

LOSCHMIDT
LABORATORIES



Analýza proteinových sekvencí



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

- ❑ 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

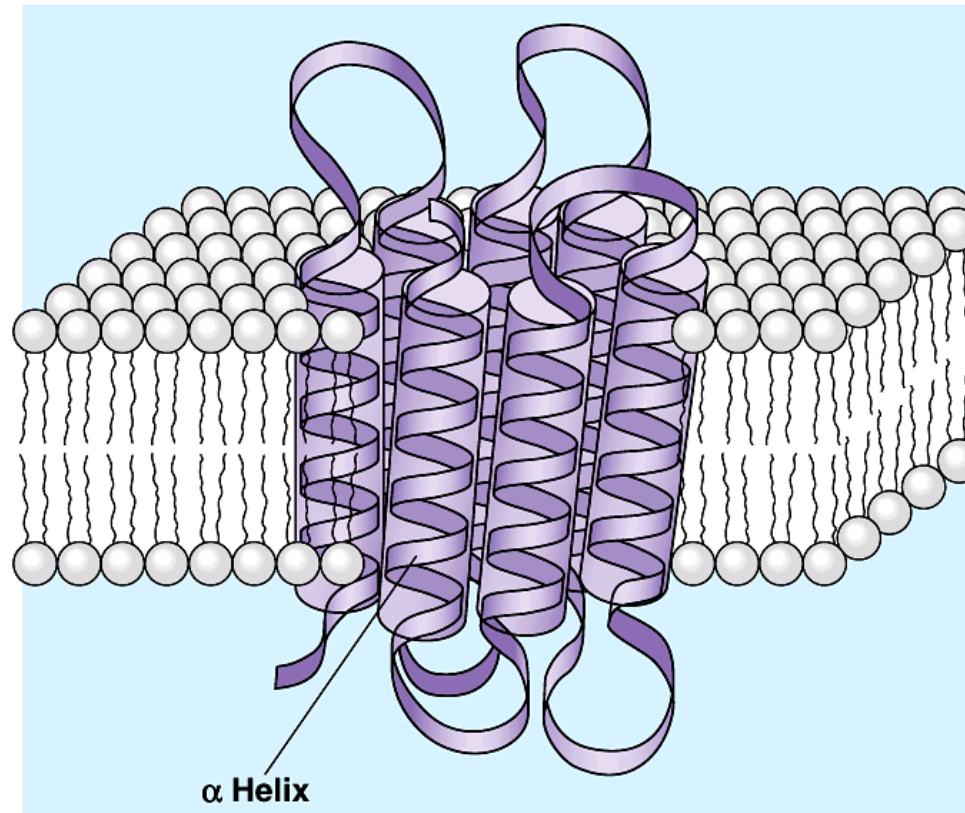
- [AACompldent](#)  - Identify a protein by its amino acid composition
- [AACompSim](#)  - Compare the amino acid composition of a UniProtKB/Swiss-Prot entry with all other entries
- [TagIdent](#)  - Identify proteins with isoelectric point (pI), molecular weight (Mw) and sequence tag, or generate a list of proteins close to a given pI and Mw
- [Multident](#)  - 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

