LOSCHMIDT LABORATORIES

Analýza proteinových sekvencí

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Osnova

- Predikce fyzikálně-chemických vlastností
- Predikce membránových regionů
- Predikce motivů a domén
- Databáze motivů a domén
- Prohledávání databází motivů a domén

Predikce fyzikálně-chemických vlastností

- ExPASy (Expert Protein Analysis System)

Predikce fyzikálně-chemických vlastností

- ExPASy (Expert Protein Analysis System)
  - Molekulová hmotnost
  - Izoelektrický bod
  - Extinkční koeficient
  - Postranslační modifikace
  - Místa proteasové digesce
  - Poločas rozkladu
  - Nestabilita

Predikce fyzikálně-chemických vlastností

- ExPASy (Expert Protein Analysis System)

Identification with isoelectric point, molecular weight and/or amino acid composition

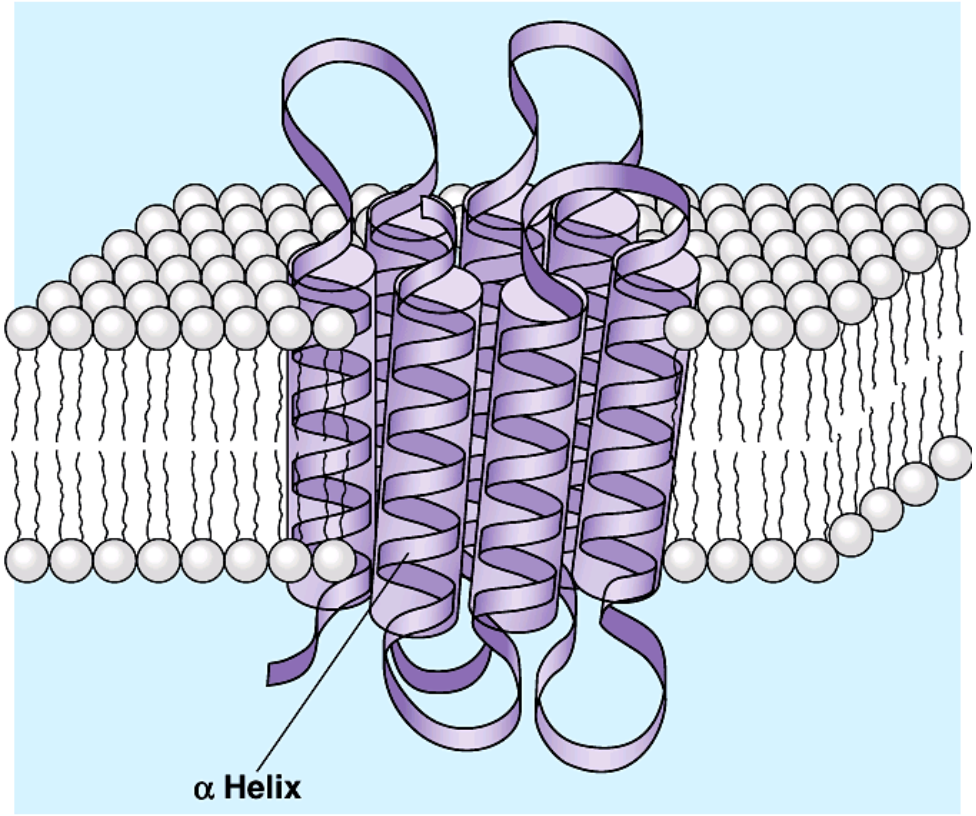
- AACompIdent - Identify a protein by its amino acid composition
- AACompSim - Compare the amino acid composition of a UniProtKB/Swiss-Prot entry with all other entries
- TagIdent - Identify proteins with isoelectric point (*pI*), molecular weight (*Mw*) and sequence tag, or generate a list of proteins close to a given *pI* and *Mw*
- MultiIdent - Identify proteins with isoelectric point (*pI*), molecular weight (*Mw*), amino acid composition, sequence tag and peptide mass fingerprinting data

Other prediction or characterization tools

- ProtParam - Physico-chemical parameters of a protein sequence (amino-acid and atomic compositions, isoelectric point, extinction coefficient, etc.)
- Compute pI/Mw - Compute the theoretical isoelectric point (*pI*) and molecular weight (*Mw*) from a UniProt Knowledgebase entry or for a user sequence
- GlycanMass - Calculate the mass of an oligosaccharide structure
- PeptideCutter - Predicts potential protease and cleavage sites and sites cleaved by chemicals in a given protein sequence
- PeptideMass - Calculate masses of peptides and their post-translational modifications for a UniProtKB/Swiss-Prot or UniProtKB/TrEMBL entry or for a user sequence
- IsotopIdent - Predicts the theoretical isotopic distribution of a peptide, protein, polynucleotide or chemical compound

