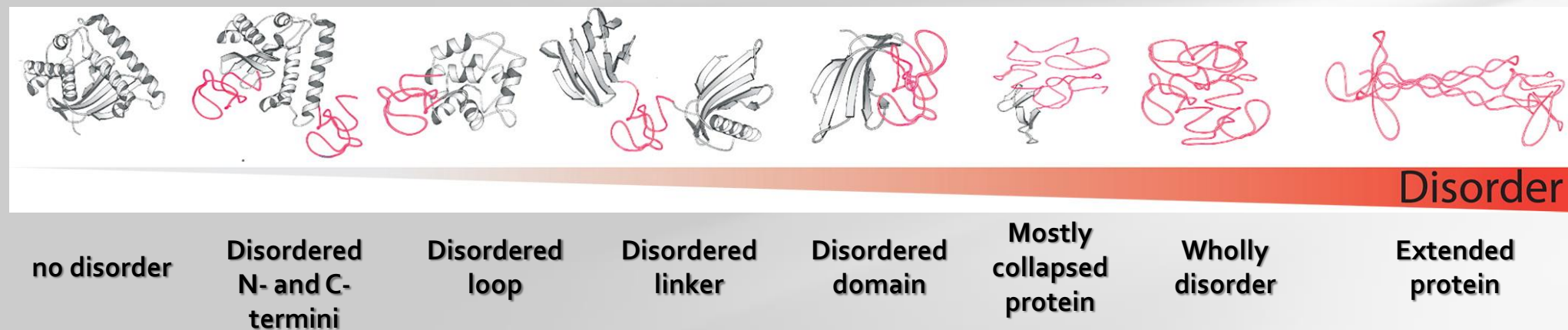
Intrinsically disordered proteins

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



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 /](http://dis.embl.de/)

MEDOR

<http://www.vazymolo.org/MeDor/>

GLOBPLOT₂

<http://globplot.embl.de/>

FoldIndex

<http://bip.weizmann.ac.il/fldbin/findex/>

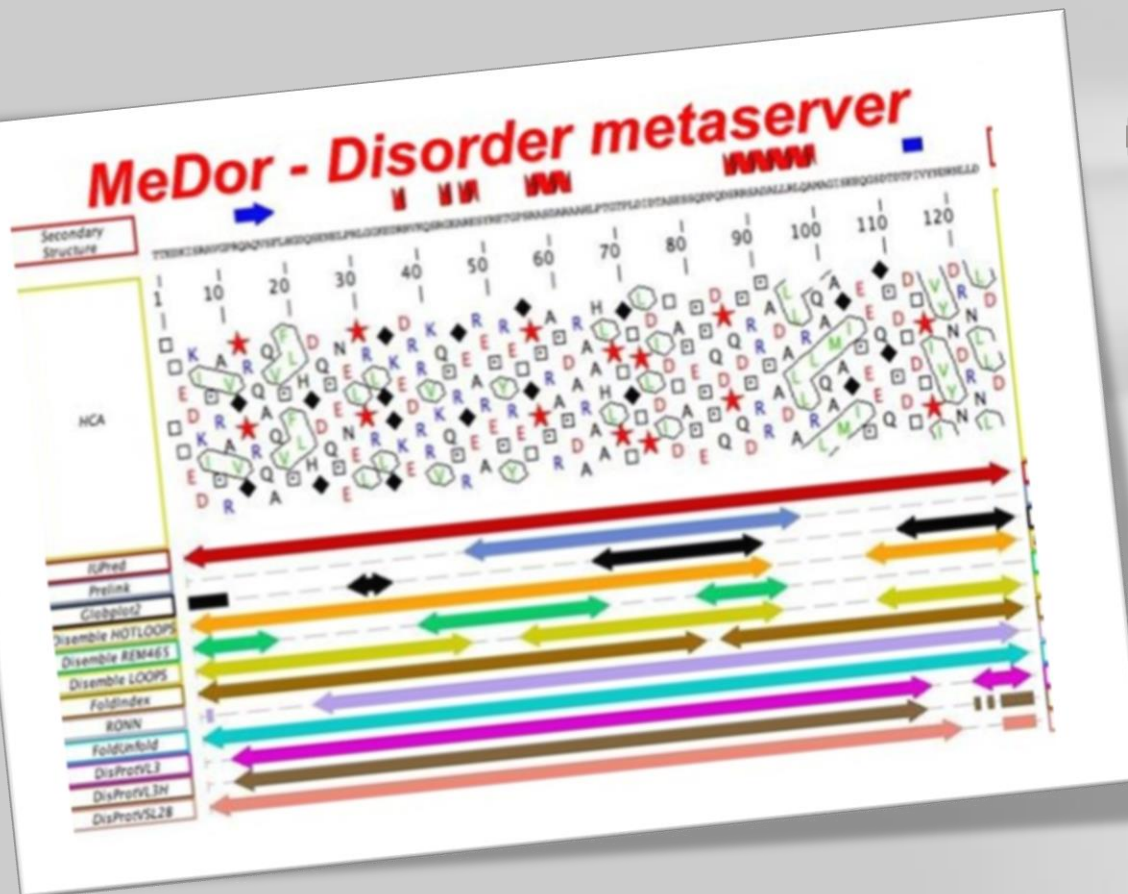
Prediction of order/disorder

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



VaZyMolO Interfaces

[VaZyMolO Home Page](#)

VaZyMolO Interfaces provides a BLAST engine and a browser to our module sequence library.

VaZyMolO is a database dealing with viral sequences at the proteic level. Its aim is to define modules suitable for high expression, solubility and crystallization. Thus it integrates tools starting from amino acids composition, hydrophobic clusters analysis, secondary prediction, modelling, homology with solved structures, data mining concerning biochemistry (function and motifs, active sites, cleavage sites etc). Domains are defined on the structural definition of a domain (which can fold by themselves and show activity); but a module can be constituted by several domains.

[How VaZyMolO is organised?](#)

Three layers in VaZyMolO

Virions are organised into three layers: surface proteins, matrix proteins, and non-structural proteins. The VaZyMolO database organisation has been directly inspired by this organisation and is therefore organised into three layers reflecting surface (layer S), matrix (layer M), and non-structural proteins (layer F).

Image: Original virion representation with courtesy of PVL Laboratory, Dept. of Biological Sciences, University of Warwick, Coventry, UK.

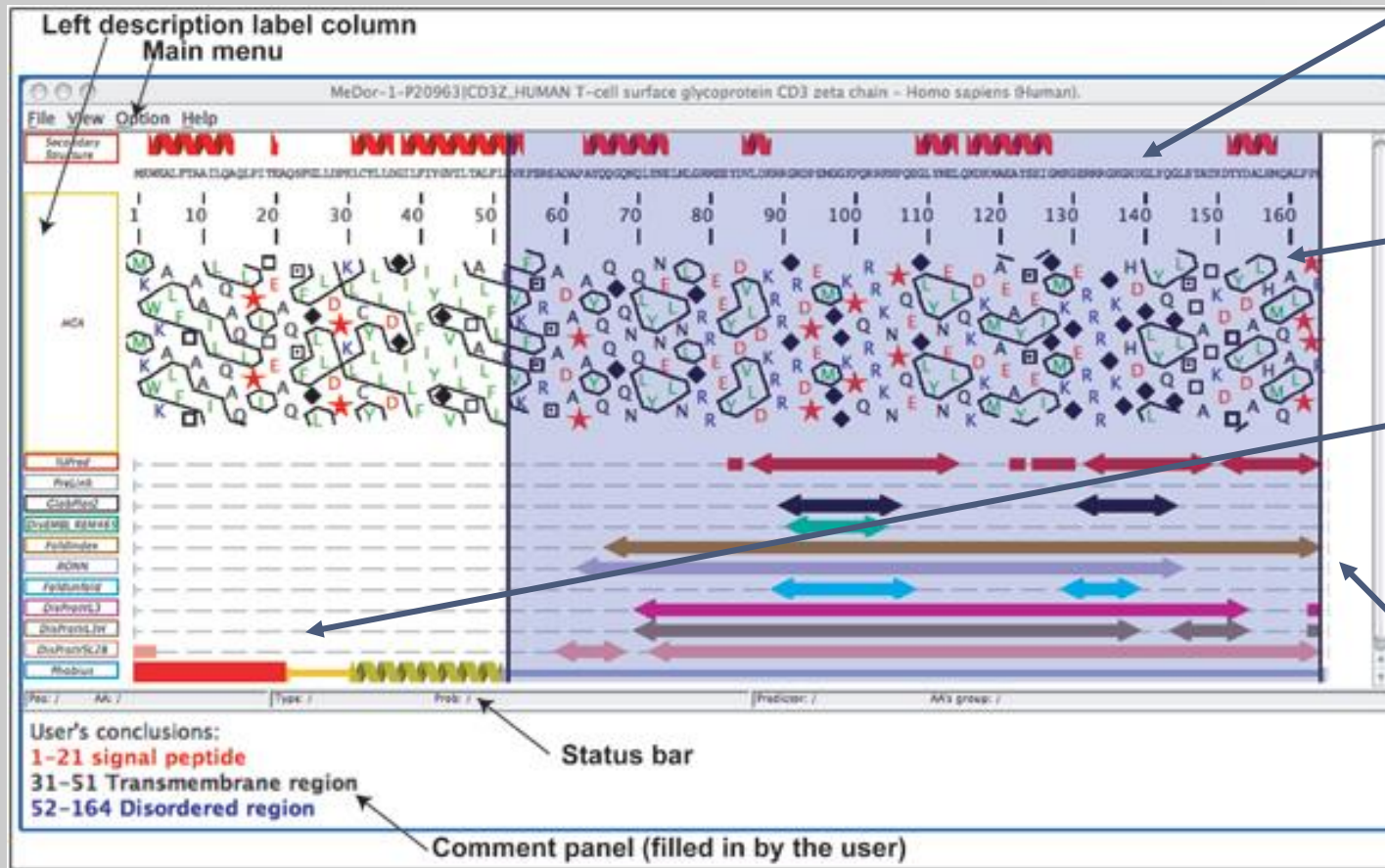
[How to start?](#)

There are 2 ways to use the VaZyMolO interfaces:

- You can seek for information by using our database browser available from the tab entitled "VaZyMolO Browser". Click on a protein name or id to access modular information. Then click on a module to get further details about it.
- If you already have a sequence of interest, you can use our "VaZyMolO Blast and tools" that will enable you with the use of several tools for sequence analysis and a BLAST engine against our database that will retrieve similarities with our data.

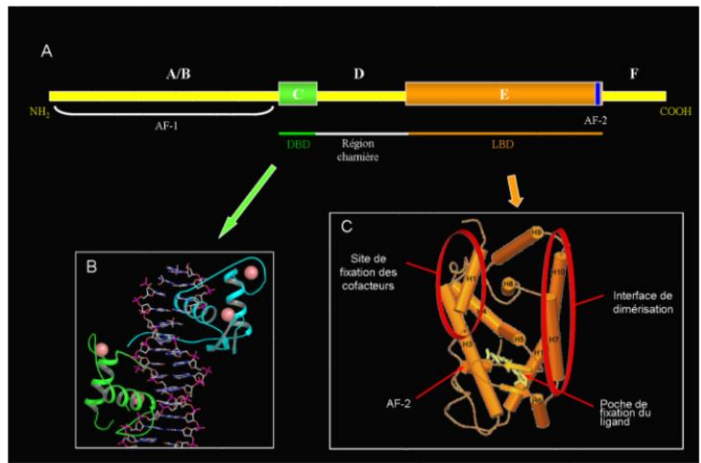
Prediction of order/disorder

Example of a MeDor output



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

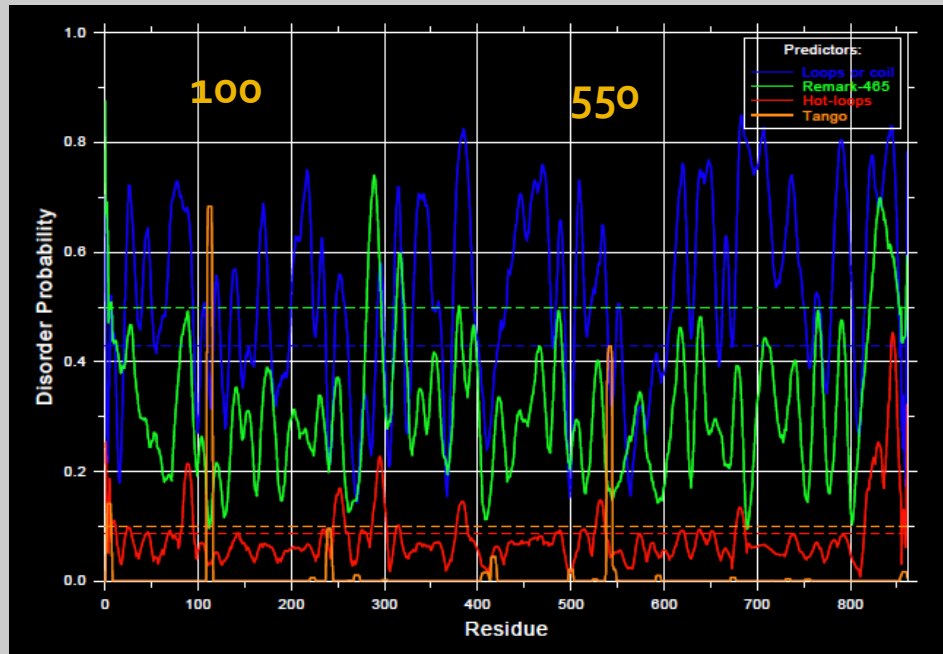
TANGO

<http://dis.embl.de/>

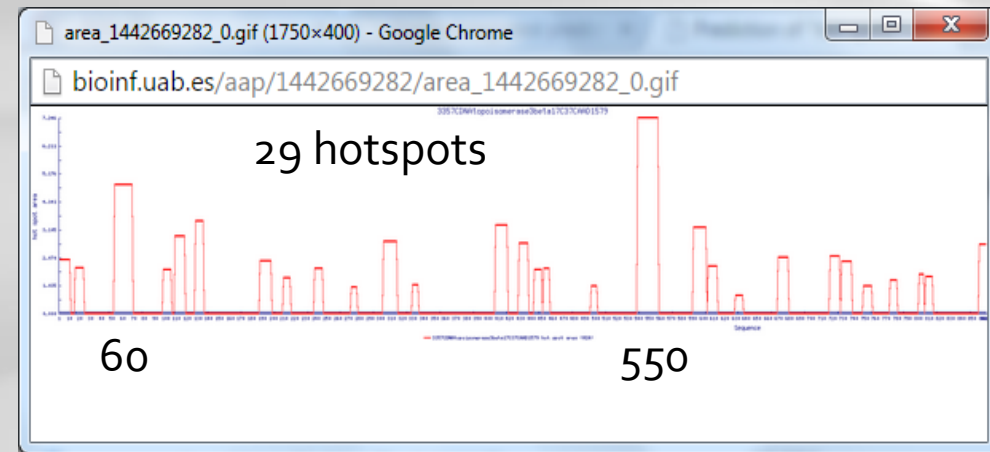
AGGRESKAN

<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 **SOL**ubility 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 **ExpreS** sion and **SOL**ubility) - 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 (**SOL**uble **PR**otein 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 Entropy Reduction 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-c13o6-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.