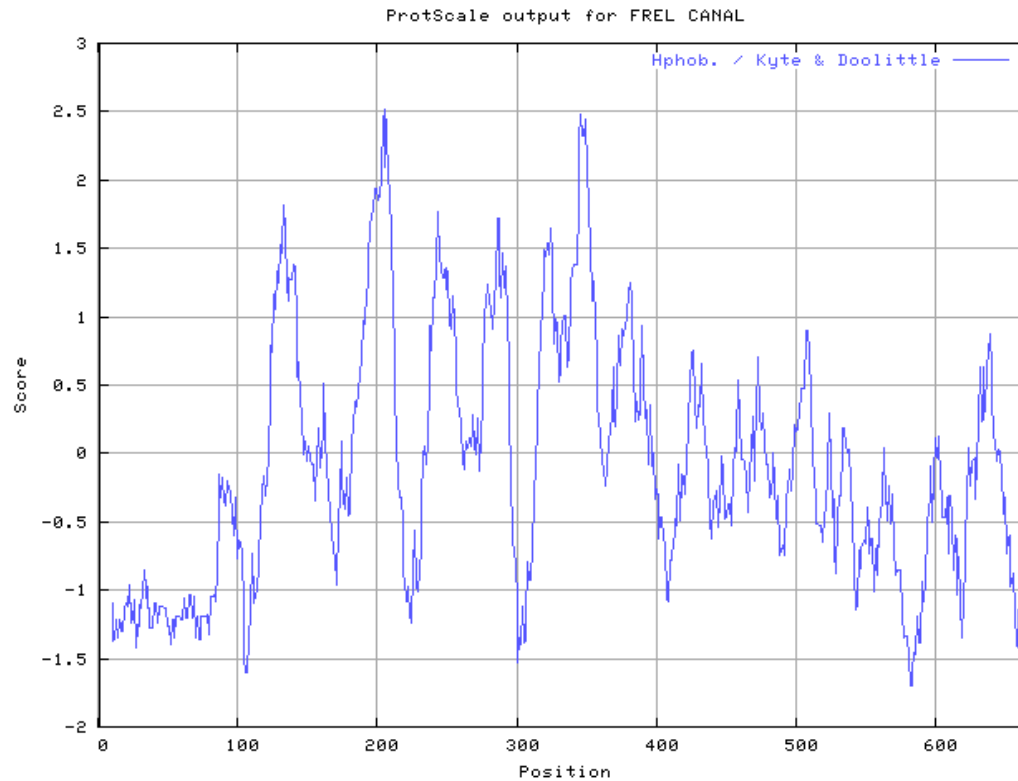
- Hydrofóbní segmenty v membránových proteinech



Predikce membránových regionů

□ ProtScale

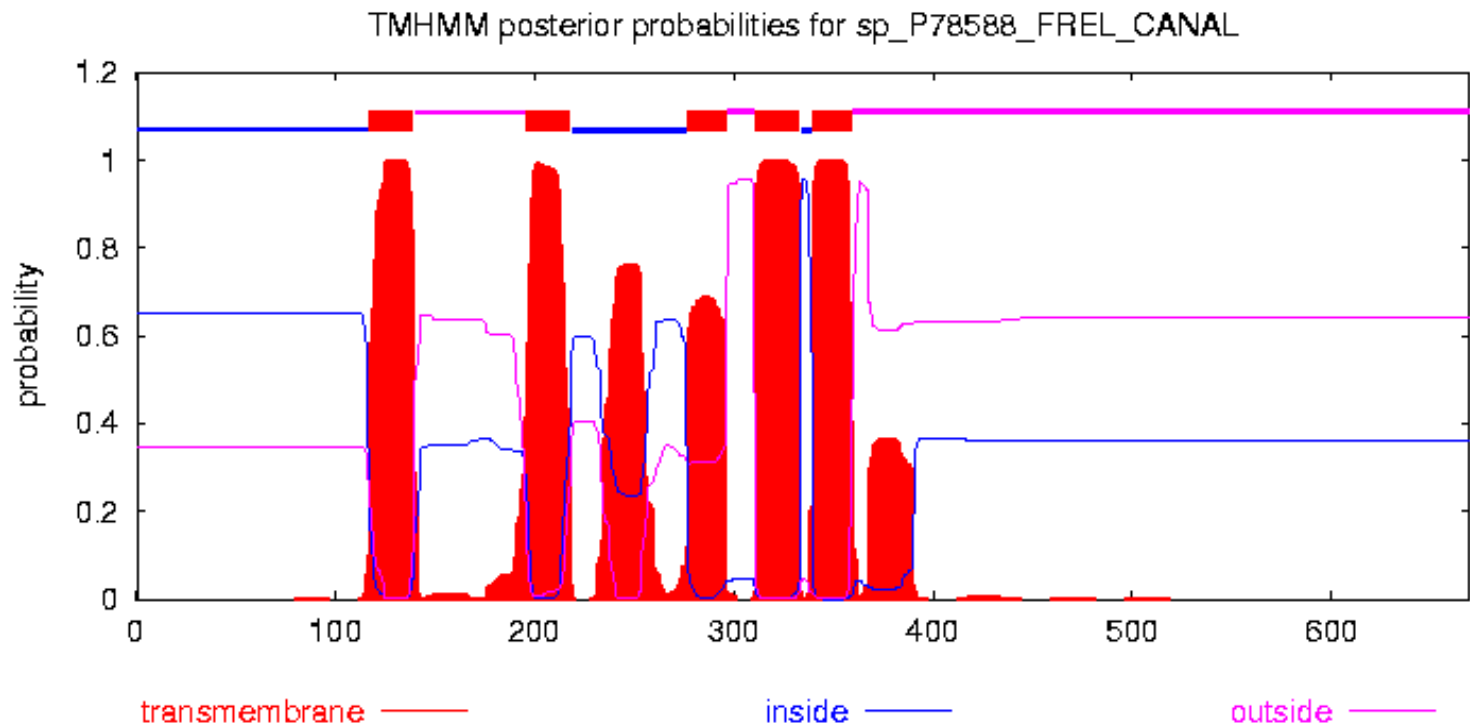
- Predikce hydrofobicitního profilu ze sekvence