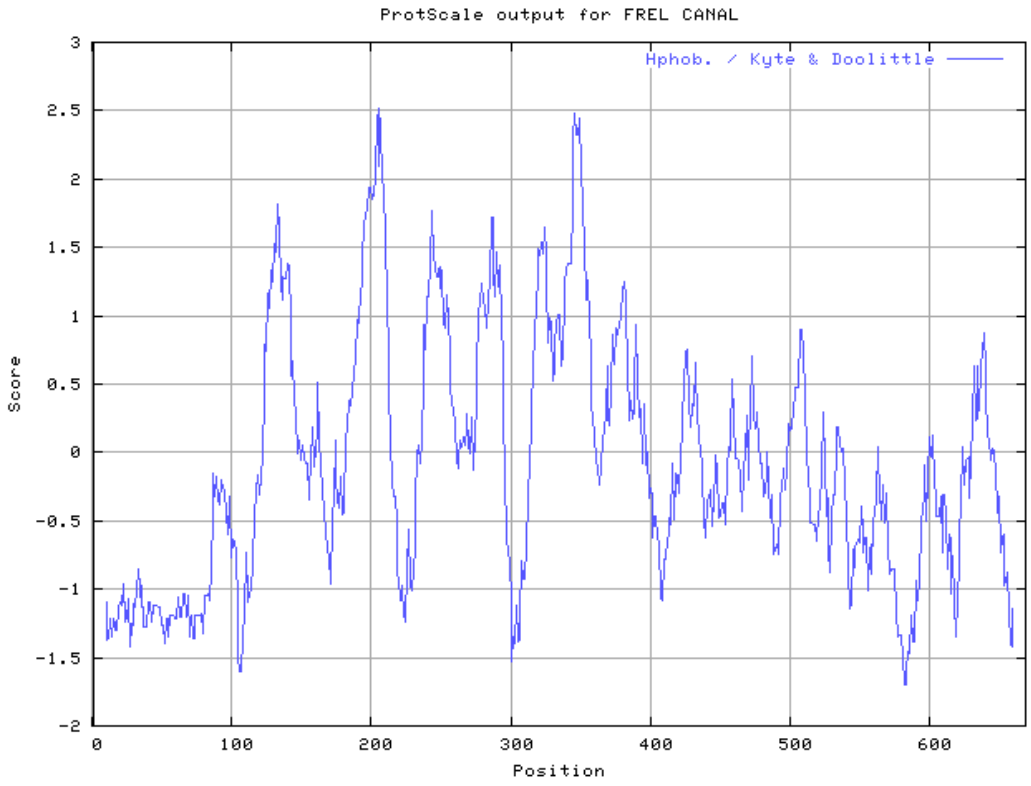
Predikce membránových regionů

- Hydrofóbní segmenty v membránových proteinech



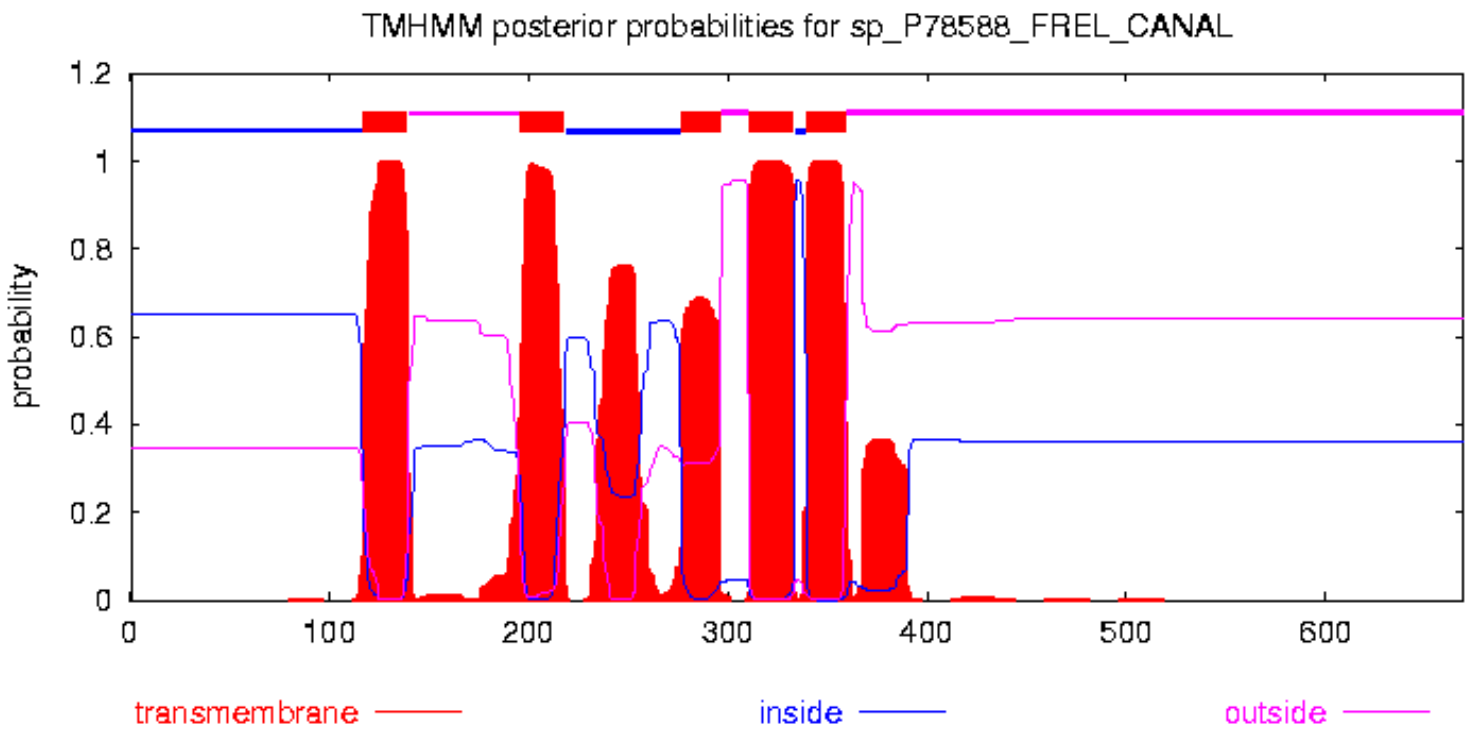
Predikce membránových regionů

- ProtScale
  - Predikce hydrofobicitního profilu ze sekvence



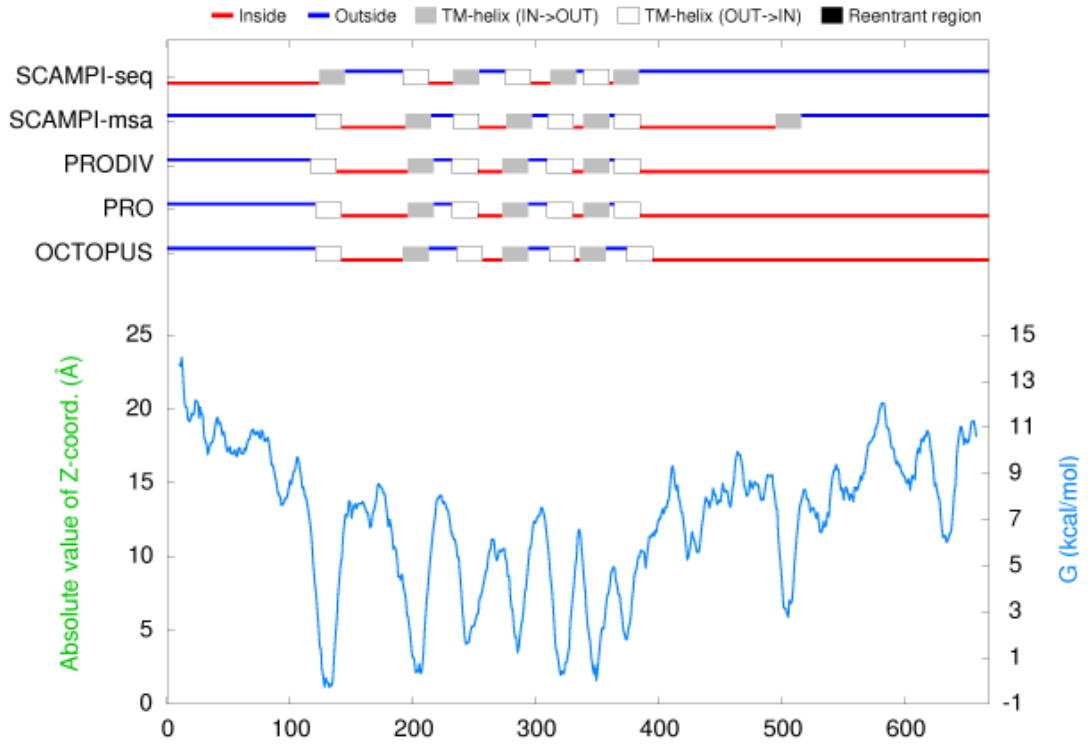
Predikce membránových regionů

- TMHMM
  - Predikce pravděpodobnostní metodou Skrytých Markovových Modelů



Predikce membránových regionů

- TOPCONS
  - Konsenzuální predikce topologie membránových proteinů



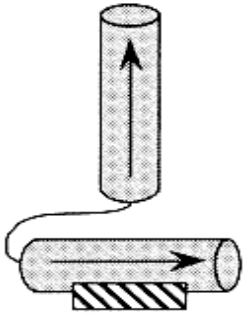
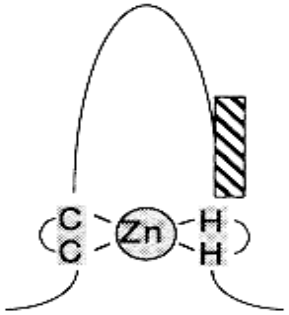