[ccSQL 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 PProtein 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

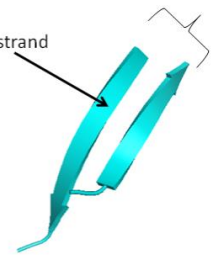
<http://bioinf.cs.ucl.ac.uk/psipred/>

Alpha helix

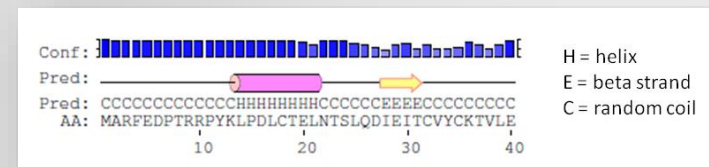


Beta sheet

Beta strand

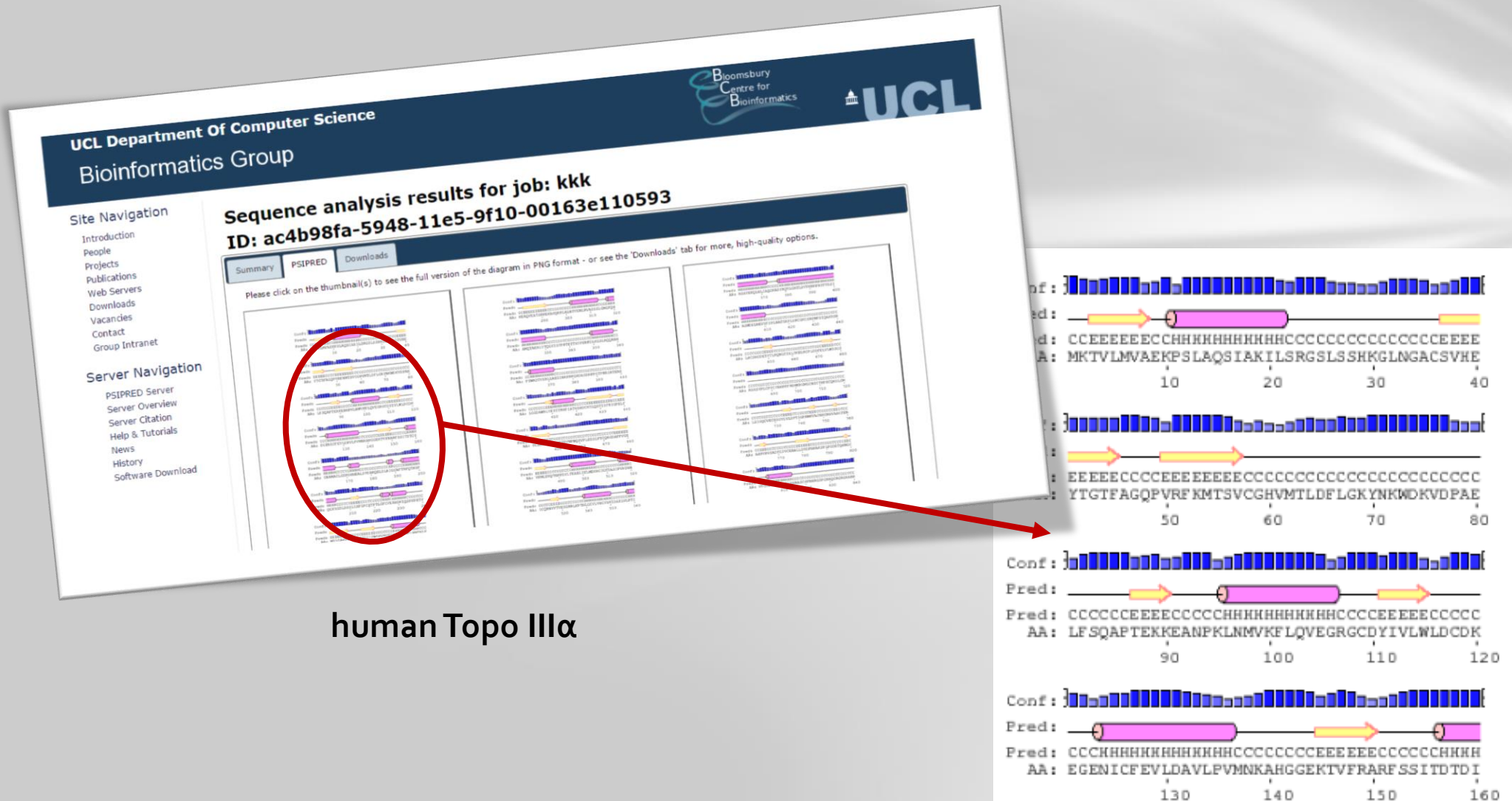


Prediction all along the sequence
with a confidence index



Secondary structure prediction

Example of a PSIPRED output

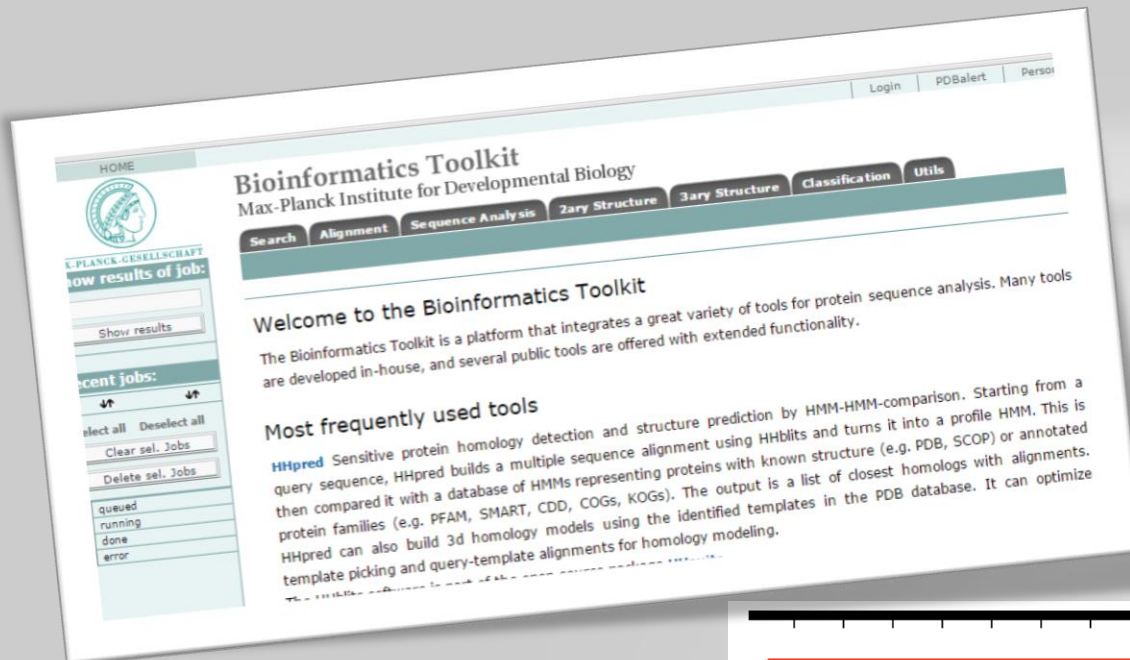


From sequence to structure prediction

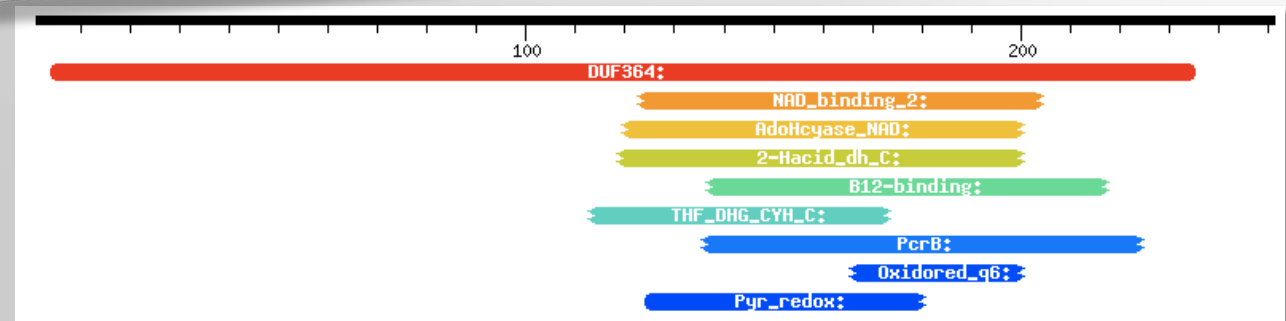
Homology detection and structure prediction by HMM-HMM comparison

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

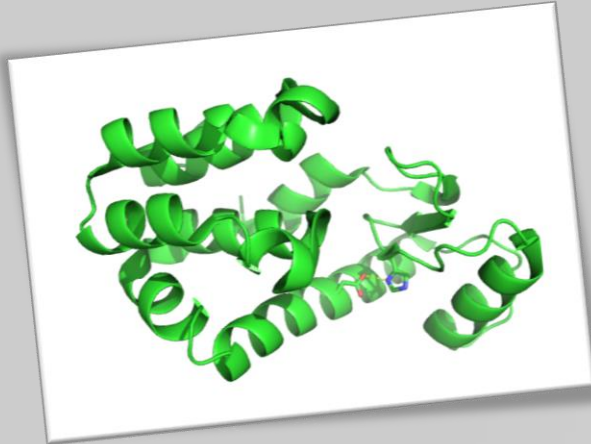
Organisation in domains

Protein organisation in structural and functional domains

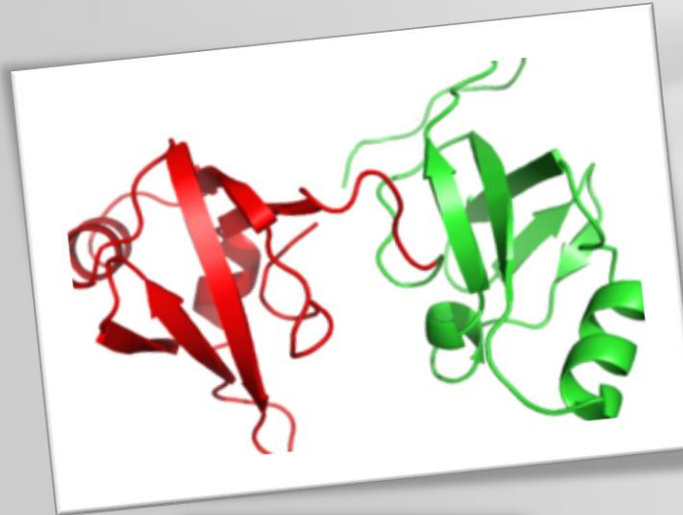
- Conserved part of a protein sequence forming an independent structure that can evolve, operate and exist independently of the rest of the protein chain
- Each domain forms a compact three-dimensional structure and is often independently stable and folded
- Many proteins consist of several structural 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

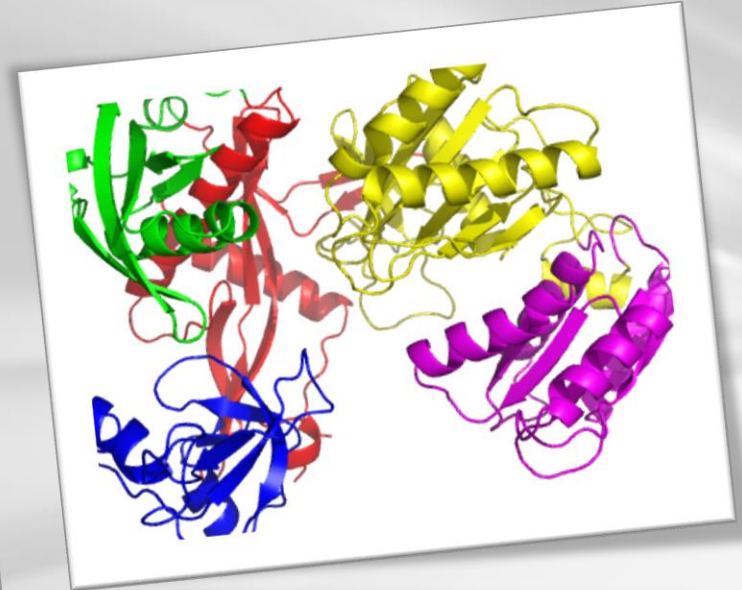
Organisation in domains



One domain



Two domains

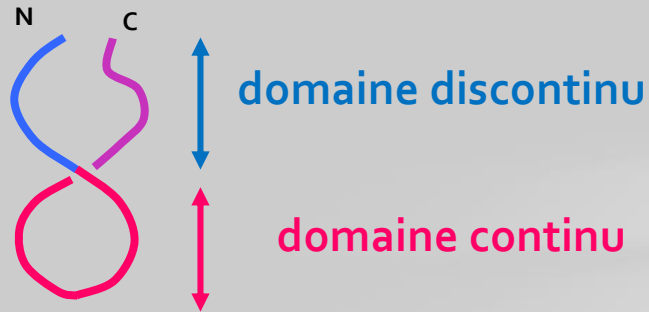


Multidomain protein

Well separated or
tightly packed

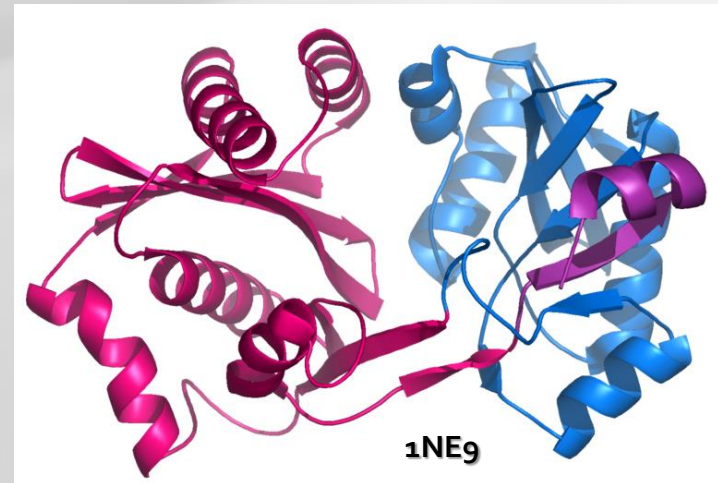
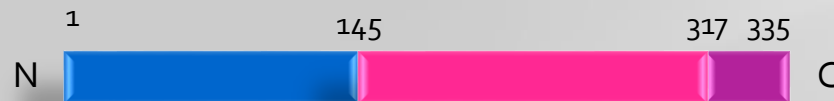
From sequence to structure prediction

