Předpověď 3D-struktury a topologie bílkovin, strukturní a funkční klasifikace

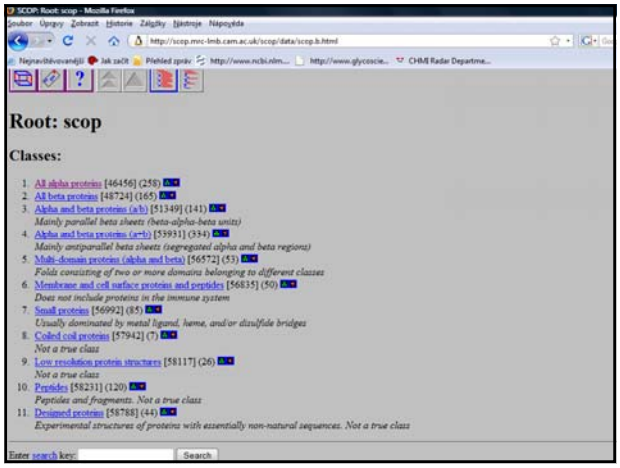
SCOP **S**tructural **C**lassification **o**f **P**roteins
(http://scop.mrc-lmb.cam.ac.uk/scop)

The **SCOP** database, created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.

Family: *Clear evolutionarily relationship*
Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. *However, in some cases similar functions and structures provide definitive evidence of common descent in the absence of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%.*

Superfamily: *Probable common evolutionary origin*
Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies. *For example, actin, the ATPase domain of the heat shock protein, and hexakinase together form a superfamily.*

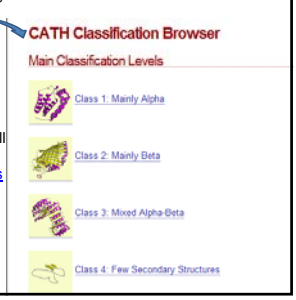
Fold: *Major structural similarity*
Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. *Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.*



CATH Protein Structure Classification (http:// www.cathdb.info)

CATH is a hierarchical classification of protein domain structures, which clusters proteins at four major levels: Class (C), Architecture (A), Topology (T) and Homologous superfamily (H). The boundaries and assignments for each protein domain are determined using a combination of automated and manual procedures which include computational techniques, empirical and statistical evidence, literature review and expert analysis

Class (C), Architecture (A) - the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures., Topology (T) - the same overall shape and connectivity of the secondary structures in the domain core Homologous superfamily (H) - share a common ancestor (Similarities are identified either by high sequence identity or structure comparison)



What are protein domains?

Since the first protein structures were solved, it was apparent that the polypeptide chain could often fold into one or more distinct regions of structure. Such substructures, or domains, are considered as the basic units of folding, function and evolution and often have similar chain topologies (Holm & Sander, 1994). Protein domains are often considered as independent or, at the least, semi-independent units, able to fold and in some cases retain function if separated from the parent chain. The independent, modular nature of many domains means that they can often be found in proteins with the same domain content, but in different orders, or in different proteins in combination with entirely different domain structures.

The concept of the protein domain is just as valid at the sequence level as the structural level. This can be shown by the fact that the alignment of sequences containing similar domains, but in different orders can result in poor and possibly misleading alignments.
However alignment of the shared domains if extracted from the parent sequence may reveal a high level of sequence similarity, demonstrating an evolutionary link between the domain sequences.

domain boundary/disorder/globularity prediction:

LinkPred (at NIMR)
SnapDRAGON domain boundary prediction (at NIMR)
PASS (at RIKEN)
Domain Guess by Size (DGS) (at NCBI)
UMA (Udwary-Merski Algorithm) (at Johns Hopkins Univ.)
DomPred (at UCL)
Domain boundary prediction based on entropy profile (at IPR, Moscow]
GlobPlot (at EMBL) Prediction of protein disorder/order/globularity
DisEMBL (at EMBL) Protein disorder prediction

Příklad: Předpověď spojovacích úseků mezi doménami - program DomCut

Předpovídání doménových a spojovacích oblastí v sekvencích proteinů

Domény = funkční jednotky, z nichž jsou bílkoviny složeny

Linker = spojovací úsek aminokyselinového řetězce spojujícího dvě sousední domény