Predikce membránových regionů

□ TMHMM

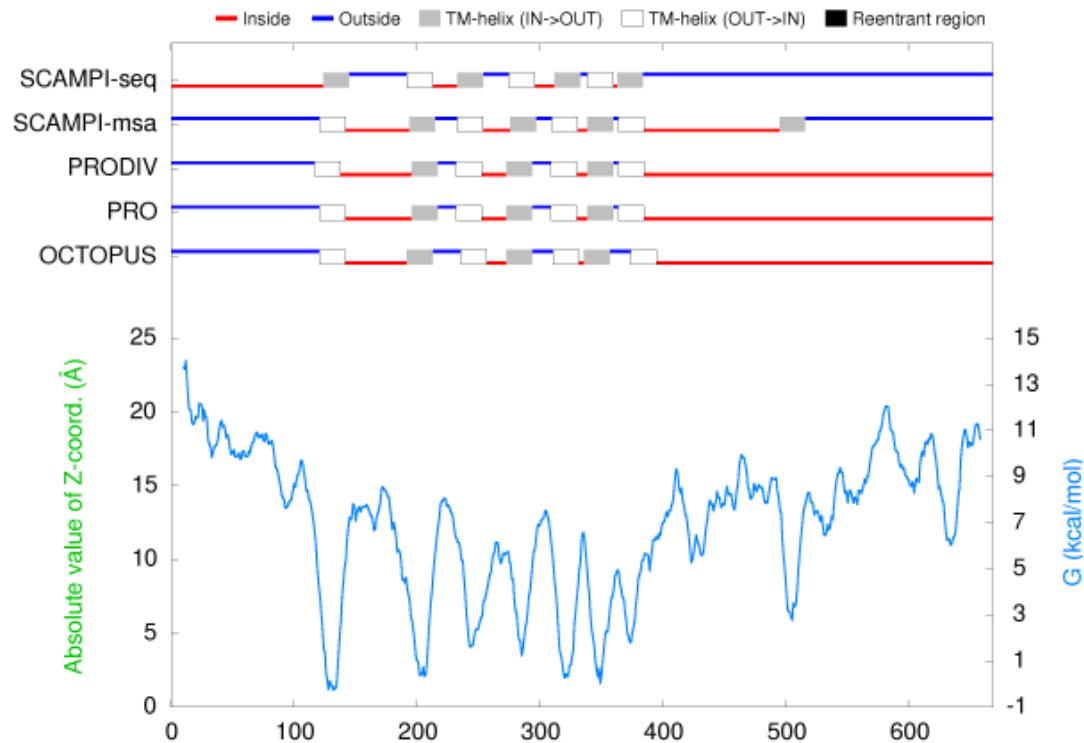
- Predikce pravděpodobnostní metodou Skrytých Markovových Modelů