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

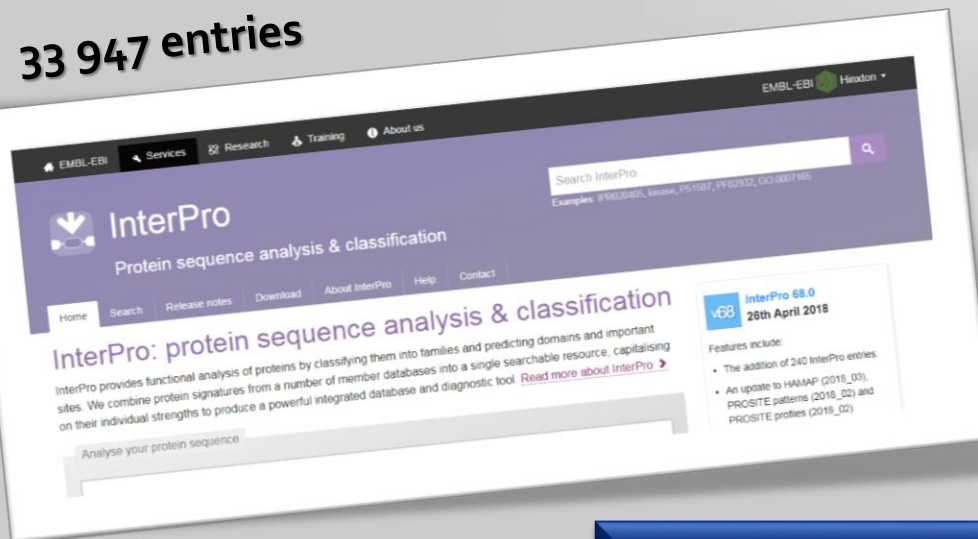
<http://www.uniprot.org/>

Specialized

InterPro

<https://www.ebi.ac.uk/interpro/>

33 947 entries



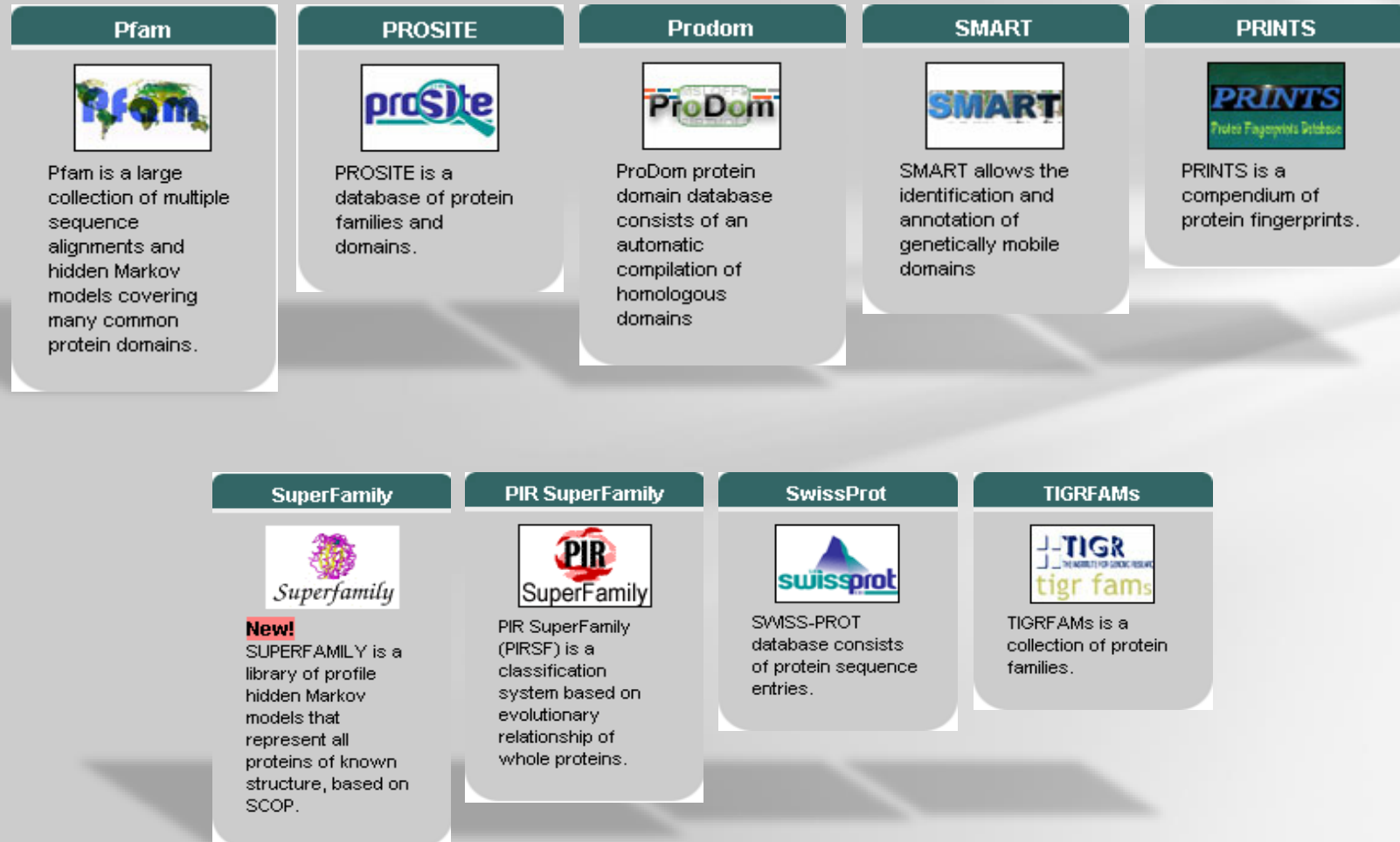
*InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of **member databases** into a single searchable resource, capitalising on their individual strengths to produce a **powerful integrated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

InterProScan

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

From sequence to structure prediction

Integrated databanks



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

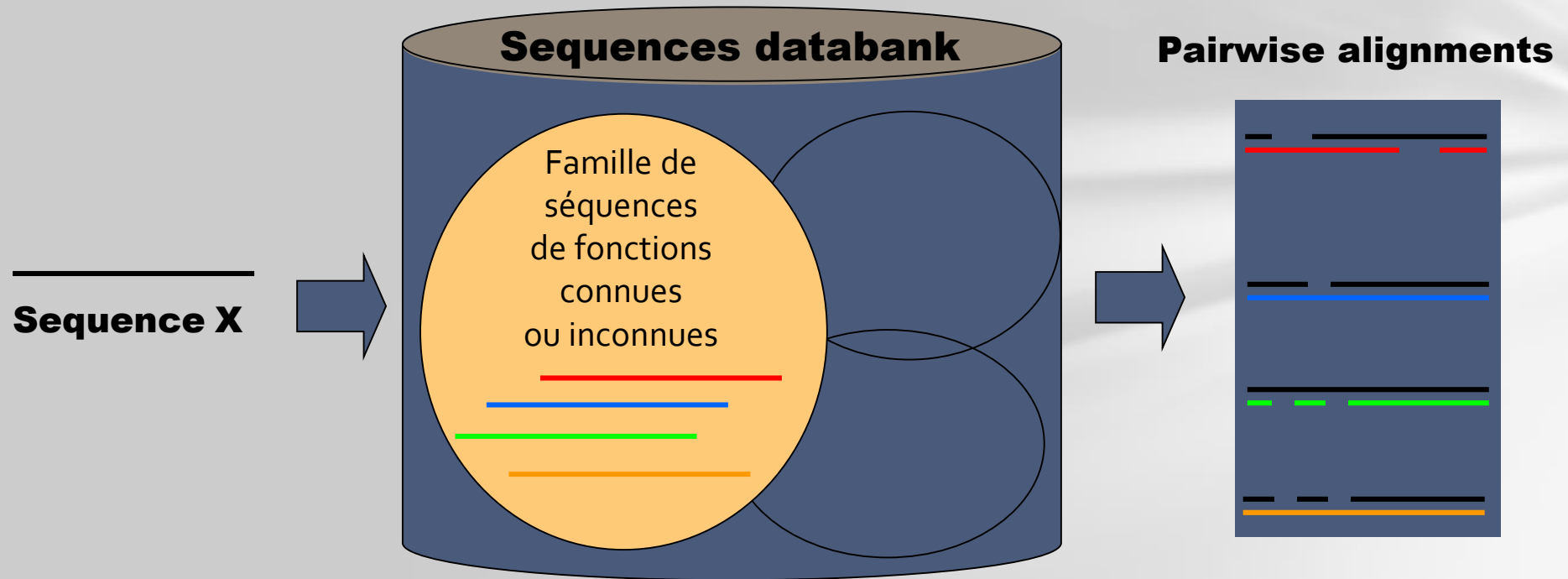
Sequence retrieval

The principles

BLAST

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

Basic Local Alignment Search Tool



Web BLAST



blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide

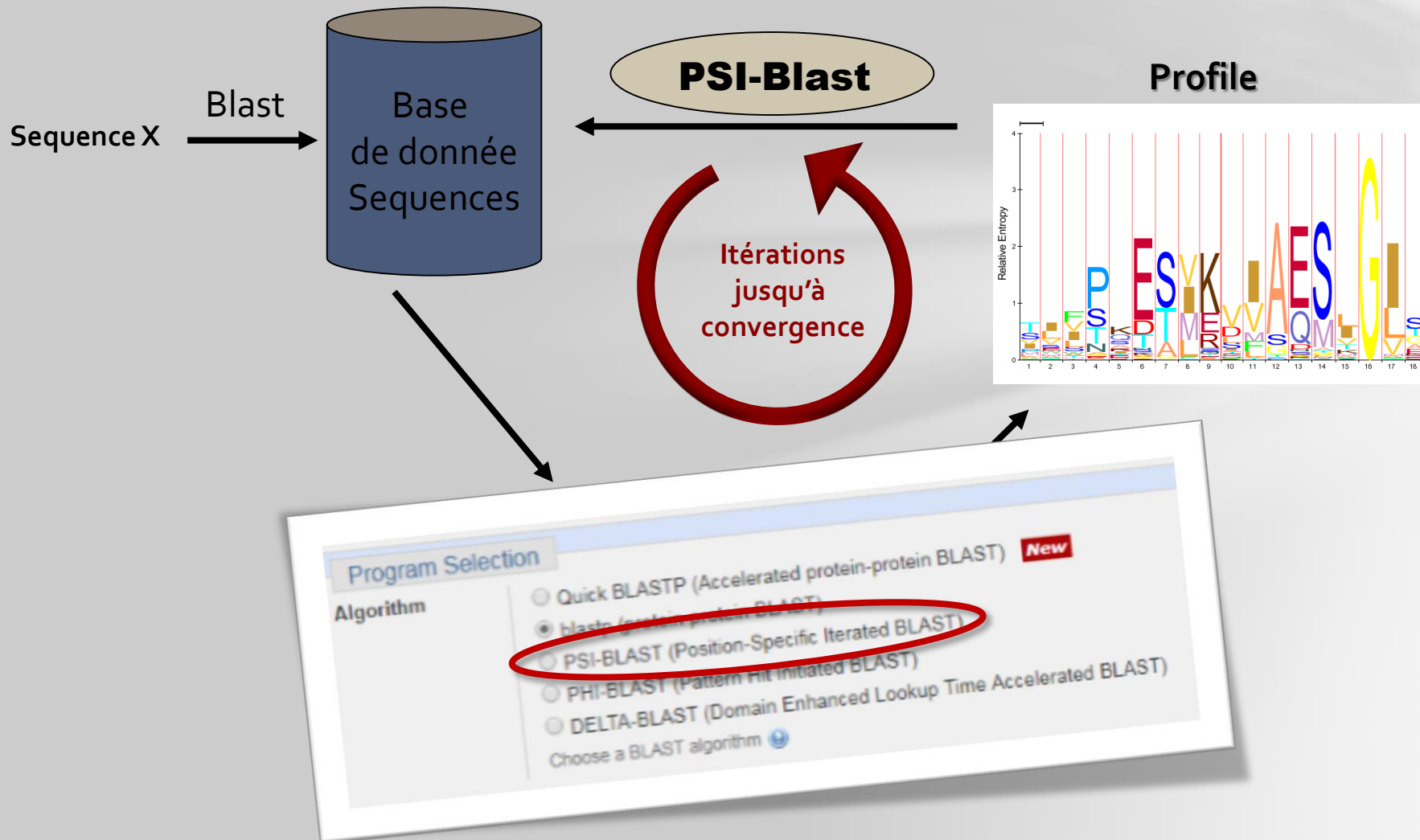


Sequence retrieval

The principles

BLAST

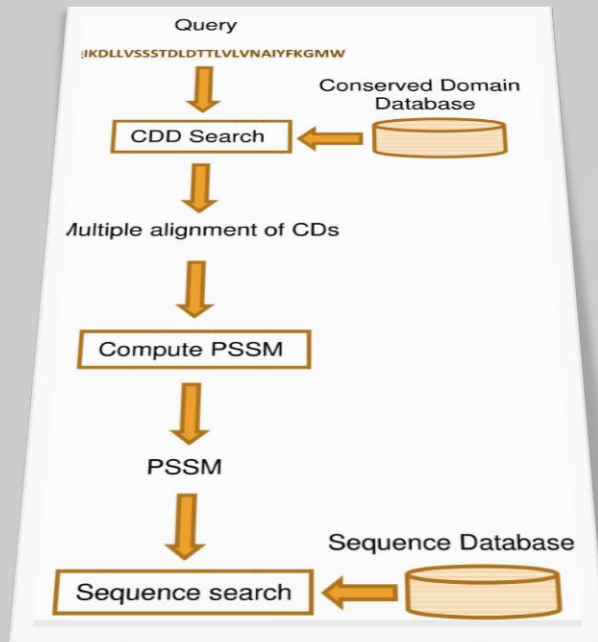
<https://blast.ncbi.nlm.nih.gov/Blast.cgi>



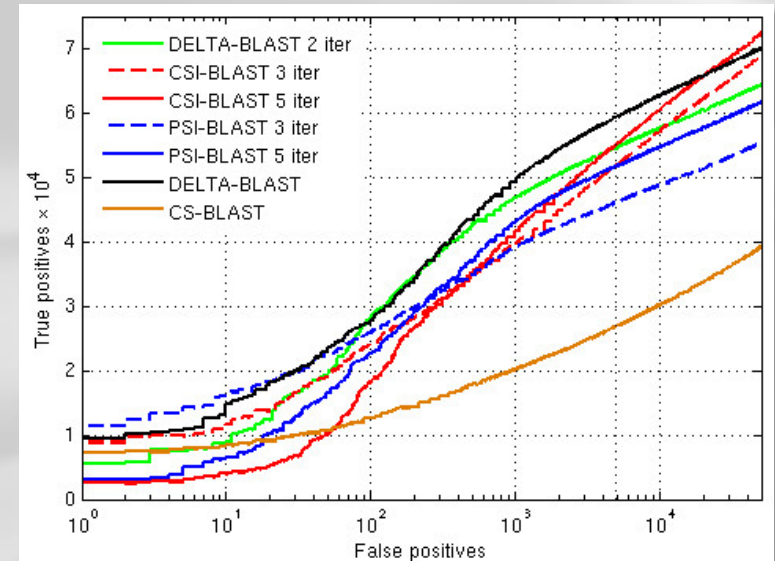
Sequence retrieval

Domain enhanced lookup time accelerated BLAST

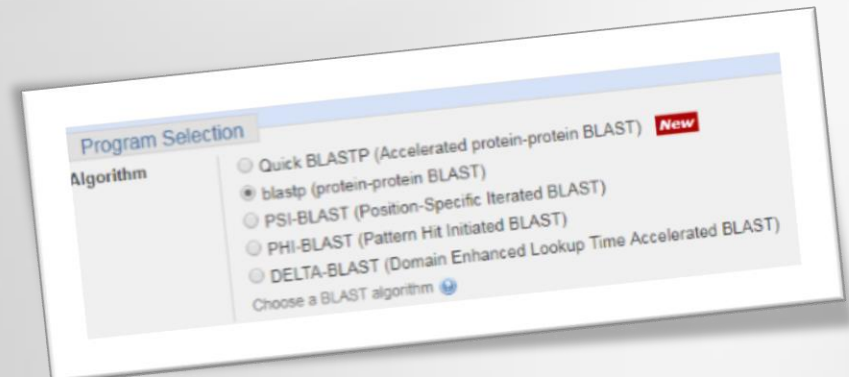
<https://blast.ncbi.nlm.nih.gov/Blast.cgi>