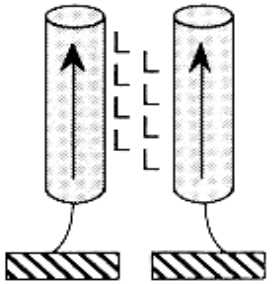
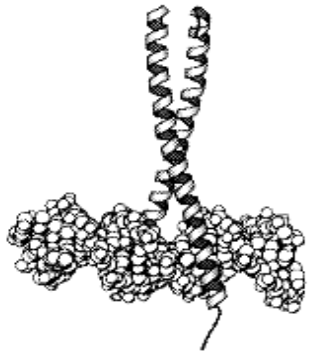
Predikce motivů a domén

- Konzervované vzorce sekvencí jsou spojené s konkrétní proteinovou rodinou, biologickými vlastnostmi nebo funkcí

Predikce motivů a domén

- Konzervované vzorce sekvencí jsou spojené s konkrétní proteinovou rodinou, biologickými vlastnostmi nebo funkcí
  - Klasifikace proteinových sekvencí
  - Identifikace strukturních a evolučních vztahů
  - Funkční anotace nových proteinů
  - Identifikace vazebných míst pro ligandy
  - Predikce postranslačních modifikací
  - Predikce sub-celulární lokalizace