Predikce membránových regionů

□ TOPCONS

- Konsenzuální predikce topologie membránových proteinů



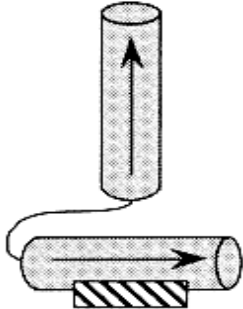
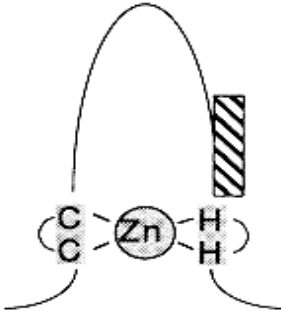
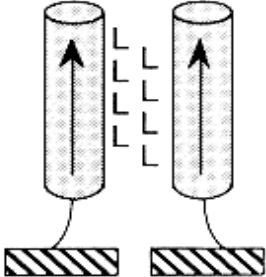


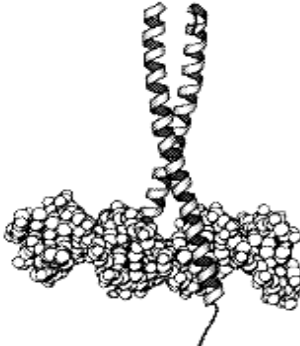
Predikce motivů a domén

- 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	$[DENSTAP]-K-[LIVMWAGN]-$ $\{FYWC\}PHKR-[LIVT]-[LIV]-x(2)-$ $[STAV]-[LIVMSTAC]-x-[VMFYH]-$ $[LIVMTA]-\{P\}-\{P\}-[LIVMSR]$	$C-x(2,4)-C-x(3)-[LIVMFYWC]-$ $x(8)-H-x(3,5)-H$	$L-x(6)-L-x(6)-L-x(6)-L$
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