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



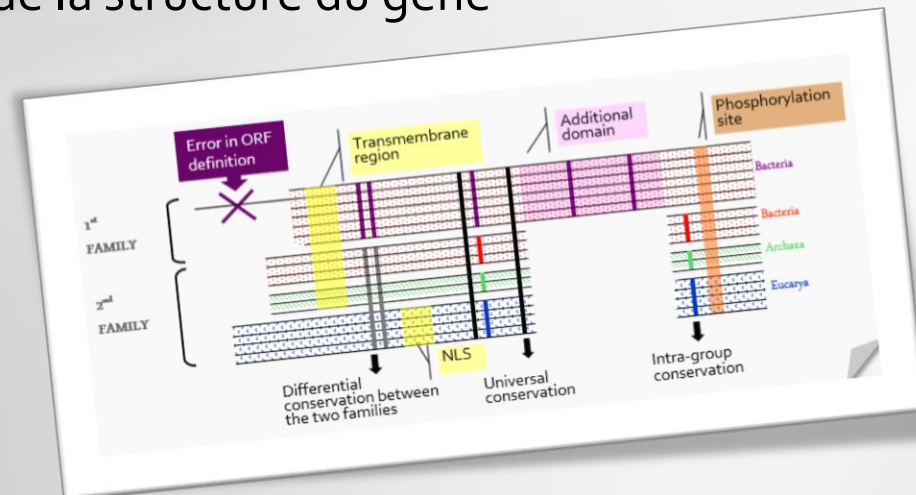
Multiple alignment of complete sequences

➔ **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

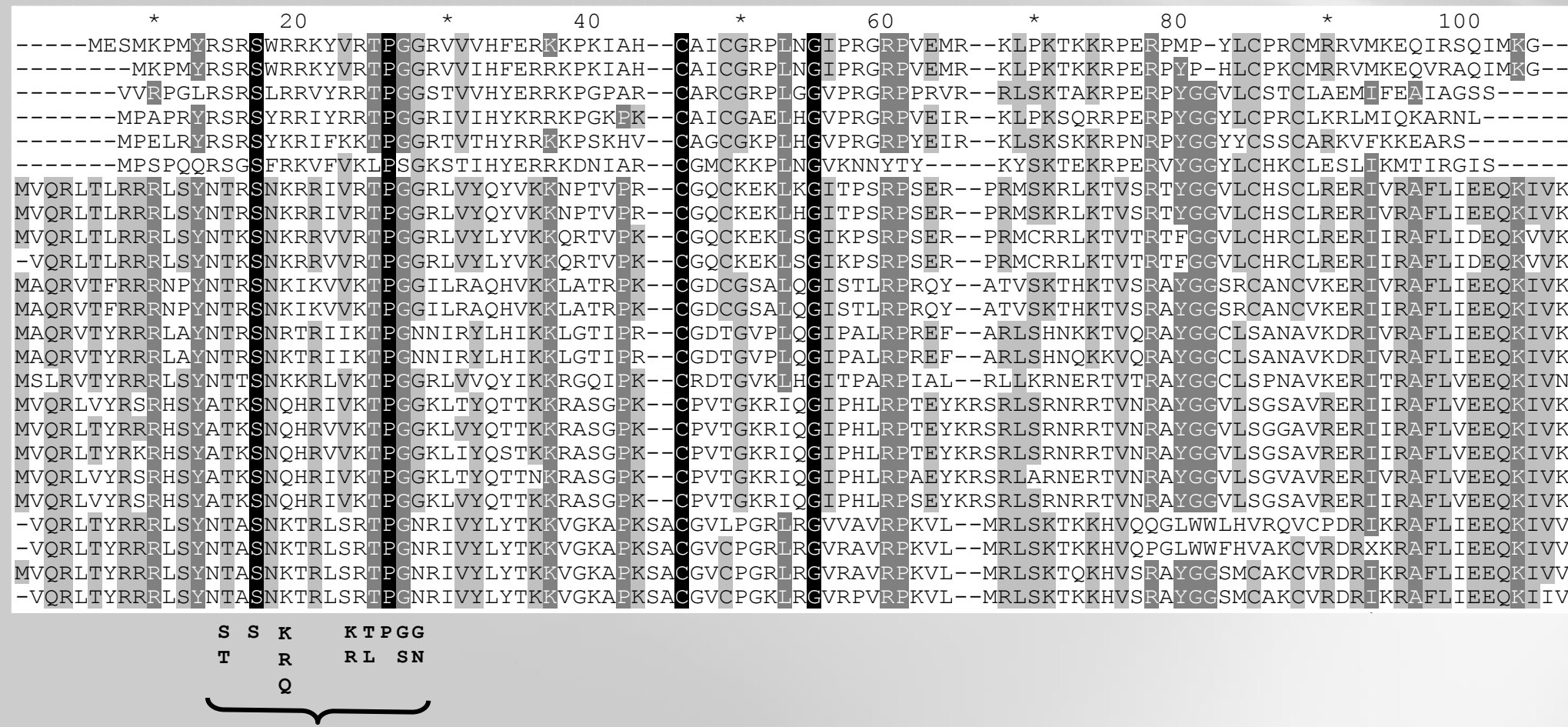
➔ **Validation de la séquence protéique**
détection des erreurs de séquençage, de *frameshift*
détection des erreurs dans la prédiction de la structure du gène
codon initiateur, sites d'épissage exon/intron

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



Multiple alignment of complete sequences

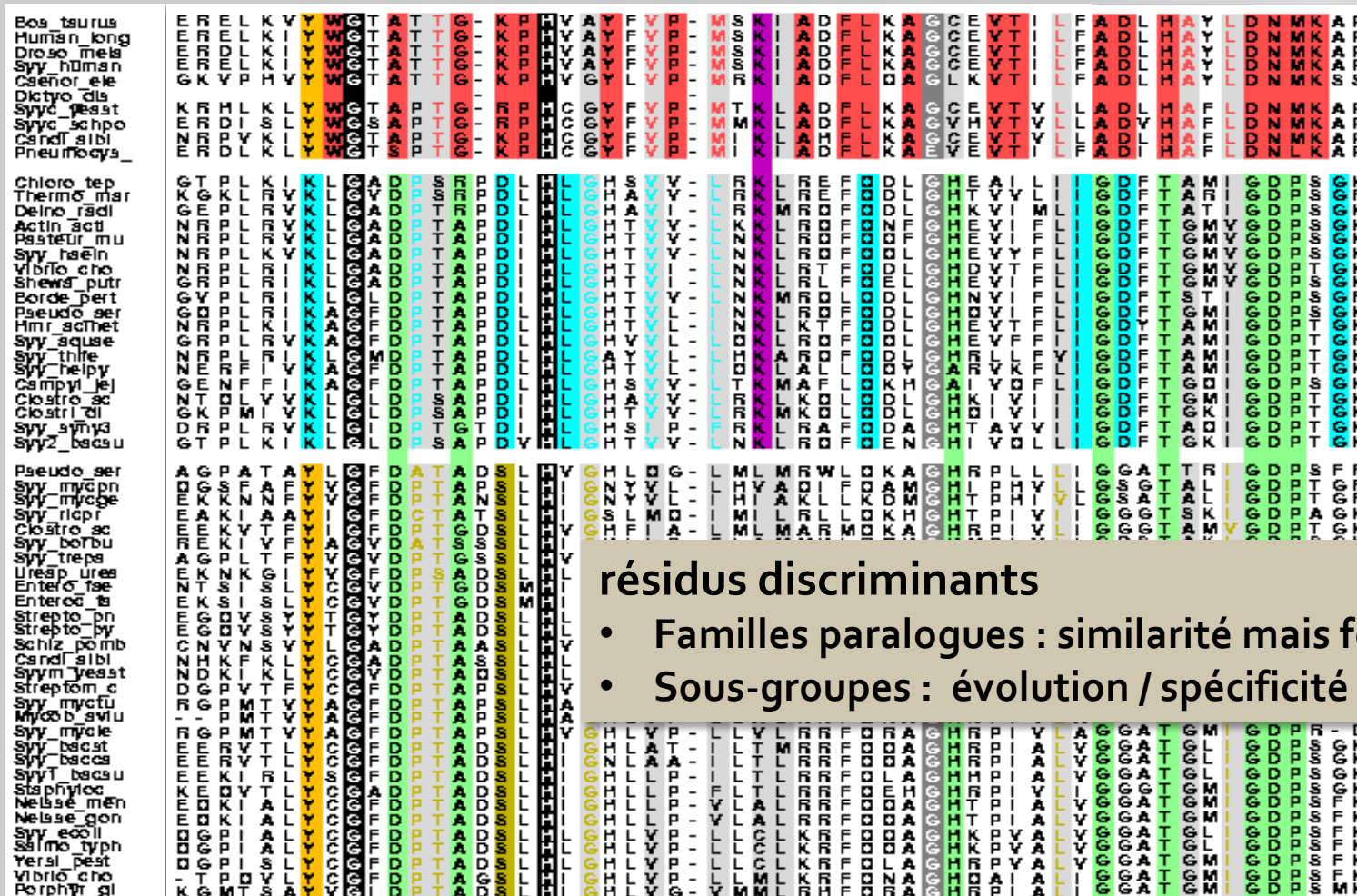
Exemple : Famille des protéines ribosomales L34e



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 pour 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>

ESPrpt

<http://esprpt.ibcp.fr/ESPrpt/ESPrpt/>

WebLogo

<http://weblogo.berkeley.edu/logo.cgi>

Multiple alignment of complete sequences

<http://lbgi.fr/pipealign>

The screenshot shows the PipeAlign2 web interface. At the top right is a 'Contact' link. The main heading is 'PipeAlign2'. Below it is a welcome message and a description of the tool. The 'Create a new session' section contains a text input field for 'Paste your query sequence here', a 'Send' button, an 'Upload your file' section with a 'Choisir un fichier' button and an 'Upload' button, and an 'Or use a file from the server' section with a dropdown menu and a 'Use' button. At the bottom of this section is a 'Restore a previous session' section with a 'Session id' input field and a 'Go' button. The footer includes logos for ICUBE and BISTRO, and a list of authors and the year 2003.

PipeAlign2

Welcome to the PipeAlign online tool !

PipeAlign2 is a protein family analysis tool integrating a multi-step process ranging from the search for sequence homologues in protein and 3D structure databases to the structural / functional annotation of the family. The complete, automatic pipeline takes a single sequence or a set of sequences as input and constructs a high-quality, validated multiple sequence alignment (MSA) in which sequences are clustered into potential functional subgroups. Sequence annotations are automatically extracted from the public database (Uniprot, InterPro, etc.), cross-validated and integrated in the MSA. For the more experienced user, the PipeAlign2 server also provides numerous options to run only a part of the analysis, with the possibility to modify the default parameters of each step.

Create a new session

Paste your query sequence here ?

Your sequence

Send

Upload your file ?

Choisir un fichier Aucun fichier choisi

Upload

Or use a file from the server

/

Use

Restore a previous session

Session id Go

ICUBE BISTRO

PipeAlign: A new toolkit for protein family analysis by Plewniak F, Blanchetti L, Breilvet Y, Carles A, Chalmel F, Lecompte O, Mochel T, Moulinier L, Muller A, Muller J, Pigent V, Ripp R, Thierry JC, Thompson JD, Wicker N, Poch O. Nucleic Acids Res. 2003.

Multiple alignment starting from one sequence

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

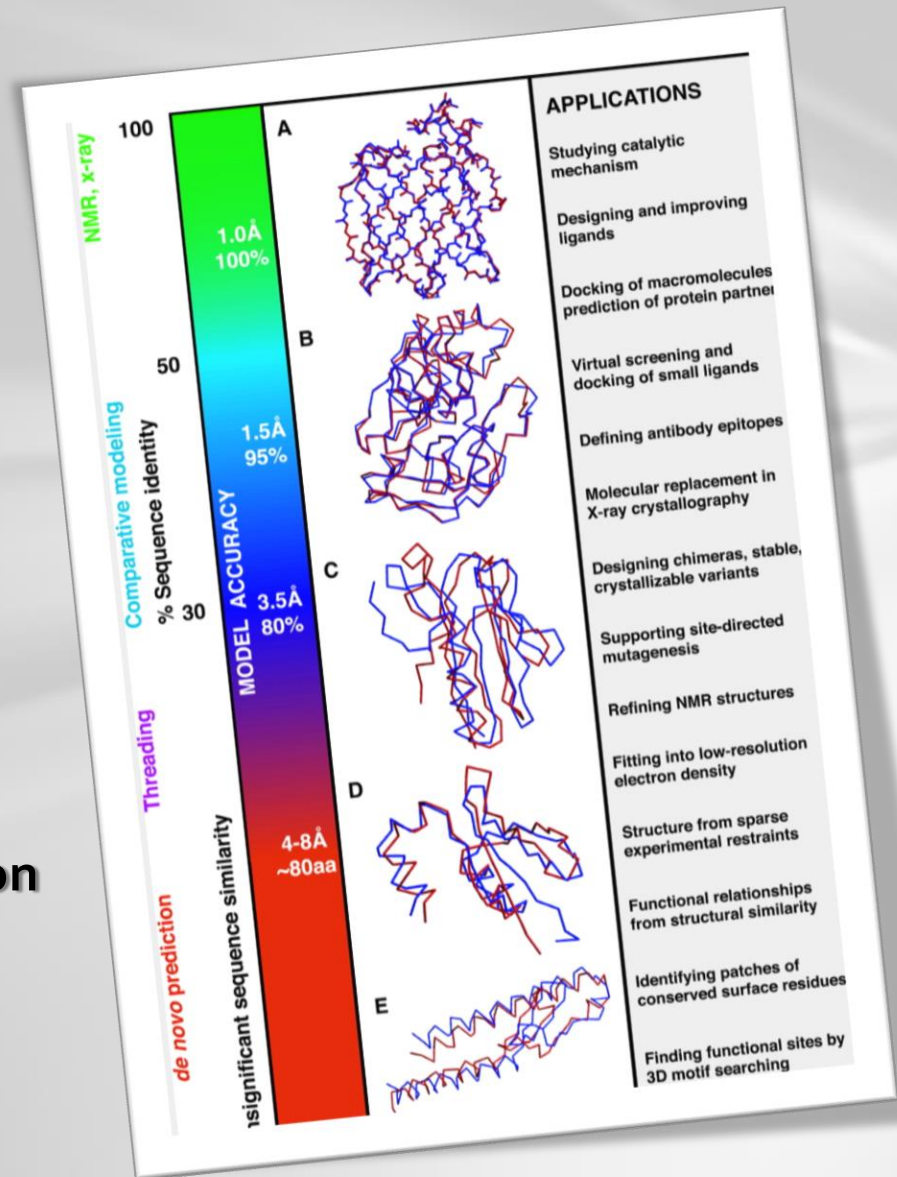
Informations données par la séquence

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

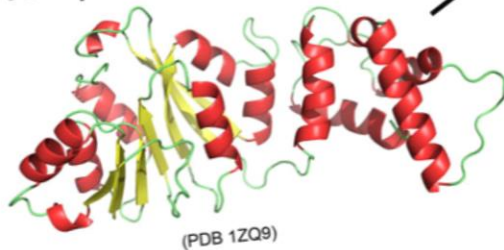


Homology modeling

Target sequence

HLLKNPGILDKIIYAAIKSSDIVLEIGOGTGNLTVKLLPLAKKV
ITIDIDSRMISEVKKRCLYEGYNNLEVYEGDAIKTVFPKFDVCTA
NIPYKISSPLIFKLLISRPLFKCAVLMFQKFAERMLANVGDSNY
SRLTINVKLFCKVTVCNVRSSFPNPPKVDVSVIKLIPKSSFL
TNFDENDNLLRICFSRKRKTLHAIFKRNVLNMLEHNYKNWCTLN
KQVPVNFPPKKYCLDVLEHLMCEKRSINLDENDFLKLLLEFNKK
GIHFF
(T0295 from CASP7)