Predikce motivů a domén

Name	Helix-loop-helix (Myc type)	Cys-His zinc finger	Leucine zipper
Sequence	[DENSTAP]-K-[LIVMWAGN]-{FYWCPHKR}-[LIVT]-[LIV]-x(2)-[STAV]-[LIVMSTAC]-x-[VMFYH]-[LIVMTA]-{P}-{P}-[LIVMSR]	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H	L-x(6)-L-x(6)-L-x(6)-L
Structure			
Function	DNA Binding	DNA Binding	DNA Binding
Example			
	3CRO	2DRP	1YSA

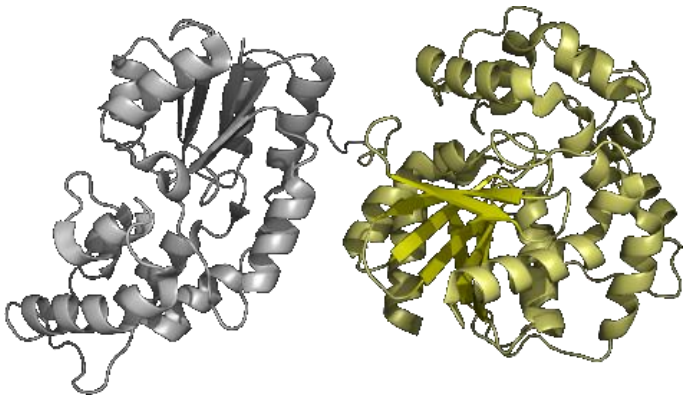
Predikce motivů a domén

- Konzervované vzorce sekvencí jsou spojené s konkrétní proteinovou rodinou, biologickými vlastnostmi nebo funkcí
- Motivy
  - Zpravidla krátké – 10-20 aminokyselinových zbytků

    CGDAEEGDACCDGA

Predikce motivů a domén

- Konzervované vzorce sekvencí jsou spojené s konkrétní proteinovou rodinou, biologickými vlastnostmi nebo funkcí
- Motivy
- Domény
  - Delší než motivy – 40-700 aminokyselinových zbytků
  - Nezávislé strukturní a funkční jednotky

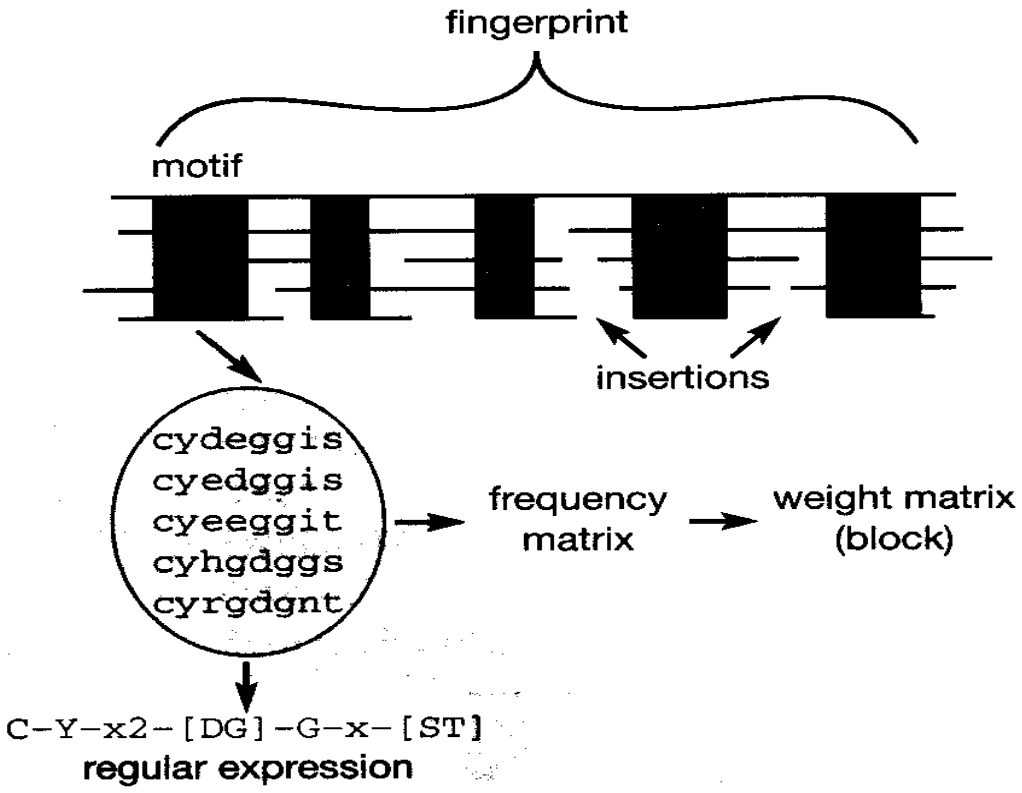


Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence

Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence



Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence

	1	2	3	4	5	6	7	8	9	10
I	Y	D	G	G	A	V	–	E	A	L
II	Y	D	G	G	–	–	–	E	A	L
III	F	E	G	G	I	L	V	E	A	L
IV	F	D	–	G	I	L	V	Q	A	V
V	Y	E	G	G	A	V	V	Q	A	L
	y	d	G	G	A/I	V/L	V	e	A	l

Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence
  - Regulární výrazy

Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence
  - Regulární výrazy

    E-X(2)-[FHM]-X(4)-{P}-L

Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence
  - Regulární výrazy

    E-X(2)-[FHM]-X(4)-{P}-L

    zbytek E je následován 2 libovolnými zbytky

Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence
  - Regulární výrazy

    E-X(2)-[FHM]-X(4)-{P}-L

    zbytek E je následován 2 libovolnými zbytky

    následovanými F nebo H nebo M zbytky

Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence
  - Regulární výrazy

    E-X(2)-[FHM]-X(4)-{P}-L

    zbytek E je následován 2 libovolnými zbytky

    následovanými F nebo H nebo M zbytky

    následovanými 4 libovolnými zbytky

Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence
  - Regulární výrazy

    E-X(2)-[FHM]-X(4)-{P}-L

    zbytek E je následován 2 libovolnými zbytky

    následovanými F nebo H nebo M zbytky

    následovanými 4 libovolnými zbytky

    následovanými jakýmkoliv zbytkem kromě P

Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence
  - Regulární výrazy

    E-X(2)-[FHM]-X(4)-{P}-L

    zbytek E je následován 2 libovolnými zbytky

    následovanými F nebo H nebo M zbytky

    následovanými 4 libovolnými zbytky

    následovanými jakýmkoliv zbytkem kromě P

    následovaným zbytkem L

Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence
  - Regulární výrazy — Počet přesných hitů

D-A-V-I-D	71
D-A-V-I-[DENQ]	252
[DENQ]-A-V-I-[DENQ]	925
[DENQ]-A-[VLI]-I-[DENQ]	2739
[DENQ]-[AQ]-[VLI]2-[DENQ]	51506

Predikce motivů a domén

- Vytvářeny z multinásobného přiložení příbuzných sekvencí
- Uloženy v databázích ve formě konsenzuální sekvence
  - Regulární výrazy
  - Statistické modely (profily, bloky, Skryté Markovovy Modely)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
2	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
3	P	-1	-2	-2	-2	-3	-2	-1	-2	-2	-3	-3	-1	-3	-4	8	-1	-1	-4	-3	-3
4	S	1	-1	0	-1	-1	0	0	-1	-1	-3	-3	0	-2	-3	-1	5	1	-3	-2	-2
5	C	-1	-4	-3	-4	9	-3	-4	-3	-3	-2	-2	-3	-2	-3	-3	-1	-1	-3	-3	-1
6	T	0	-1	0	-1	-1	-1	-1	-1	-2	-2	-3	-1	-2	-3	-1	4	3	-3	-2	-2
7	Y	-2	-3	-3	-4	-3	-2	-3	-4	1	-1	-1	-3	-1	5	-4	-2	-2	1	7	-2
8	Y	-1	-1	-1	-1	-2	0	-1	-2	6	-2	-1	-1	-1	1	-1	-1	-1	0	5	-2
9	V	-1	-2	-2	-2	-1	-2	-2	-2	-2	1	2	-2	0	-1	-2	-2	-1	-2	-1	4
10	S	-1	-1	-1	-1	-3	3	3	-2	-1	-2	1	0	-1	-2	-2	2	-1	-3	-2	-2

Databáze motivů a domén

- Manuální
  - Informativní díky kvalitním anotacím
  - Nízký počet položek
- Automatické
  - Méně informativní
  - Vysoký počet položek

Databáze motivů a domén

The Main Domain Collections			
Name	***Web Address***	***Number of Domains***	***Generation***
PROSITE-Profile (IP)	www.expasy.org/prosite	616	Manual
PfamA (IP)	www.sanger.ac.uk/Software/Pfam	7973	Manual
PRINTs (IP)	www.bioinf.man.ac.uk/dbbrosers/PRINTS	1900	Manual
PRODOM (IP)	protein.toulouse.inra.fr/prodom/current/html/home.php	736000	Automatic
SMART (IP)	smart.embl-heidelberg.de	685	Manual
COGs	www.ncbi.nlm.nih.gov/COG/new/	4852	Manual
TIGRFAM (IP)	www.tigr.org/TIGRFAMs	2453	Manual
BLOCKs	blocks.fhcrc.org/	12542	Automatic

Databáze motivů a domén

- PROSITE
  - Motivy navrženy manuálně kvalifikovanými odborníky
  - Motivy často krátké pro zvýšení specifiy
  - Shody nutno interpretovat opatrně!

Databáze motivů a domén

- PROSITE
  - Hity *versus* hity s vysokým výskytem
  - Rozpoznání hitů = délka vzorce, informace o organismu, identifikace podobných vzorců, konzervovanost vzorce v přiložení

hits by patterns: [6 hits (by 3 distinct patterns) on 1 sequence]

P12259
(FA5_HUMAN)

RecName: Full=Coagulation factor V; AltName: Full=Activated protein C cofactor; AltName: Full=Proaccelerin, labile factor; Contains: RecName: Full=Coagulation factor V heavy chain; Contains: RecName: Full=Coagulation factor V light chain; Flags: Precursor;. *Homo sapiens (Human)*

PS00079 **MULTICOPPER_OXIDASE1** *Multicopper oxidases signature 1* : Hits on PDB 3D structures: [1FV4-H, 1FV4-L, 1Y61-H, 1Y61-L]