DomCut

- Metoda programu DomCut vychází ze statisticky potvrzeného předpokladu odlišného složení doménových a linkerových úseku v řetězcích aminokyselin.
- Jestliže známe relativní frekvence výskytu jednotlivých AK v linkerových a doménových úsecích, můžeme u neznámé sekvence odhadnout zda je ten či onen úsek spíše linker nebo doména, podle toho, zda v něm převládají AK vyskytující se více v linkerech nebo v doménách.
- Pro vyjádření přednosti AK v linkerech je definován tzv. „linker index" *S* (f_i^{linker} a f_i^{domain} je frekvence zastoupení aminokyseliny *i* v úsecích linkeru a domény)
- :

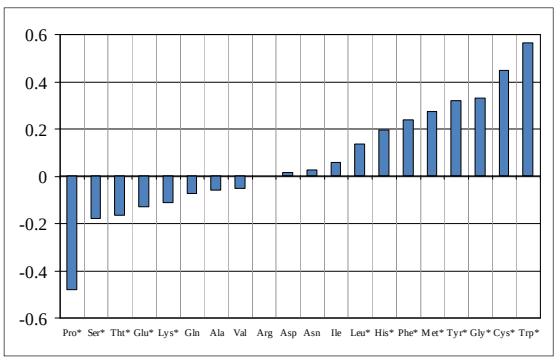
$$S_i = - ln \frac{f_i^{linker}}{f_i^{domain}}$$

Četnost výskytu jednotlivých aminokyselin v doménách a linkerech

- Záporná hodnota znamená, že daná AK se častěji vyskytuje v linkerových úsecích
- Výjimku tvoří Gly, který je hojně zastoupený v doménách, ale je častým prvkem v linkerových oblastech – zajišťuje „ohebnost"

Aminokyselina		f_i^{linker} (%)	f_i^{domain} (%)	S_i	Aminokyselina		f_i^{linker} (%)	f_i^{domain} (%)	S_i
Proline	Pro*	7.95	4.93	−0.478	Asparagine	Asn	4.29	4.41	0.027
Serine	Ser*	8.32	6.97	−0.177	Isoleucine	Ile	4.86	5.16	0.060
Threonine	Thr*	6.68	5.67	−0.163	Leucine	Leu*	7.62	8.75	0.138
Glutamic acid	Glu*	7.53	6.62	−0.128	Histidine	His*	2.13	2.59	0.195
Lysine	Lys*	6.30	5.64	−0.112	Phenylalanine	Phe*	2.92	3.71	0.240
Glutamine	Gln	4.35	4.04	−0.073	Methionine	Met*	1.47	1.94	0.275
Alanine	Ala	7.03	6.64	−0.058	Tyrosine	Tyr*	2.49	3.44	0.322
Valine	Val	7.33	6.96	−0.052	Glycine	Gly*	5.46	7.60	0.331
Arginine	Arg	5.39	5.39	0.000	Cysteine	Cys*	1.62	2.53	0.447
Aspartic acid	Asp	5.39	5.47	0.016	Thryptophan	Trp*	0.89	1.56	0.564

DomCut - grafické znázornění S_i faktoru



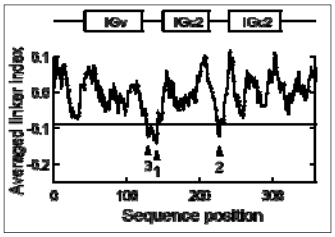
DomCut – příklad predikce spojovacích úseků

Aminokyselinová sekvence Q24372

- není podobná s žádnou z referenční množiny (podobnost <40%)
- úseky linkerů mezi doménami odpovídají jasně odhadům (prohlubně pod prahovou hodnotou –0,09)

Záznam trEMBL:

Lachesin, Contains 2 Ig-like C2-type domains, 1 Ig-like V-type domain.