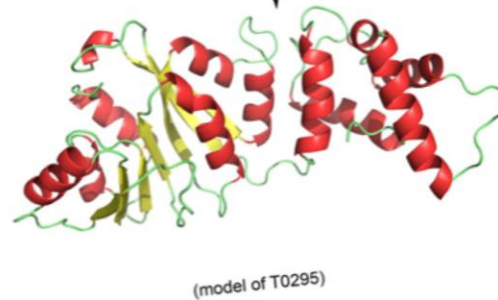
Template structure



Sequence alignment

```
1ZQ9 QHILKNPLIINSIIDKAALEPTDVLVGVPGTGNMTVKLLKAKKVACELDPRLVAELH 60
T0295 -HLLKNPGILDKIIYAAIKSSDIVLEIGOGTGNLTVKLLPLAKKVITIDIDSRMISEVK 59
      *:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*
1ZQ9 KRVOGTFVASKLQVLVGDVLKTDLPFFDTVCVANLPYQISSPFVKLLHRPFPCAILMF 120
T0295 KRCLYEGYN-NLEVYEGDAIKTVFPKFDVCTANIPYKISSPLIFKLLISRPLFKCAVLMF 118
      **      *:*  **:*  *  **:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*
1ZQ9 QREFALRLVAKFGDKLYCRLSINTQLARVDHLMKVGKNFRPPPKVESSVVRKIEPNPP 180
T0295 QKEFAERMLANVGDSNYSRLTINVKLFCKVTVCNVRSSFPNPPKVDVSVIKLIPKSS 178
      *:*:*  *:*  **  *:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*
1ZQ9 PPINFQEWGDLVRITFVRKNKTLAAPKSSAVQQLLEKNYRIHCSVHNIIIPEDFSIADK 240
T0295 FLTNFDENDNLLRICFSRKRKTLHAIFKRNVLNMLEHNYKNWCTLN-KQVPVNFPPKKY 237
      *:*:*:*:*:*  **  *:*:*:*:*  **  *:*:*:*:*  *:*:*  *:*:*  *
1ZQ9 IQQILTSTGFSOKRMSMDIDDFIRLLHGFNAEGIHFS 278
T0295 CLDVLEHLMCEKRSINLDENDFLKLLLEFNKKGIHFF 275
      :*  .:.*:*:*  .:*  *:*:*:*  **  *:*:*:
```

exploit the 3D similarity between
a known template structure
and the target sequence
to build models



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

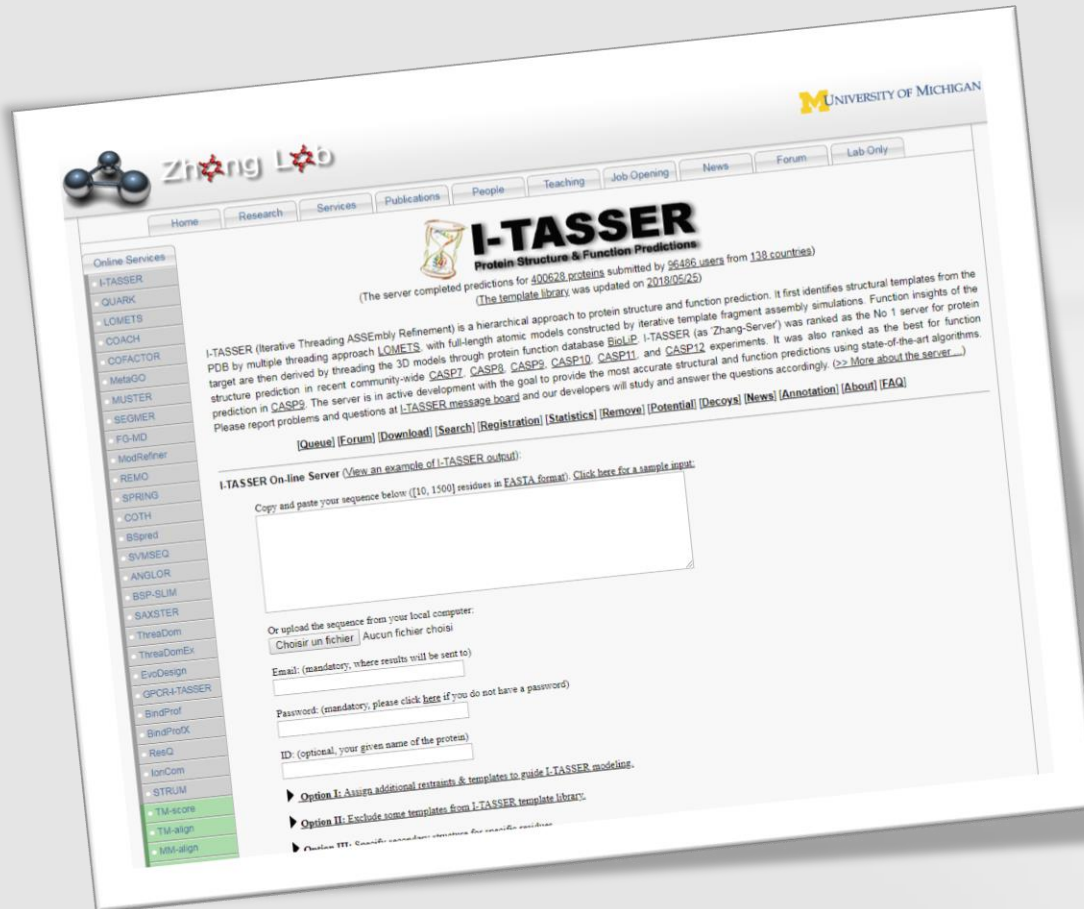
is a *de novo* approach aimed at predicting peptide structures from amino acid sequences, based on a HMM structural alphabet

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

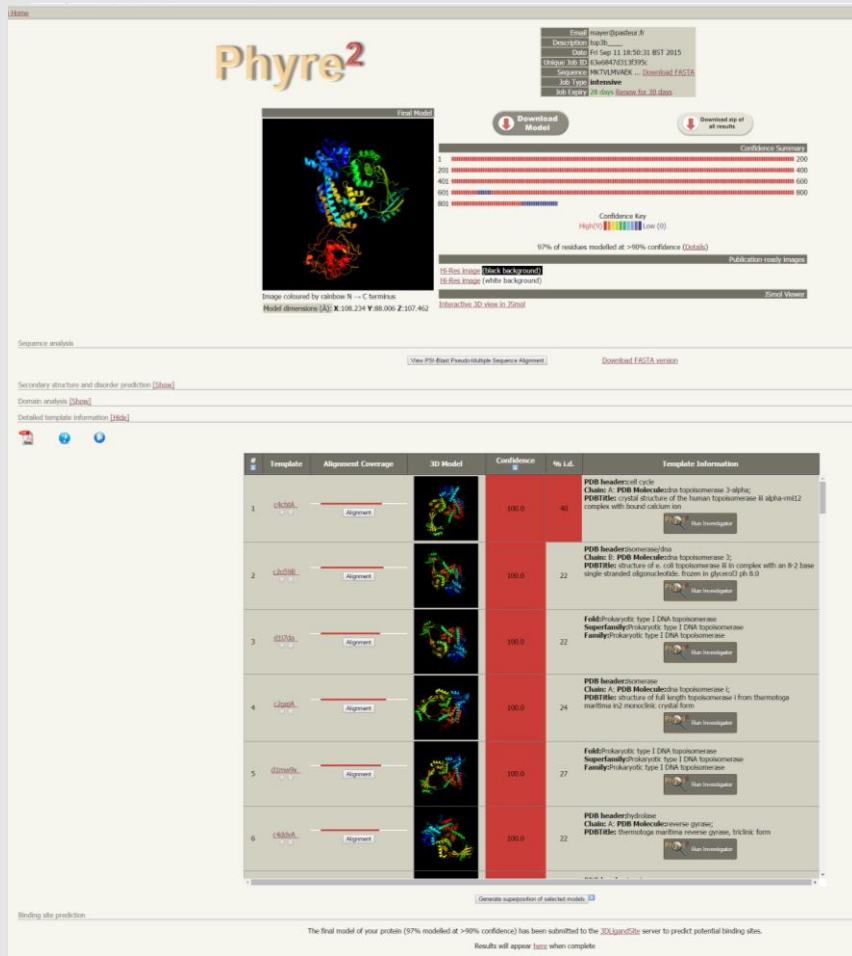
I-TASSER

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

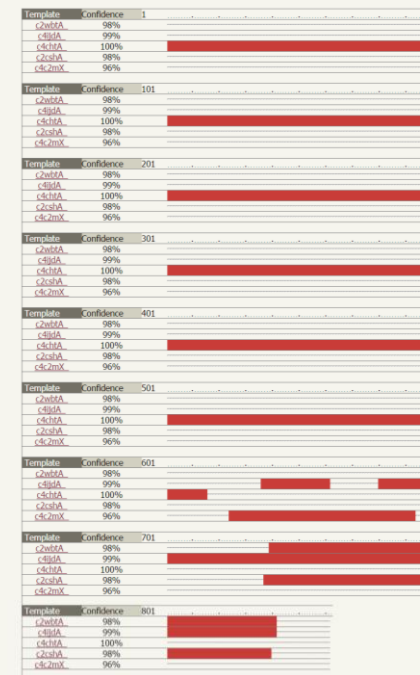
Homology modeling

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**



5 templates were selected to model your protein based on heuristics to maximise confidence, percentage identity and alignment coverage. Below is a table indicating where your sequence was covered by each template, colour-coded by the confidence of the match to that template overall.
29 residues were modelled by ab initio. Please note: ab initio modelling is **highly** unreliable.



Phyre is for non-commercial use only

Commercial users please contact [Michael Sternberg](#)

Please cite: The Phyre2 web portal for protein modeling, prediction and analysis.

Kelley LA et al. *Nature Protocols* 10, 845-858 (2015) [pdf] [Citation link]

If you use the binding site predictions from 3DLipaseSite, please also cite:

3DLipaseSite: predicting ligand-binding sites using similar structures.
Wass MN, Kelley LA and Sternberg MJ *Nucleic Acids Research* 38, W469-73 (2010) [PubMed]

© Structural Bioinformatics Group
Imperial College London
Laurence Kelley, Michael Sternberg
Disclaimer
Terms and Conditions

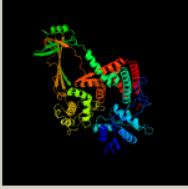
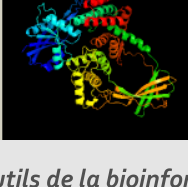
Component software
Template selection: [clustal](#) 1.3.1
Secondary structure prediction: [Pipred](#) 2.3
Disorder prediction: [Disorder](#) 2.4
Transmembrane prediction: [TMtool](#) 5.01
Multi-template modelling and ab initio: [I-TASSER](#) 1.0



Homology modeling

Phyre2

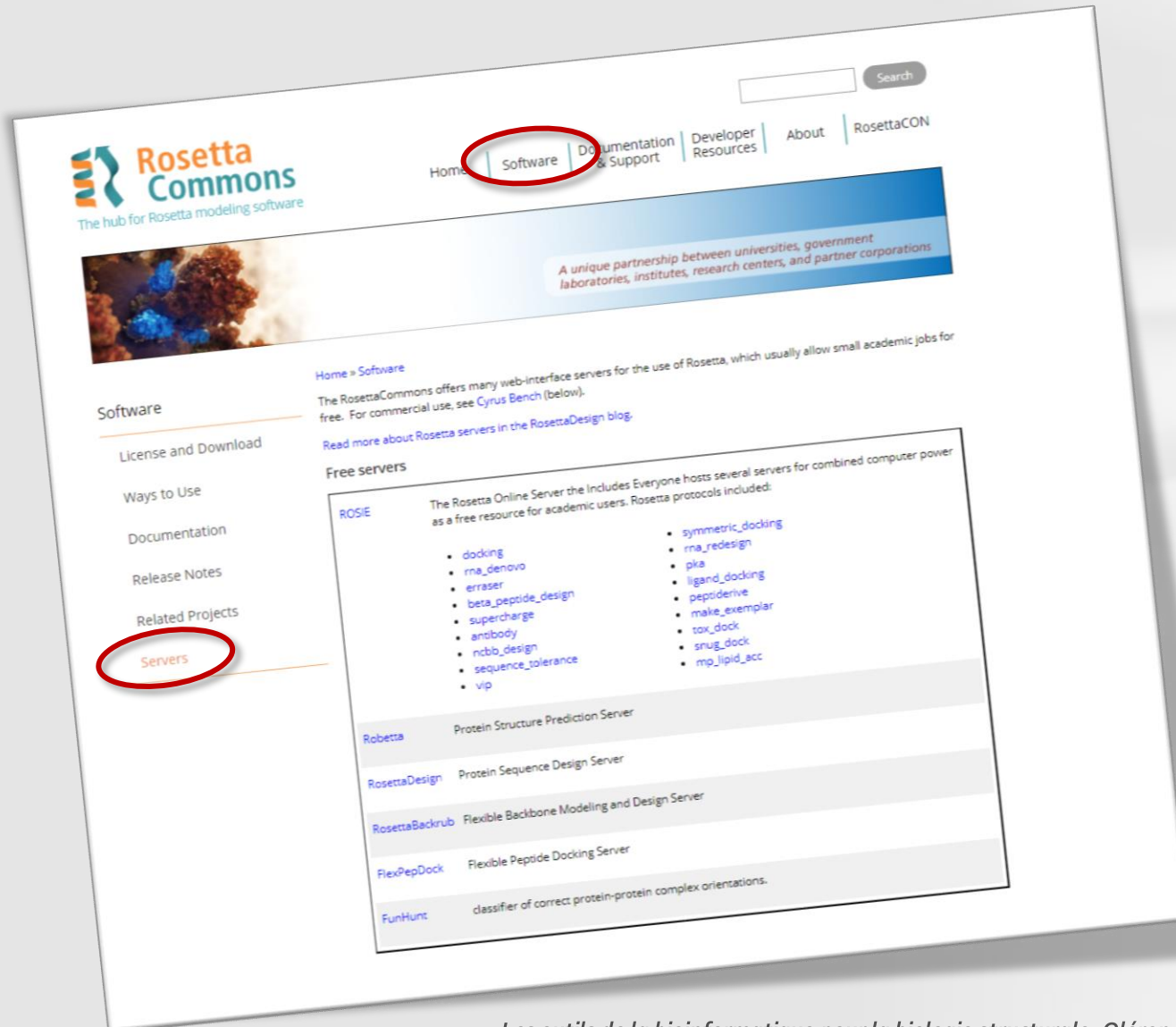
Sequence analysis
Secondary structure and domain analysis (Show)
Detailed template information
Binding site prediction

#	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 Phyre2 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 Phyre2 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 Phyre2 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 Phyre2 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 Phyre2 Run Investigator
6	c4ddvA ☐ ☐	Alignment		100.0	22	PDB header: hydrolase Chain: A: PDB Molecule: reverse gyrase; PDBTitle: thermotoga maritima reverse gyrase, triclinic form Phyre2 Run Investigator

ls of
, and are

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/>

ERRAT plot gives which residue is creating problem in the model

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/>

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/>

The screenshot displays the CAMEO3D website interface. At the top, the header includes the CAMEO logo, the tagline 'Continuous Automated Model EvaluatiOn', and navigation links for 'Try our BETA!', 'Home', '3D - Protein Structure', 'QE - Model Quality Estimation', 'CP - Contact Prediction', and a 'More' dropdown. A 'Login' button is located in the top right corner. Below the header, three main service boxes are featured: '3D - Protein Structure' (280 weeks, 4922 targets, 8 predictors), 'QE - Model Quality Estimation' (170 weeks, 23262 structural models, 18 predictors), and 'CP - Contact Prediction' (12 weeks, 68 targets, 4 predictors). A star icon indicates that predictions are evaluated against reference structures from the PDB on a weekly basis. A section titled 'CAMEO is a community project' explains the project's goals and invites user suggestions. A 'Join CAMEO today...' section lists registered servers and provides contact information for direct inquiries.