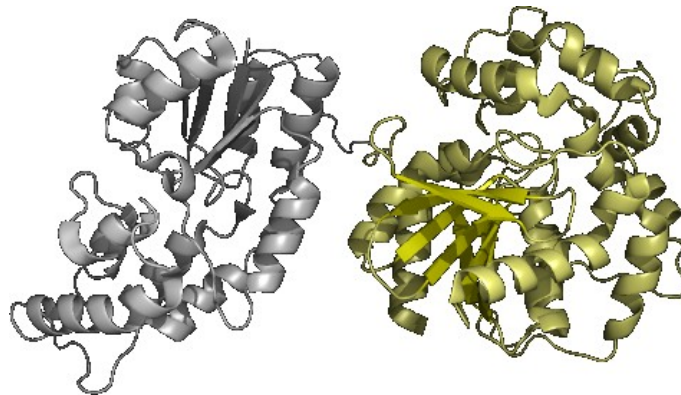
□ Motiv

- 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

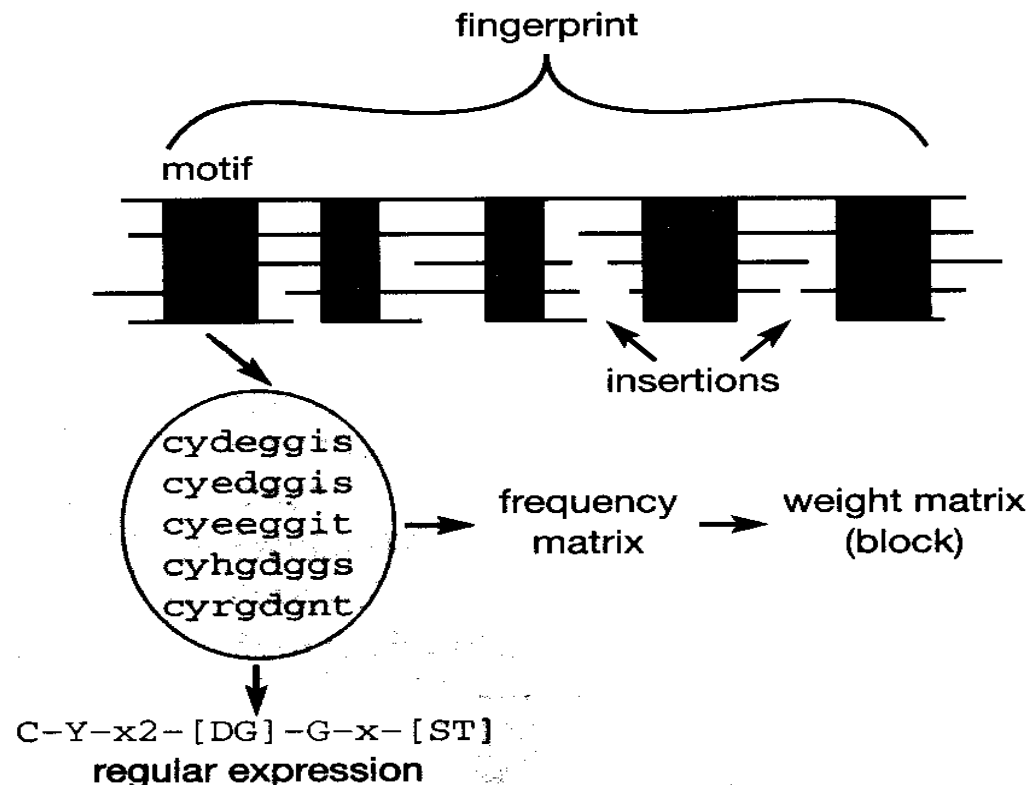


Predikce motivů a domén

- 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



- 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/dbbrowsers/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