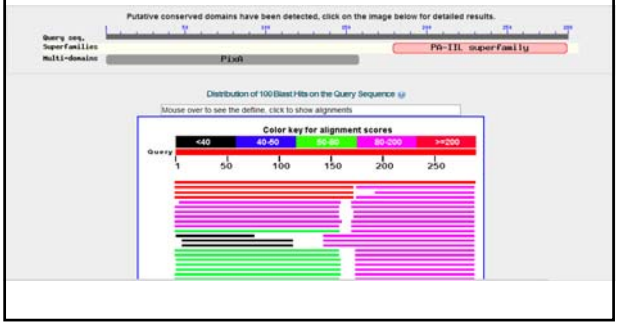
M. Suyama and O. Ohara; DomCut: prediction of inter-domain linker regions in amino acid sequences, Bioinformatics, Vol. 19,no. 5 2003 pg. 673-674

http://www.bork.embl.de/~suyama/domcut

NCBI – Blast (Basic Local Alignment Search Tool)
(National Centre for Biotechnology Information)

Prohledávání databází známých aminokyselinových sekvencí

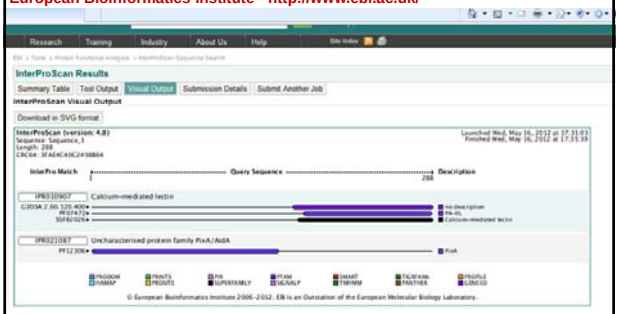
➢ **celý protein**



InterPro protein sequence analysis & classification

InterPro is an integrated database of predictive protein signatures used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures.

European Bioinformatics Institute - http://www.ebi.ac.uk/



Předpověď 3D-struktury/ foldu

- Threading - „navlékání“
- Homology modeling
- *Ab inicio* metody

Threading

- „navlékání“ = rozpoznání a přiřazení proteinového foldu aminokyselinové sekvenci
- sekvence je porovnávána s databází existujících foldů (3D profilů) a na jejich základě jsou konstruovány 3D- modely
- 3D profil - každému reziduu v 3D struktuře je přiřazena environmentální proměnná (obsah polárních atomů v postranním řetězci, skrytá plocha, sekundární elementy, apod.) vycházející z předpokladu, že okolí rezidua je více konzervováno než aminokyselina samotná.
- Reziduum může být také popsáno pomocí svých interakcí
- Výsledná kvalita modelu shoda je popsána pomocí Z-skóre nebo energie
- U multidoménových struktur je potřeba aminokyselinovou sekvenci rozdělit na jednotlivé domény a analyzovat je separátně

PHYRE (3D-PSSM)

http://www.bmm.icnet.uk/

Threading at 2D level and scoring at 3D level :
matching of secondary structure elements, and propensities of the residues in the query sequence to occupy varying levels of solvent accessibility

[Kelley et al. (2000). J. Mol. Biol. 299, 499-520]

ProFIT

http://www.proceryon.com/index.html

Threading and scoring at 3D level :

based on the use of Knowledge-Based Potentials that are derived from the database of existing structures

[Sippl & Flöckner (1996) Structure 4, 15-19]