CAMEO | Continuous Automated Model EvaluatiOn

★ Try our BETA! Home 3D - Protein Structure QE - Model Quality Estimation CP - Contact Prediction More ▾

CAMEO continuously evaluate the accuracy and reliability of predictions

3D - Protein Structure
280 weeks, 4922 targets, 8 predictors.

QE - Model Quality Estimation
170 weeks, 23262 structural models, 18 predictors.

CP - Contact Prediction
12 weeks, 68 targets, 4 predictors.

★ Predictions in all categories are evaluated against reference structures released by the PDB on a weekly basis.

CAMEO is a community project

→ CAMEO continuously applies quality assessment criteria established by the protein structure prediction community. Since the accuracy requirements for different scientific applications vary, there is no "one fits all" score. CAMEO therefore offers a variety of scores

- assessing different aspects of a prediction (coverage, local accuracy, completeness, etc.) to reflect these requirements.

→ CAMEO is a community project - please feel free to suggest additional/alternative ways how CAMEO can support users and developers of structure prediction.

Join CAMEO today...

We invite developers of prediction methods to participate in the continuous evaluation by registering their servers **[REGISTER]**. We also invite developers of scoring and evaluation methods to suggest alternative scoring schemes. Please contact us **directly**.

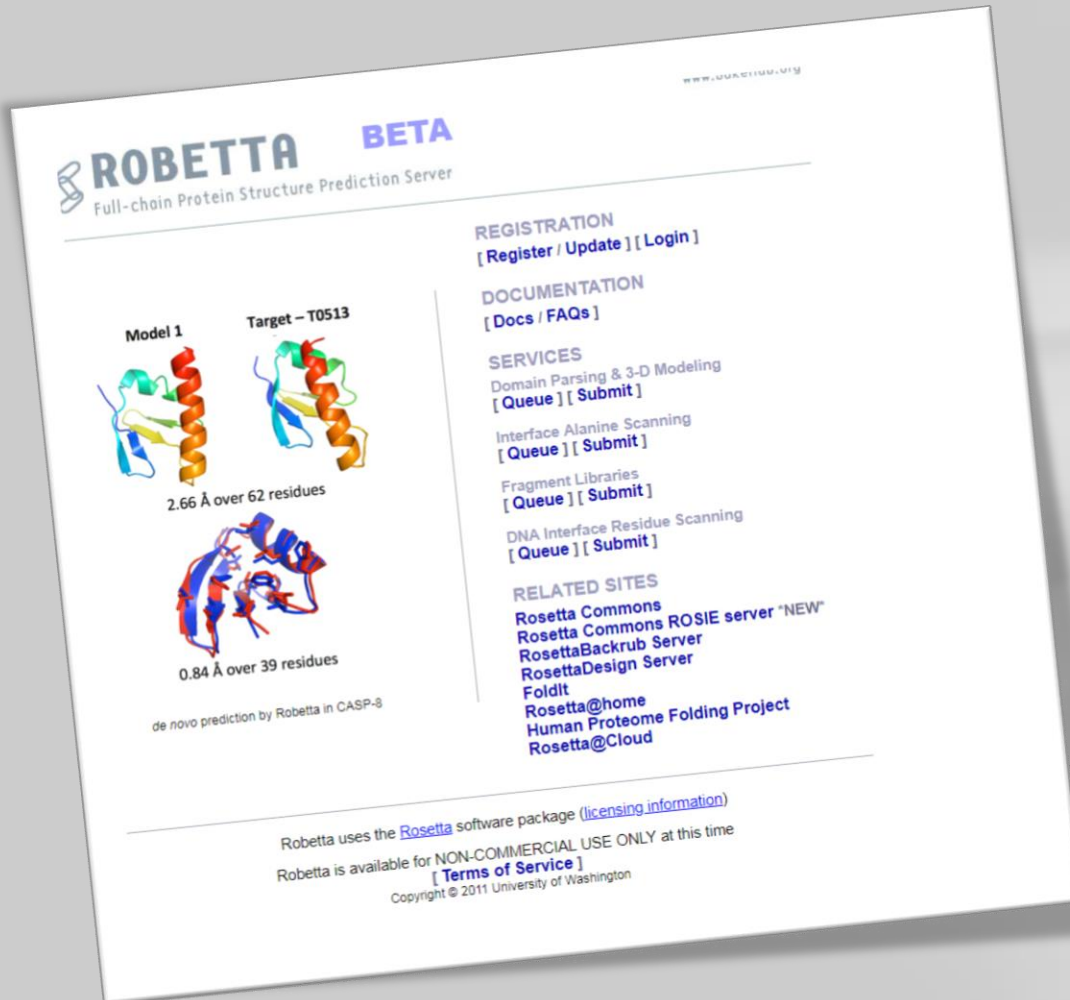
Servers of the following groups are registered so far:

A. Sali *, L. McGuffin *, T. Schwede *, J. Soeding *, D. Baker *, A. Fiser *, M. Sternberg *, Y. Zhang *, C. Floudas *, S. Tosatto *, J. Xu *, Y. Zhou *, O. Brock *, B. Wallner *, A. Elofsson *, D. Labudde *, C. Venclovas *, J. Cheng *, O. Taştan Bishop *, Y. An-Suei *

Ab initio modeling

Full-chain protein structure prediction server

<http://rosetta.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**

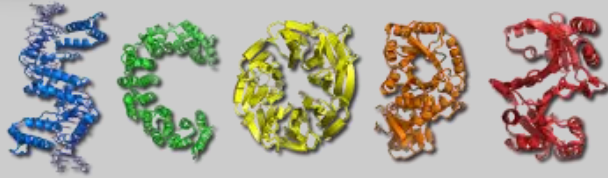
<http://www.ebi.ac.uk/msd-srv/ssm/>

The screenshot shows the PDBeFold website. The header includes the EMBL-EBI logo, 'Protein Data Bank in Europe', and 'Bringing Structure to Biology'. The main content area is titled 'PDBeFold. Structure Similarity.' and lists 'PDBeFold functionality:' with several bullet points: pairwise comparison and 3D alignment of protein structures, multiple comparison and 3D alignment of protein structures, examination of a protein structure for similarity with the whole PDB archive or SCOP archive, best Co-alignment of compared structures, download and visualisation of best-superposed structures using Rasmol, Rastop, and jmol, and linking the results to other services. A 'Launch PDBeFold' button is present. A sidebar on the left contains 'PDBeFold links' (FAQ, Visualisation, Performance, Privacy, Version log, PDBeFold Links, Comparisons, Publications, PDBeFold tutorial) and 'Other links' (PDBePISA, CCP4, CoorLib, Rasmol, Rastop). The footer contains a disclaimer and a feedback link.

The screenshot shows the PDBeFold submission form. The header includes the EMBL-EBI logo, 'Protein Data Bank in Europe', and 'Bringing Structure to Biology'. The main content area is titled 'PDBeFold' and contains a 'Submission Form' with a 'pairwise' tab selected. The form has two sections: 'Query' and 'Target'. The 'Query' section has a 'Source' dropdown set to 'PDB entry' and a 'PDB code' input field with '1J26' entered. The 'Target' section has a 'Source' dropdown set to 'Whole PDB archive'. Below these sections are checkboxes for 'match individual chains', 'match connectivity', 'best matches only', and 'unique matches only'. There are also input fields for 'Chains', 'Lowest acceptable match (%)', and 'Precision'. A 'Submit your query' button is at the bottom. The footer contains the PDBeFold logo, 'PDBe is a member of' with logos for EMBL-EBI, EBI, and PDB, and a list of links for 'Services', 'Research', 'Training', 'Industry', and 'About us'.

Structural similarity

→ Protein structure classification

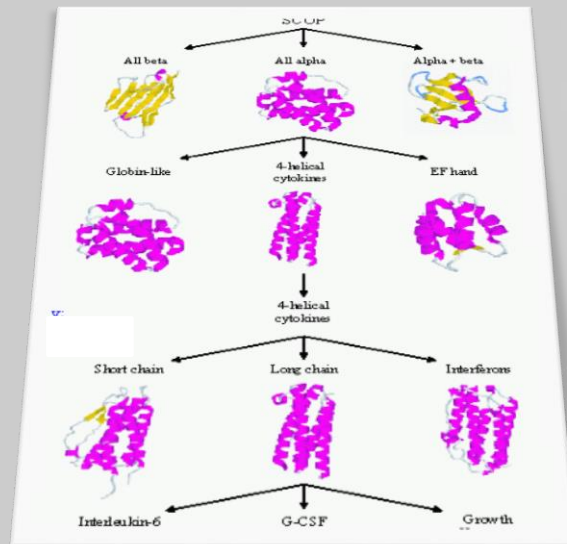


<http://scop2.mrc-lmb.cam.ac.uk/>



<http://www.cathdb.info/>

Phylogenetic tree of all the existing folds

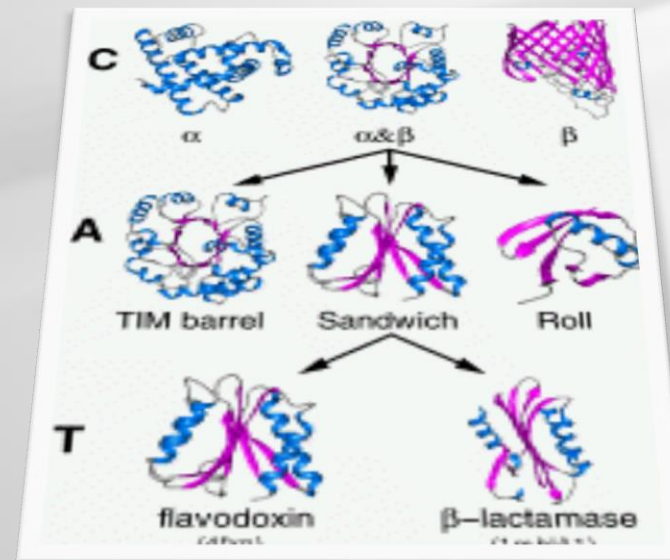


3902 Familles

7 Classes

1395 Plis

1962 Superfamilles



2700 superfamilles

3 Classes

40 Architectures

1400 Topologies

Structural similarity

➔ Matching of protein structures

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

The screenshot displays the RCSB PDB website interface. At the top, there is a navigation bar with links for Deposit, Search, Visualize, Analyze, Download, Learn, and More. A search bar is prominently featured, allowing users to search by PDB ID, author, macromolecule, sequence, or ligands. Below the search bar, the PDB logo and a brief description of the database are visible. The main content area shows the details for PDB entry 3IFZ, including its title, resolution, and a list of structural similarities. A dropdown menu is open, showing various comparison methods such as Pairwise Sequence Alignment (blast2seq, Smith-Waterman, Needleman-Wunsch) and Pairwise Structure Alignment (jFATCAT - rigid, jFATCAT - flexible, JCE algorithm, JCE Circular Permutation, external server: FATCAT, external server: MammOTH, external server: TM-Align, external server: TopMatch, external server: Dali).

RCSB PDB Deposit Search Visualize Analyze Download Learn More

Search by PDB ID, author, macromolecule, sequence, or ligands

Go

Advanced Search | Browse by Annotations

140591 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Structure Summary 3D View Annotations Sequence Sequence Similarity Structure Similarity Experiment

Display Files Download Files

3IFZ
crystal structure of the first part of the Mycobacterium tuberculosis DNA gyrase reaction core: the breakage and reunion domain at 2.7 Å resolution

Structural Similarities for the Entities in PDB 3IFZ

The following structural similarities have been found using the jFATCAT-rigid algorithm [1,2]. To reduce the number of hits, a 40% sequence identity clustering has been applied and a representative chain taken from each cluster. If the representative chain consists of multiple domains, each domain is included in the search. If available, the SCOP 1.75 domain assignment [3] is used. Otherwise algorithmic domain assignments are computed using the ProteinDomainParser [4].

Currently viewing only **significant results** (P-value < 0.001). Show all available results.

Cluster data are up-to-date as of **Sep-27-2017**

Entity #1 | Chains A,B

Description: DNA gyrase subunit A protein | **Length:** 508
No structure alignment results are available for 3IFZ A, 3IFZ B explicitly. These chains are represented by chain 3ILWA which is 100% sequence identical.

Show structure comparison results for representative 3ILWA

View how chain 3IFZ B compares with the representative chain PDB:3ILWA. Select a comparison method

Info & Help Documentation
View Table Legend
A detailed description of the procedure for the all vs. all alignments is available.

--- Select Comparison Method ---

Pairwise Sequence Alignment
blast2seq
Smith-Waterman
Needleman-Wunsch

Pairwise Structure Alignment
jFATCAT - rigid
jFATCAT - flexible
JCE algorithm
JCE Circular Permutation
external server: FATCAT
external server: MammOTH
external server: TM-Align
external server: TopMatch
external server: Dali

--- Select Comparison Method ---

Structural similarity

➔ Matching of protein structures

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

Entity #1 | Chains A,B

Description: DNA gyrase subunit A protein | **Length:** 508
No structure alignment results are available for 3IFZ.A, 3IFZ.B explicitly.
These chains are represented by chain 3ILW.A which is 100% sequence identical.