304 - 324: GkWiIsSlTPkhLqAGMqayI

1880 - 1900: GwWlLnTeVGenQrAGMqtpF

Databáze motivů a domén

- BLOCKs
  - Bloky = segmenty multinásobného přiložení bez mezer

    korespondující s nejkonzervovanějšími regiony v proteinech

Family		Strand	Blocks	Combined E-value
IPB000639	Epoxide hydrolase signature	1	5 of 6	7.7e-22
IPB003089	Alpha/beta hydrolase fold signature	1	3 of 4	4.5e-07
IPB000073	Alpha/beta hydrolase fold	1	2 of 2	0.0041
IPB002410	Prolyl aminopeptidase (S33) family	1	1 of 3	0.032
IPB002828	Survival protein SurE	1	1 of 7	0.5

Databáze motivů a domén

- BLOCKs
  - Bloky = segmenty multinásobného přiložení bez mezer

    korespondující s nejkonzervovanějšími regiony v proteinech

Family		Strand	Blocks	Combined E-value
IPB000639	Epoxide hydrolase signature	1	5 of 6	7.7e-22
IPB003089	Alpha/beta hydrolase fold signature	1	3 of 4	4.5e-07
IPB000073	Alpha/beta hydrolase fold	1	2 of 2	0.0041
IPB002410	Prolyl aminopeptidase (S33) family	1	1 of 3	0.032
IPB002828	Survival protein SurE	1	1 of 7	0.5

==========

>IPB000639 5/6 blocks Combined E-value= 7.7e-22: Epoxide hydrolase signature

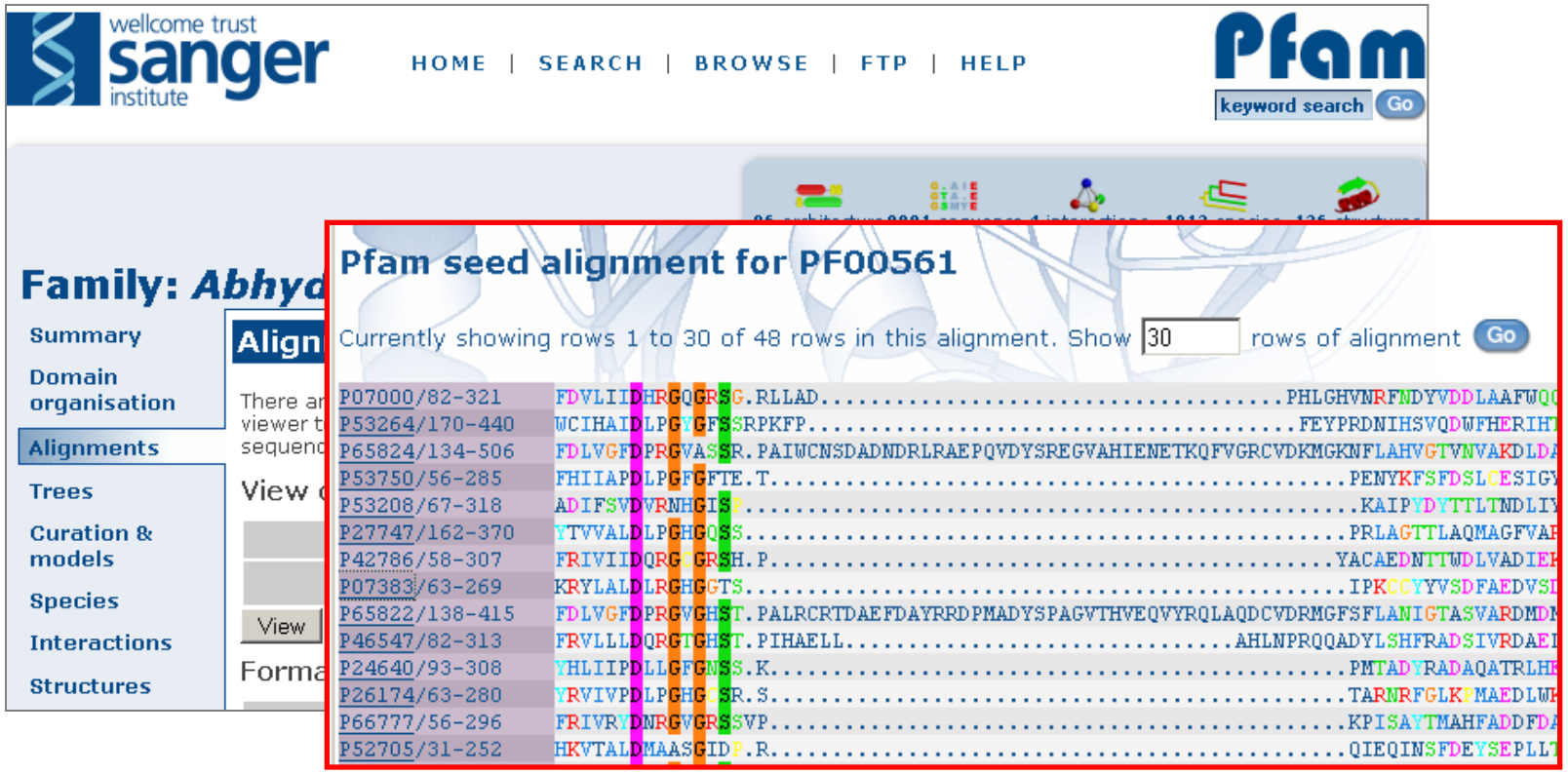
Block	Frame	Location (aa)	Block E-value
IPB000639A	0	36-54	0.0002
IPB000639B	0	55-70	0.012
IPB000639C	0	104-117	0.032
IPB000639D	0	118-131	2.2
IPB000639F	0	267-289	0.0024