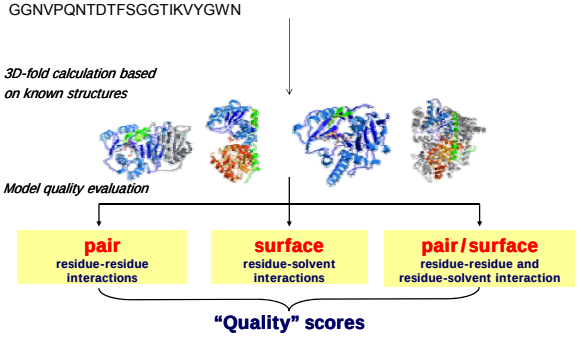
Threading

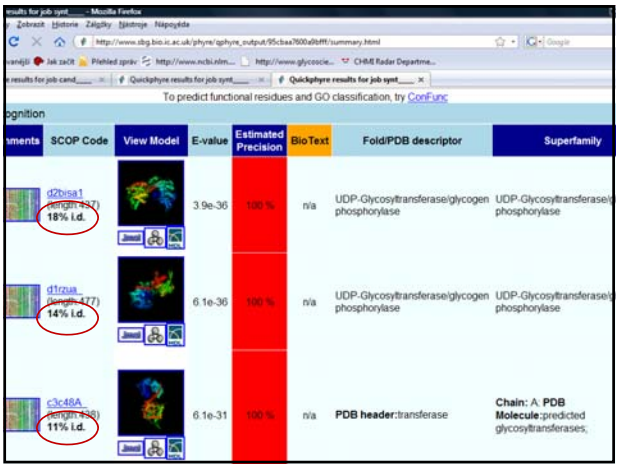
Protein **H**omology/analog**Y** **R**ecognition **E**ngine

(nástupce 3D-PSSM)

- sekvenční „alignment“ s porovnávanou strukturou
- Využívá PSSMs (position-specific scoring matrix) generovanou metodou PSI-Blast jak pro cílovou sekvenci tak sekvencemi ze známých struktur.
- Kopírování 3D souřadnic a přepis jednotlivých reziduí podle zkoumané sekvence
- Následně porovná shodu profilů cílové sekvence a porovnáváné struktury společně se shodou jejich sekundárních struktur.
- Jediné zásahy do aminokyselinové páteře templátu jsou při modelování inzercí a delecí v sekvenci oproti porovnávané struktuře.

SDVDIEAGQTLVQVVNISNGETWVAIQLPAQYRSFDLVFENVSPSTSGSVLVAQMA
PQSGGVYGSNYSGSGWGNDLGGGGFYGYSEAKWMCLWPANRSGPNSKTGIYG
TCKLMNLNQSNAVPSVTSNLFAPTAYKNEPGYANVGGCCQKIRGLASSIQFAFALH
GGNVPQNTDTFSGGTIKVYGWN





Homology modeling

•přiložení cílové sekvence se sekvencí homologního proteinu se známou 3D strukturou

• extrakce uhlíkové páteře ze struktury templátu a umístění postranních řetězců

•modelování otoček a smyček

• minimalizace energie

•validace modelované struktury

MODELLER

Mostly used program in academic environment for serious homology modeling

SWISS-MODEL

An automated knowledge-based protein modelling server

Computation of this workunit has stopped.

Please see the following log report for details:
Started: Wed May 13 06:59:31 2009 (sms_automode) Reading user input sequence **No Templates found.**
==========
Simple automated template selection could not identify suitable templates. Please use advanced Template Selection under **[Tools]** to select a template and prepare a workunit using the project mode.

Fold Databases
SCOP Structural Classification of Proteins (http://scop.mrc-lmb.cam.ac.uk/scop/)
Dali/FSSP (http://www.ebi.ac.uk/dali/)
CATH Protein Structure Classification (http:// www.cathdb.info)

Structural Alignment Tools
Vast (http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html)
CE (http://cl.sdsc.edu/ce.html)
DALI (http://www.ebi.ac.uk/dali)

Fold Prediction
3D-PSSM and PHYRE Protein Fold Recognition (http://www.sbg.bio.ic.ac.uk/~phyre/)
CPHmodels homology modeling (http://www.cbs.dtu.dk/services/CPHmodels/)
Geno3D (http://geno3d-pbil.ibcp.fr/cgi-bin/geno3d_automat.pl?page=/GENO3D/geno3d_home.html)
3D-JIGSAW (http://www.bmm.icnet.uk/~3djigsaw/)
ESyPred3D (http://www.fundp.ac.be/urbm/bioinfo/esypred/)

Fully Automatic Homology Modelling
Robetta full-chain protein structure prediction server (http://robetta.bakerlab.org/)
Swiss-Model (http://www.expasy.org/swissmod/SWISS-MODEL.html)

A zpět k doménám – proč potřebujeme jejich predikci?

•Prohledávání sekvenčních databází bez predikce domén může být neúspěšné

•Automatická predikce struktury se zaměří jen na nejlépe „definovanou" část

•….