Hide structure comparison results for representative 3ILW.A

Results for domain PDP:3ILWAc

PDP:3ILWAc (chain 1) vs. representatives of other sequence clusters (chain 2)

Rank	Results	Domain 2	Title	P-value	Score	Rmsd	Len1	Len2	%ID	%Cov1	%Cov2
1	view	PDP:2XCSB	DNA GYRASE SUBUNIT B, DNA	0.0	530.38	2.08	225	203	52	88	98
2	view	PDP:4I3Haf	Topoisomerase IV subunit B, DNA	1.11E-16	524.57	2.28	225	209	34	92	99
3	view	PDP:1ZVUA	Topoisomerase IV subunit A	2.44E-15	378.45	1.43	225	159	34	70	99
4	view	PDP:4GFHA	DNA topoisomerase 2	5.36E-7	270.08	3.15	225	184	15	73	90
5	view	d2fcwa1	Alpha-2-macroglobulin receptor-as	4.72E-4	206.01	3.58	225	105	5	31	67
6	view	PDP:4OYDB	Computationally designed Inhibitor	6.47E-4	210.05	6.15	225	117	3	52	100
7	view	PDP:4GFQA	Ribosome-recycling factor	8.49E-4	188.21	4.7	225	111	2	45	91
8	view	PDP:3LF9Aa	4E10_D0_1S1A_001_C (T161)	9.1E-4	200.09	5.48	225	120	5	48	90

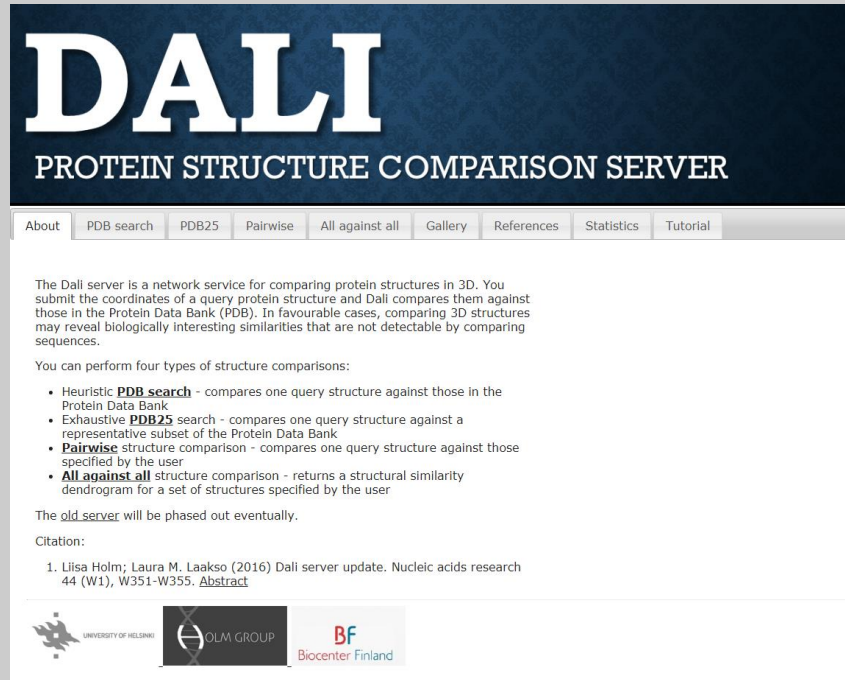
Filter Results Reload Results Page 1 of 1 15 View 1 - 8 of 8

View how chain 3IFZ.B compares with the representative chain PDP:3ILWAc. Select a comparison method: --- Select Comparison Method ---

Structural similarity

3D comparison - Dali server

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



DALI
PROTEIN STRUCTURE COMPARISON SERVER

About PDB search PDB25 Pairwise All against all Gallery References Statistics Tutorial

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

You can perform four types of structure comparisons:

- Heuristic **PDB search** - compares one query structure against those in the Protein Data Bank
- Exhaustive **PDB25** search - compares one query structure against a representative subset of the Protein Data Bank
- **Pairwise** structure comparison - compares one query structure against those specified by the user
- **All against all** structure comparison - returns a structural similarity dendrogram for a set of structures specified by the user

The old server will be phased out eventually.

Citation:

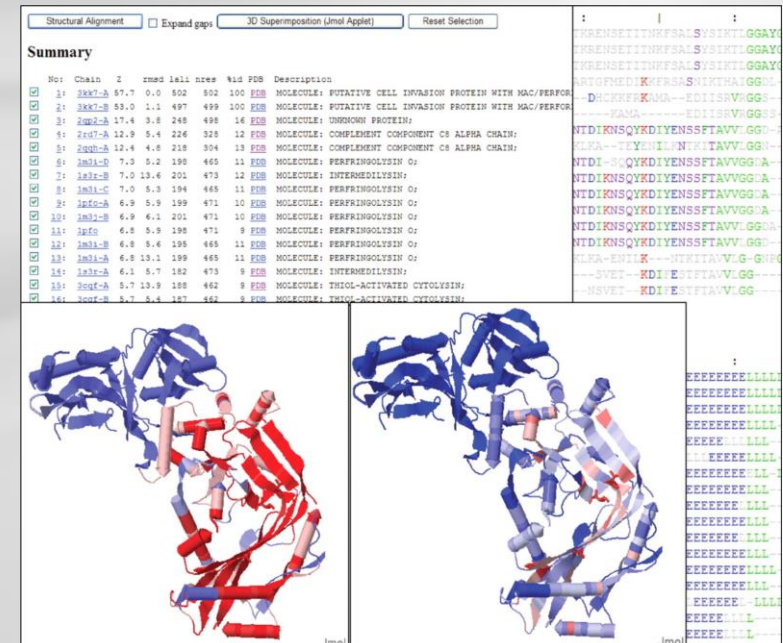
1. Liisa Holm; Laura M. Laakso (2016) Dali server update. Nucleic acids research 44 (W1), W351-W355. [Abstract](#)

UNIVERSITY OF HELSINKI HOLM GROUP Biocenter Finland

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)

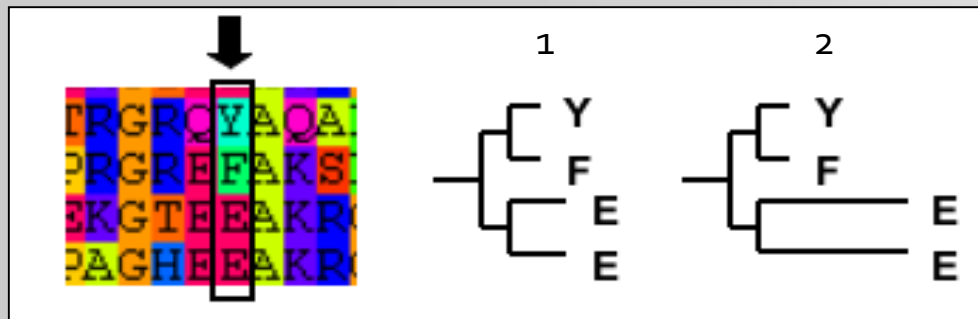
Stability: prediction of free energy changes between alternative structures

Analysis of the conservations

ConSurf : Analyse automatique des conservations

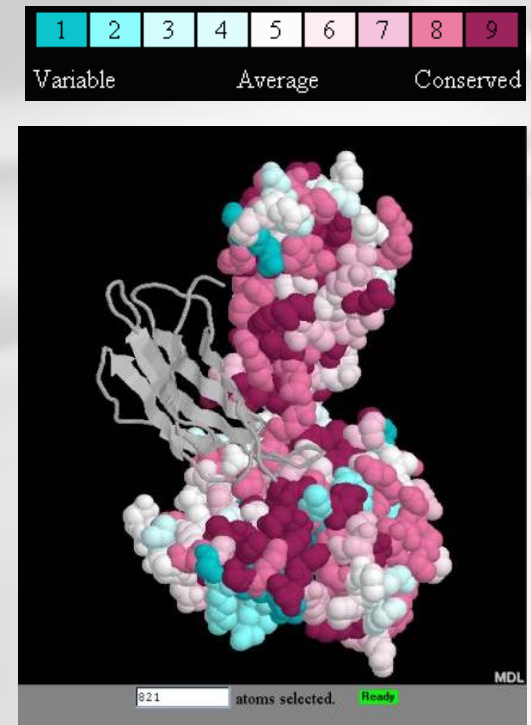
<http://consurf.tau.ac.il/>

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



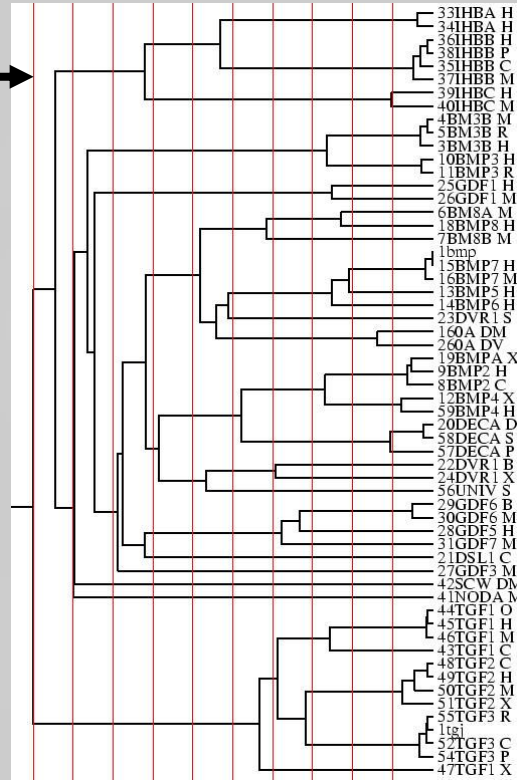
Score de conservation dans la colonne des facteurs B du fichier pdb

Structure and evolution

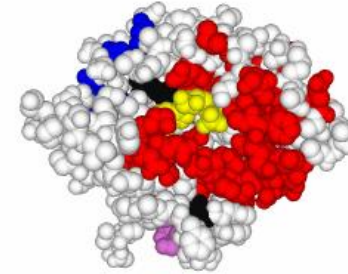
Analyse des traces évolutives

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

SNWLPEGPYPDPLTGIDSDVPLA EHGVOQAKELAHY
SNWLPGGPYPPTGIDNDVPLSEHGVEQAHIELANY
SNWLPP-PHPPNPTGIDSDPALAPHGVQOAOOLAAAY
SEFNLLGK-----IGGDSGLSVRGKQFAQALKKF
SEFNLLGK-----IGGDSGLSLRGKQFAQALKKF
S-WNQENFFCG-----WFDALSEKGAEEAKFGATA
SEFNLLGK-----IGGDSGLSVRGKQFAQALKKF
SEFNLLGK-----IGGDSGLSVRGKQFAQALKKF
SEFNLLGK-----IGGDSGLSARGKQYAYALANF
SEFNLLGK-----IGGDSGLSVRGKQFAQALKKF
STWNQENFFCG-----WFDALSEKGTTEEAKFGAKA
SOLNLKG-----IGGDAGLSTGKQYAOALAEF
SEFNLLGK-----IGGDPGLSPGKEFAKSLAQF
TTWNQENFFCG-----WFDALSEKGTTEEAKFGAKA
SAWNLNFFSG-----WYDADLSPAGHEEAKFGGQA
SEFNLLGK-----IGGDPGLSPGKEFAKSLAQF
SAWNLNFFSG-----WYDADLSPAGHEEAKFGGQA
SEFNLLGK-----IGGDSGLSVRGKQYAHLEGNF
SEWNEKNLFTG-----WVDVNLTPKGEKEATFGGEL
SAWNLNFFSG-----WYDADLSPAGHEEAKFGGQA
SEWNLNFFSG-----WVDVNLTEQGVQOEA TAGGKA

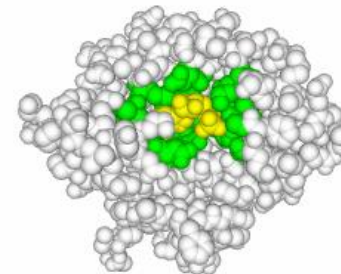


« Evolutionary Traces »



Patches colorés = classes d'aa

Epitope fonctionnel



<http://lichtargelab.org/>

Analyse de l'arbre phylogénétique
par partition de classes

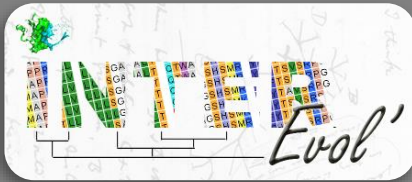
Yao, H., ...and O. Lichtarge
(2003) J Mol Biol. 326:255-261

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

InterEvol:

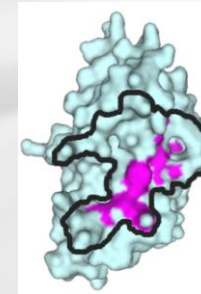
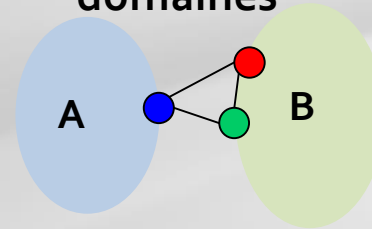
~18,000 non-redundant interfaces
among which
~4,000 heteromeric interfaces

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

<http://biodev.cea.fr/interevol>

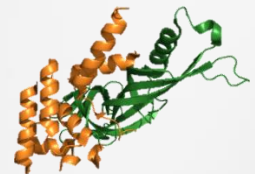
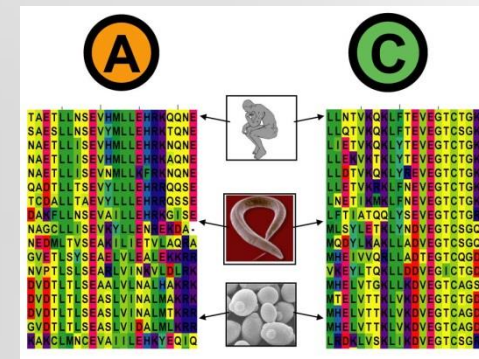
**InterEvScore : Discriminate
co-evolved interfaces**