Databáze motivů a domén

- Pfam
  - Přiložení domén vytvořené ze sekvencí databáze UniProtKB
  - Každá doména je reprezentována profilem Skrytých Markovových Modelů vytvořeným z mnohonásobného přiložení
  - Obsahuje dvě části: Pfam-A z manuálního přiložení a Pfam-B z automatického přiložení

Databáze motivů a domén

- Pfam



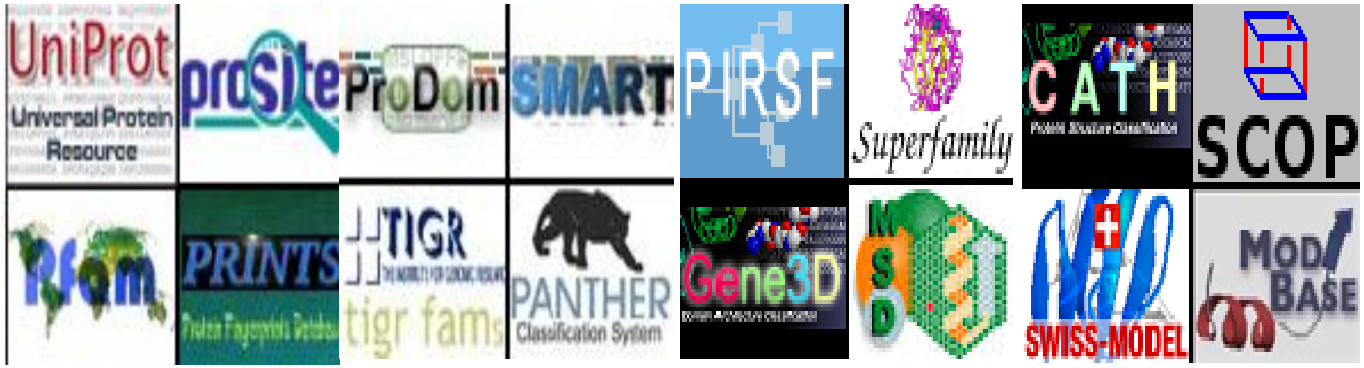
Databáze motivů a domén

- ProDom
  - Databáze proteinových domén automaticky vytvořenými ze sekvencí databáze UniProtKB
  - Navržena jako vyčerpávající sbírka domén i bez znalosti funkce

Databáze motivů a domén

- InterPro
  - Řeší problém redundance jednotlivých databází
  - Zahrnuje téměř všechny dostupné sekundární databáze:

    PROSITE, Pfam, PRINTS, ProDom, SMART,…



Prohledávání databází motivů a domén

- Simultánní prohledání několika databází
  - InterProScan
  - CD Server
  - Motif-Scan

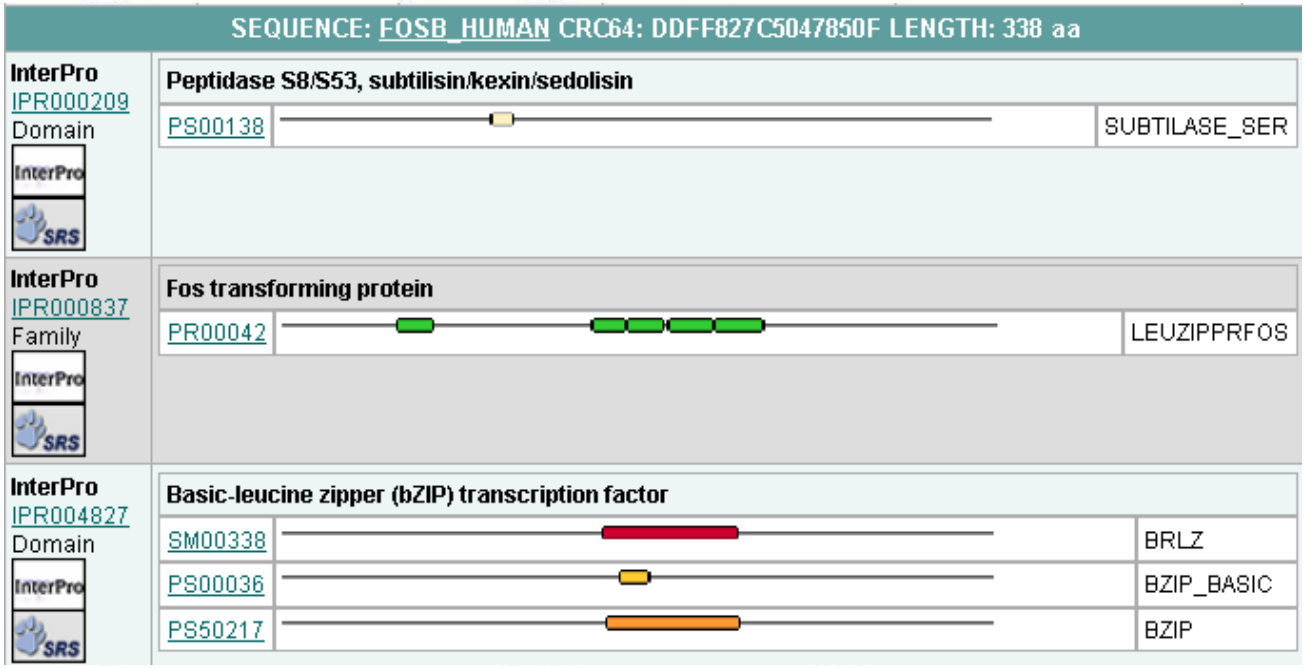
Prohledávání databází motivů a domén

- Simultánní prohledání několika databází
  - InterProScan
  - CD Server
  - Motif-Scan