□ 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 příložen

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: GkWiIsS1TPkhLqAGMqayI

1880 - 1900: GwWlLnTeVGenQrAGMqtpF

□ 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
```


□ Pfam

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

Databáze motivů a domén

□ Pfam

wellcome trust
sanger
institute

HOME | SEARCH | BROWSE | FTP | HELP

Pfam
keyword search

Family: **Abhydrolase superfamily**

Summary
Domain organisation
Alignments
Trees
Curation & models
Species
Interactions
Structures

Alignments
Currently showing rows 1 to 30 of 48 rows in this alignment. Show rows of alignment

P07000/82-321	FDVLIIDHRGQGRSG.RLLAD.....	PHLGHVNRFN DYVDDLAAFWQQ
P53264/170-440	WCIHAIDLPGYGFSSRPKFP.....	FEYPRDNIHSVQDWFHERIHF
P65824/134-506	FDLVGFDPFGVASSR.PAIWCMSDADMDRLRAEPQVDYSREGVAHIEMETKQFVGRVCVDMKMGKFLAHVGTVNVAKDLDA	
P53750/56-285	FHIIAPDLLPFGFTE.T.....	PENYKFSFD SLCE SIGY
P53208/67-318	ADIFSV D VRNHGIS P.....	KAIPYDYTTLTNDLIY
P27747/162-370	YTVVALDLPCHGQSS.....	PRLAGTTLAQMGFVAF
P42786/58-307	FRIVIIDQRGGRSH.P.....	YACAE DNTTDLVADIEP
P07383/63-269	KRYLALDLRGHGGS.....	IPKCCYVVSDF AEDVSI
P65822/138-415	FDLVGFDPFGVGHST.PALRCRTDAEFDAYRRDPMADYSPAGVTHVEQVYRQLAQDCVDRMGFSFLANIGTASVARDMDM	
P46547/82-313	FRVLLDQRCGTGHST.PIHAELL.....	AHLNPRQQADYLSHFRA DSIVRDAEI
P24640/93-308	YHLIIPDLLGFGNS.K.....	PMTADYRADAQATRLHF
P26174/63-280	YRVIVPDLPGHGQSR.S.....	TARNRFGLKPM AEDLWF
P66777/56-296	FRIVRYDNRGVGRSSVP.....	KPISAYTMAHFADDFDA
P52705/31-252	HKVTALDMAASGIDP.R.....	QIEQINSFDEYSEPLL

□ ProDom

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

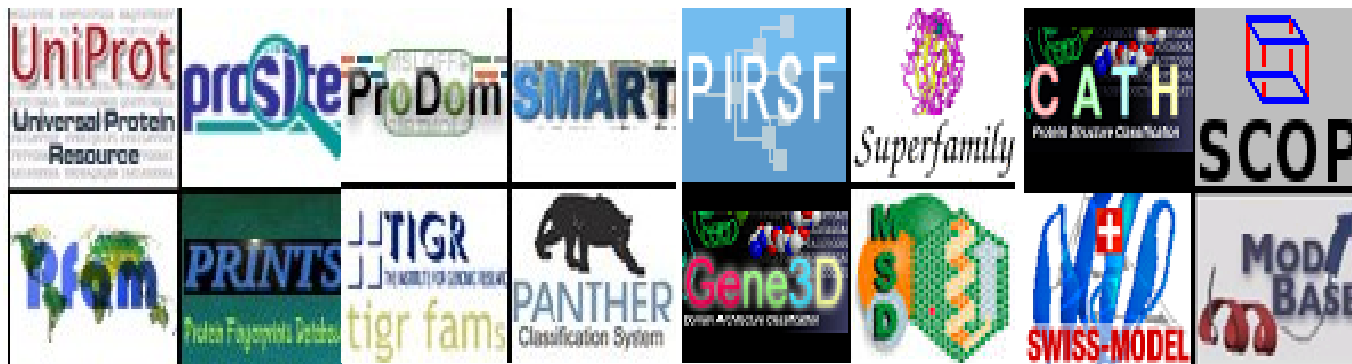


Databáze motivů a domén



□ 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

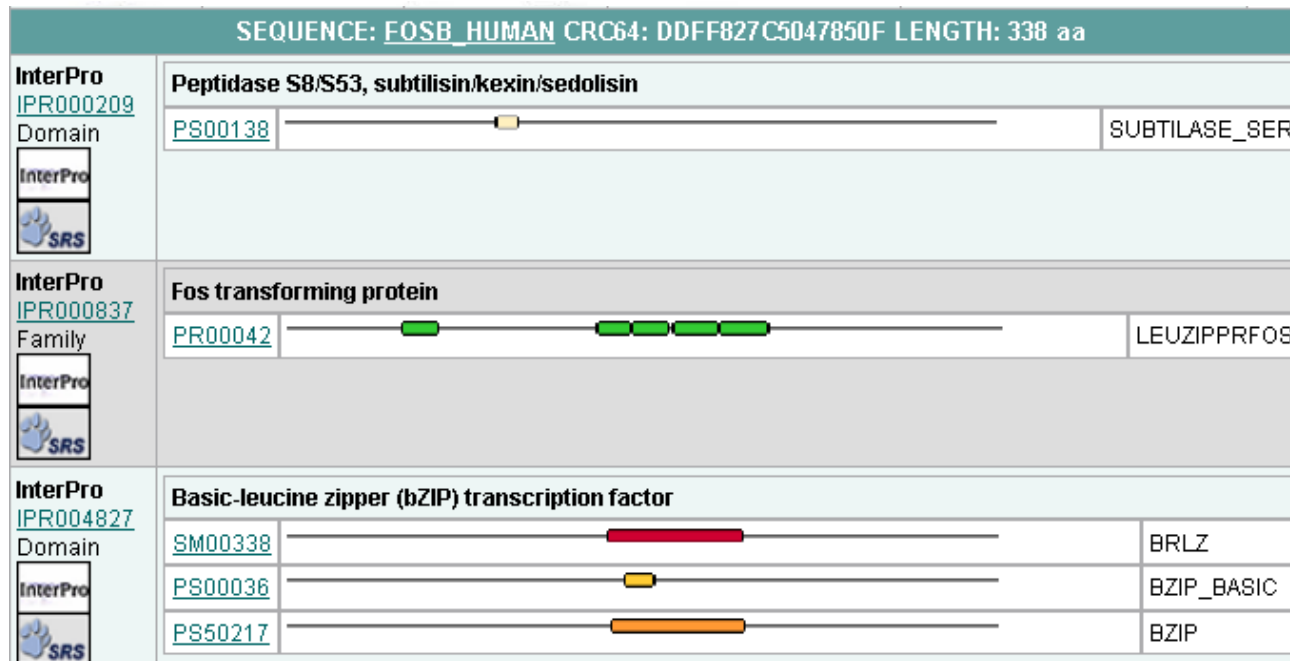
- Simultánní prohledání několika databází
 - InterProScan
 - CD Server
 - Motif-Scan
- Vysoké skóre a vysoká shoda = spolehlivá interpretace