Contacts multi-
domaines



Patches
apolaires

Information
d'évolution

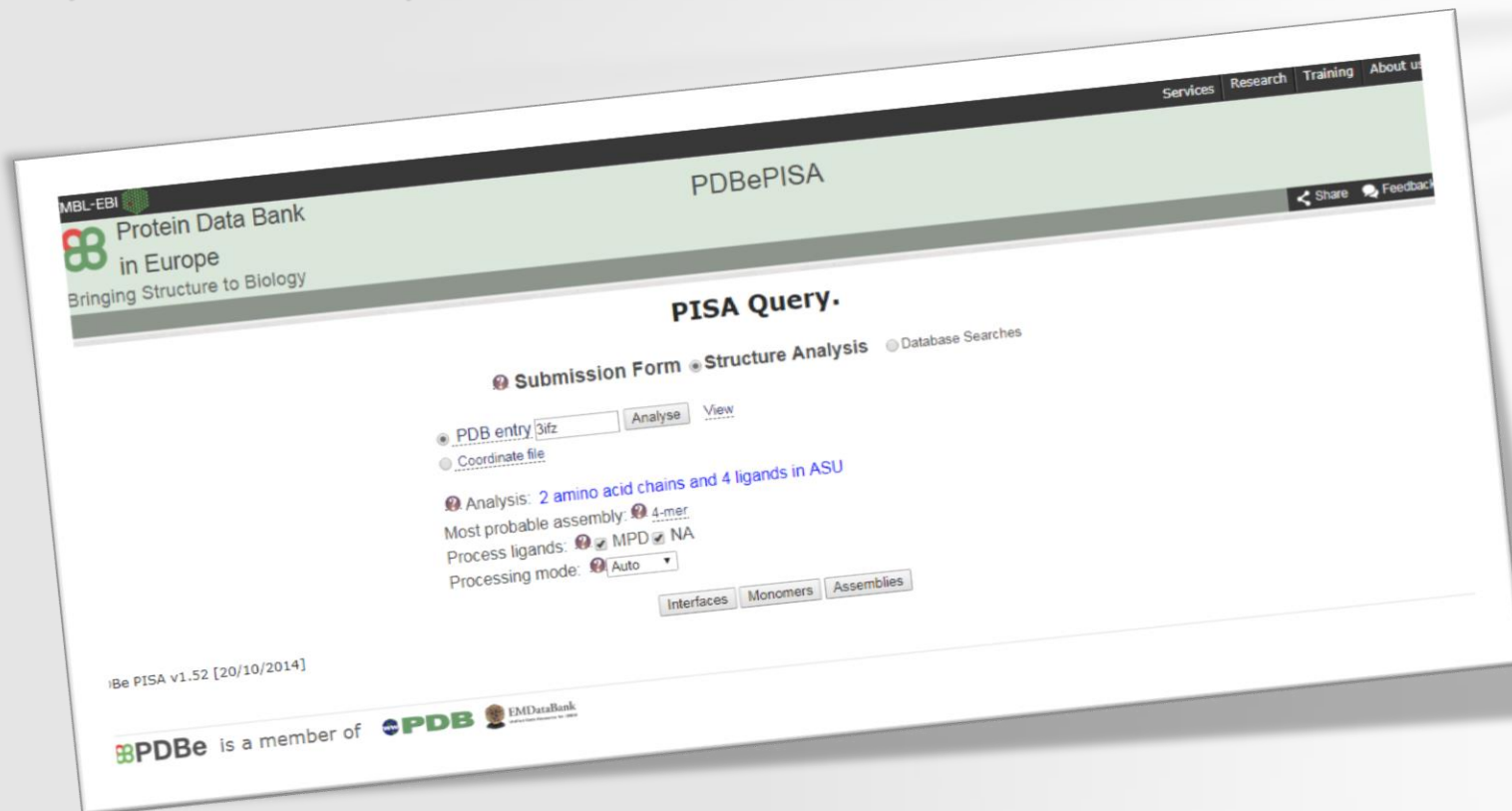


Structure and interactions

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 Å RESOLUTION

Interfaces XML View Details Download Search

##			Structure 1			x	Structure 2			interface		ΔG	ΔG	N_{HB}	N_{SB}	N_{DS}	CSS		
Id	NN	«»	Range	N _{at}	N _{res}	Surface Å ²	Range	Symmetry op-n	Sym.ID	N _{at}	N _{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	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

Prediction 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 complementary 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](#)]



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:

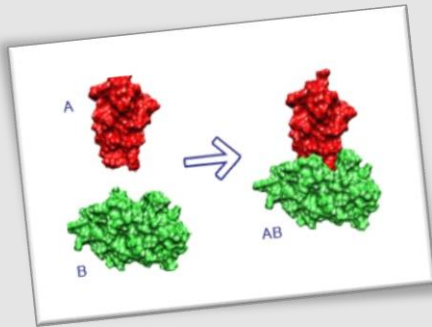
- Ambrish Roy, Jianyi Yang, and Yang Zhang. *COFACTOR: An accurate comparative algorithm for structure-based protein function annotation*. *Nucleic Acids Research*, 40:W471-W477 (2012). ([download the PDF file](#))
- Ambrish Roy, Yang Zhang. *Recognizing protein-ligand binding sites by global structural 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

Patchdock

Firedock

ClusPro

Zdock

Prediction of interaction from 2 structural models

Docking based on surface complementarity , easy to use

docking protein-protein, easy to use

tops the competition in the latest rounds of CAPRI experiment

rigid-body search of docking orientations

Stand-alone

HADDOCK

HEX

Docking with possibility to implement experimental data (mutagenesis, cross-linking, NMR chemical shift, ...)

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



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**.

Version: 10.5

[LOGIN](#)
[REGISTER](#)

STRING

[Search](#)
[Download](#)
[Help](#)
[My Data](#)

[Viewers >](#)
[Legend >](#)
[Settings >](#)
[Analysis >](#)
[Exports >](#)
[Clusters >](#)
[More](#)
[Less](#)

Basic Settings

meaning of network edges:

☒ evidence (line color indicates the type of interaction evidence)

☐ confidence (line thickness indicates the strength of data support)

☐ molecular function (line color indicates the molecular function)

active interaction sources:

☒ Textmining

☒ Experiments

☒ Databases

☒ Co-expression

☒ Neighborhood

☒ Gene Fusion

☒ Co-occurrence

minimum required interaction score:

minimum confidence (0.400)

max number of interactors to show:

1st shell:

2nd shell:

Advanced Settings

network display mode:

☐ static png (network is a simple bitmap image, not interactive)

☒ interactive svg (network is a scalable vector graphic [SVG], interactive)

display simplifications:

☐ disable structure previews inside network bubbles

☐ hide disconnected nodes in the network

☐ hide node labels

UPDATE

Server load: low (3%) [HD]

CC BY-NC-ND 4.0

The figure displays a dense network of biological entities, likely genes or proteins, represented as nodes connected by edges. The nodes are color-coded: blue, green, orange, and red. The network shows a high degree of connectivity, with many nodes having multiple neighbors. Key clusters include a group of blue nodes at the top left, a large cluster of orange and red nodes in the center, and a group of green nodes at the bottom right. The labels on the nodes include: GOL2, GZMK, KGF22, TMEM508, LANCL3, AVP1, APOM, BAIAP1, LINC43, B2M1P1, YAF3, NRP2B, CNP, RNF192, TSPAN8, SODX17, ADRFRL2, GFOD2, CONL33, ARL4A, CLYSL1, CDC20B, CSPL1, EPS1S1, EFS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI2, EVC, ZFY4, E7F5, PCOY2, CEP152, NOC46, SPAG17, EPS1S, RAB11L1, TSC22D2, SFRBP1, FCHSD1, DTNA, PLAG1, CILHRG5, CCDC67, KIAA1731, CCDC78, KIF24, KMT21, CMNC, MCN, CSPH3, HAUS3, CNTN, RNF14C, HMOX1, CADM1, S100P, MIRIP, CF2L, HAND1, CHCHD2, MARCH3, DNK2, TRINAJAP, CDCE92, GALNT14, ANKRKD61, ANKRK49, SER4, SAPCD2, PHACTR1, HIPK1, PTPN7, EREG, MSK1, CPSPF7, EVI

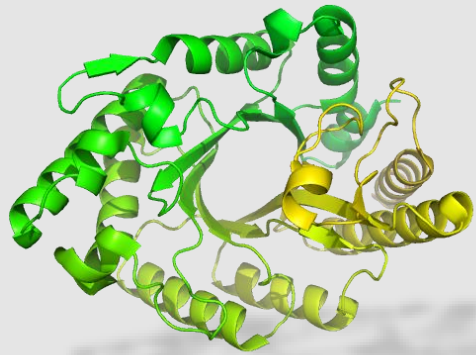
Protein thermostability

Effect of the temperature on protein stability

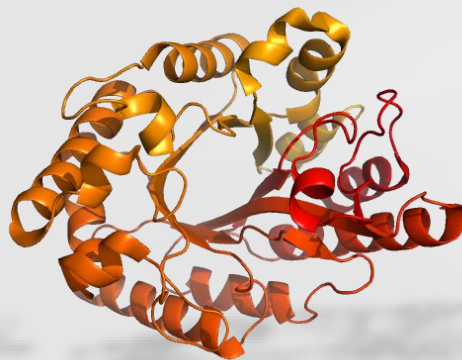
A powerful method is the comparison of mesophilic and thermostable homologous proteins

- ➔ 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

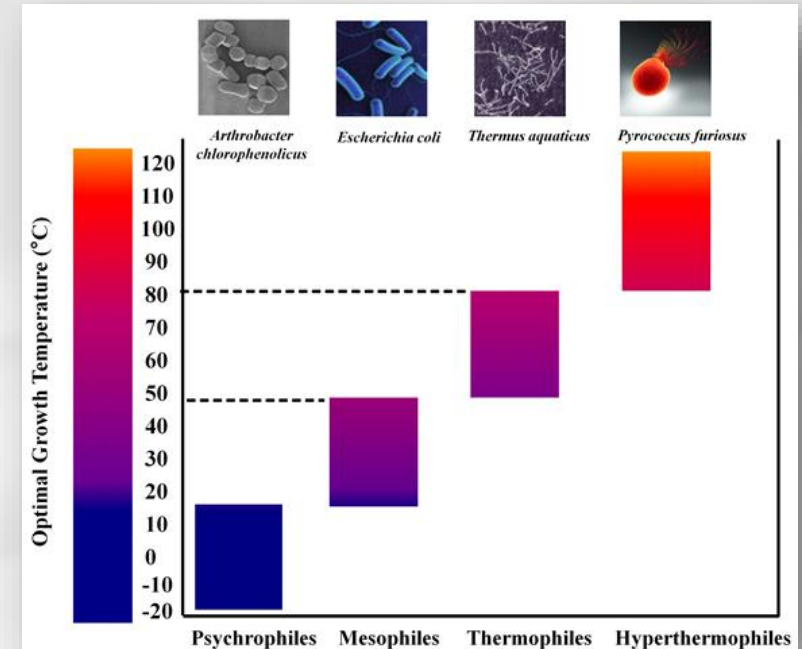
How to increase the thermostability of target proteins ?



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 T_m changes by mutations

ProTherm
Thermodynamic Database
for Proteins and Mutants

Data updated: Feb. 22 2013

[Overview](#)

ProTherm is a collection of numerical data of thermodynamic parameters such as Gibbs free energy change, enthalpy change, heat capacity change, transition temperature etc. for wild type and mutant proteins, that are important for understanding the structure and stability of proteins. It also contains information about secondary structure and accessibility of wild type residues, experimental conditions (pH, temperature, buffer, ion and protein concentration), measurements and methods used for each data, and activity information (Km and Kcat).

ProTherm is cross-linked with sequence databases (PIR and SWISS-PROT), structural database (Protein Data Bank), functional database (Protein Mutant Database), and literature database (PubMed). Moreover, the thermodynamic information is integrated with structural and functional information through the relational database, [3DinSight](#). The WWW interface enables users to search data based on various terms with different sorting options for outputs, and view three dimensional structures with automatically mapped mutation sites and surrounding amino acids. For more detail about ProTherm, please see [here](#).

Please note that this database is under constant development. There will be changes without prior notice. We welcome your comments and suggestions to improve this database.

[Home](#) | [ProTherm](#) | [ProNIT](#)
| [Biomolecules Gallery](#)

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

PDB ID Chain Wild-type Position Mutant Temperature pH (°C, 0-100) (0-14)

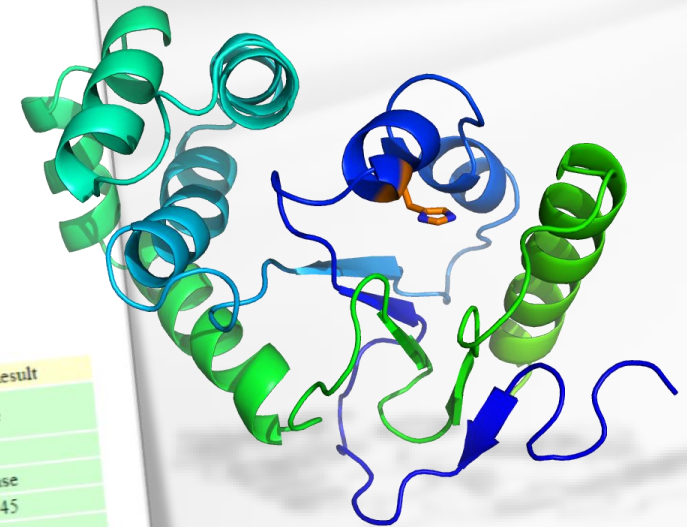
1A23 A H 32(32) Y 25 7

Protein sequence(no headers)

AQYEDGKQYTTLEKPVAGAPQVLEFFSFFCPHCYQFEVLHISDNVKKLPEGVKNTKYHVFH
GGDLGKDLTQAHAVAMALGVEDKVTPLFEGVQKTQTIASASDIRDFINAGIKGEEYDAAMNS
FVVKSLVAQQEKAADVQLRGVPAMFVNGYQLNPGSHDTSNHDFVQYADTVKYLSEKK

Clear Submit

Predictor	i-Mutant2.0 PDB	i-Mutant2.0 SEQ	AUTO-MUTE SVM	Reference	MUpro	PoPMuSiC	CUPSAT	Meta Result
Result	7	31.9	5.55	helix	helix	0.99962109	null	iStable
Conf			Increased	Increased	Increased			32.44
ΔΔG	Increase	Increase	0.76			null	Increase	Increase
RSA	6.50		1.51	3.77		null	0.66	0.73345
SS			buried	buried				



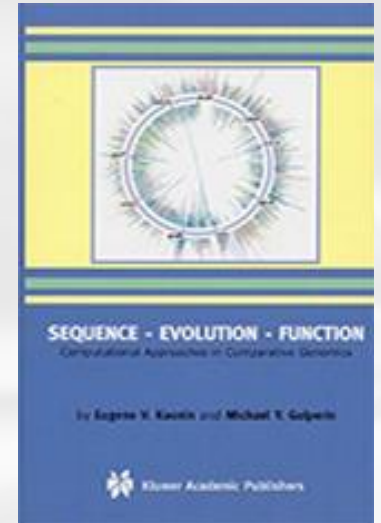
E. coli DsbA

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

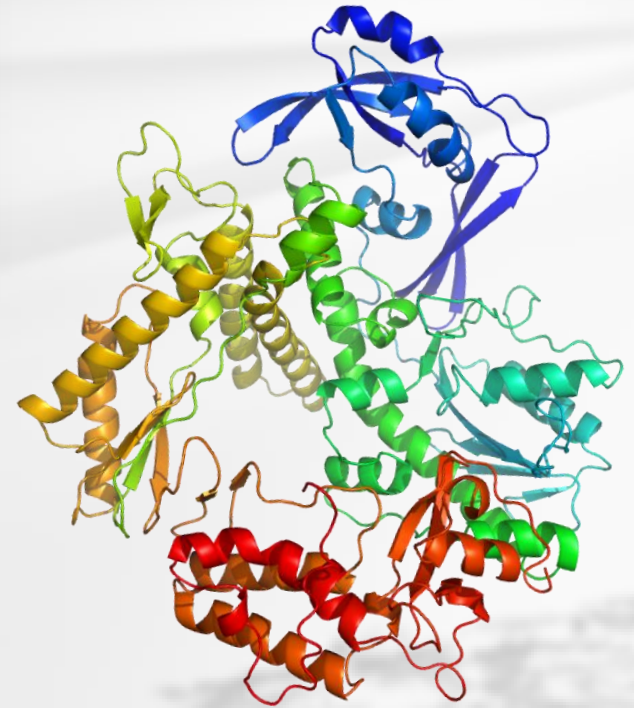
➔ **Koonin EV, Galperin MY. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics**



➔ **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