- Vysoké skóre a vysoká shoda = spolehlivá interpretace
  - Závěry téměř vždy správné
- Nízké skóre nebo částečná shoda = problematická interpretace
  - Závěry vyžadují další podpůrná data

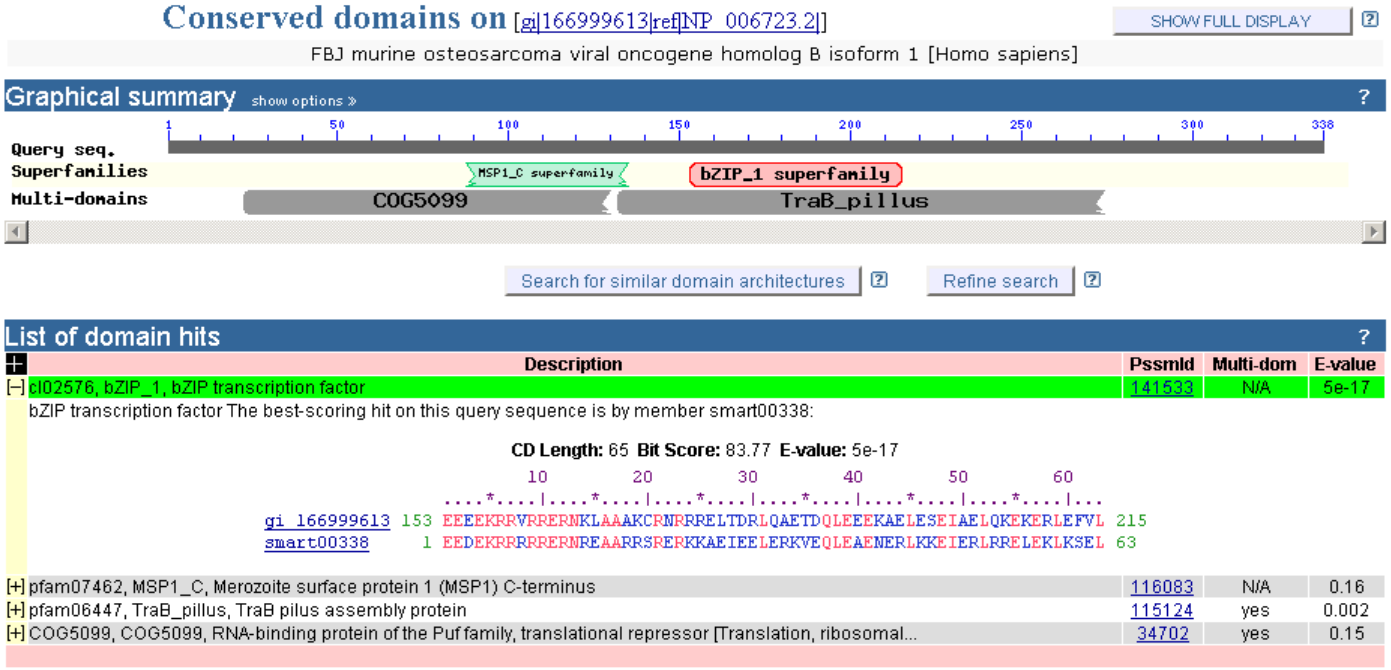
Prohledávání databází motivů a domén

- InterProScan
  - Srovnání prohledávané sekvence s InterPro databází
  - Hity a jejich umístění na sekvenci jsou vypsány přehledně



Prohledávání databází motivů a domén

- CD Server
  - Hity jsou vypsány s E-hodnotou
  - prohledává menší počet databází než InterProScan



Prohledávání databází motivů a domén

- Motif-Scan
  - Hity jsou vypsány s E-hodnotou a normalizovaným skóre
  - Relevantní hity jsou označeny “!”

Match details

match detail	match score	motif information
	Status: **!** pos.: **155-218** raw-score = **989** N-score = **13.185** E-value = **1.4e-06**	prf:BZIP *Basic-leucine zipper (bZIP) domain profile.* [entry] [graphics]

Reference

- Claverie, J-M., & Notredame, C. (2006). **Bioinformatics for Dummies** (2nd ed.). Wiley Publishing, Hoboken, p. 436.
- Xiong, J. (2006). **Essential Bioinformatics**, Cambridge University Press, New York, p. 352.

- **ExPASy**: http://www.expasy.ch/
- **ProtScale**: http://www.expasy.org/cgi-bin/protscale.pl
- **TMHMM**: http://www.cbs.dtu.dk/services/TMHMM-2.0/
- **TOPCONS**: http://topcons.net/
- **PROSITE**: http://www.expasy.org/prosite/
- **BLOCKs**: http://blocks.fhcrc.org
- **Pfam:** http://pfam.sanger.ac.uk/
- **ProDom**: http://prodom.prabi.fr/prodom/current/html/home.php
- **InterPro**: http://www.ebi.ac.uk/interpro/
- **InterProScan:** http://www.ebi.ac.uk/Tools/InterProScan/
- **CD Search:** http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
- **Motif-Scan**: http://myhits.isb-sib.ch/cgi-bin/motif_scan