 - Závěry téměř **vždy správné**
- Nízké skóre nebo částečná shoda = problematická interpretace
 - Závěry vyžadují další **podpůrná data**



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

Conserved domains on [\[gi166999613|ref|NP_006723.2\]](#) [SHOW FULL DISPLAY](#) [?](#)

FBJ murine osteosarcoma viral oncogene homolog B isoform 1 [Homo sapiens]

Graphical summary [show options >](#) [?](#)

Query seq. 1 50 100 150 200 250 300 338

Superfamilies MSP1_C superfamily bZIP_1 superfamily

Multi-domains COG5099 TraB_pillus

[Search for similar domain architectures](#) [?](#) [Refine search](#) [?](#)

List of domain hits [?](#)

Description	Pssmid	Multi-dom	E-value
gi102576 , bZIP_1 , bZIP transcription factor	141533	N/A	5e-17

bZIP transcription factor The best-scoring hit on this query sequence is by member [smart00338](#):

CD Length: 65 Bit Score: 83.77 E-value: 5e-17


```
      10      20      30      40      50      60
      .....*.....*.....*.....*.....*.....*.....*.....
gi_166999613 153 EEEKRRVRRERNKLA AAKCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVL 215
smart00338    1  EEDEKRRRRRRERNREAARRSRERKKAETEELEKRVQLEAEMERLKKETIERLRRELEKIKSEL 63
```

[+] pfam07462 , MSP1_C , Merozoite surface protein 1 (MSP1) C-terminus	116083	N/A	0.16
[+] pfam06447 , TraB_pillus , TraB pilus assembly protein	115124	yes	0.002
[+] COG5099 , COG5099 , RNA-binding protein of the Puf family, translational repressor [Translation, ribosomal...	34702	yes	0.15

Prohledávání databází motivů a domén

□ Motif-Scan

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

Match details		
<u>match detail</u>	<u>match score</u>	<u>motif information</u>
 <p>EEKRRVRRERENKLA AAKCRNRRREL TDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAH</p>	Status: ! pos.: 155-218 raw-score = 989 N-score = 13.185 E-value = 1.4e-06	prf:BZIP <i>Basic-leucine zipper (bZIP) domain profile.</i> [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