

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

# Předpověď 3D-struktury/ foldu

- Klasifikace proteinů
- Předpověď funkce
- Vytvoření modelu pro další studium

# Metody pro predikci funkce

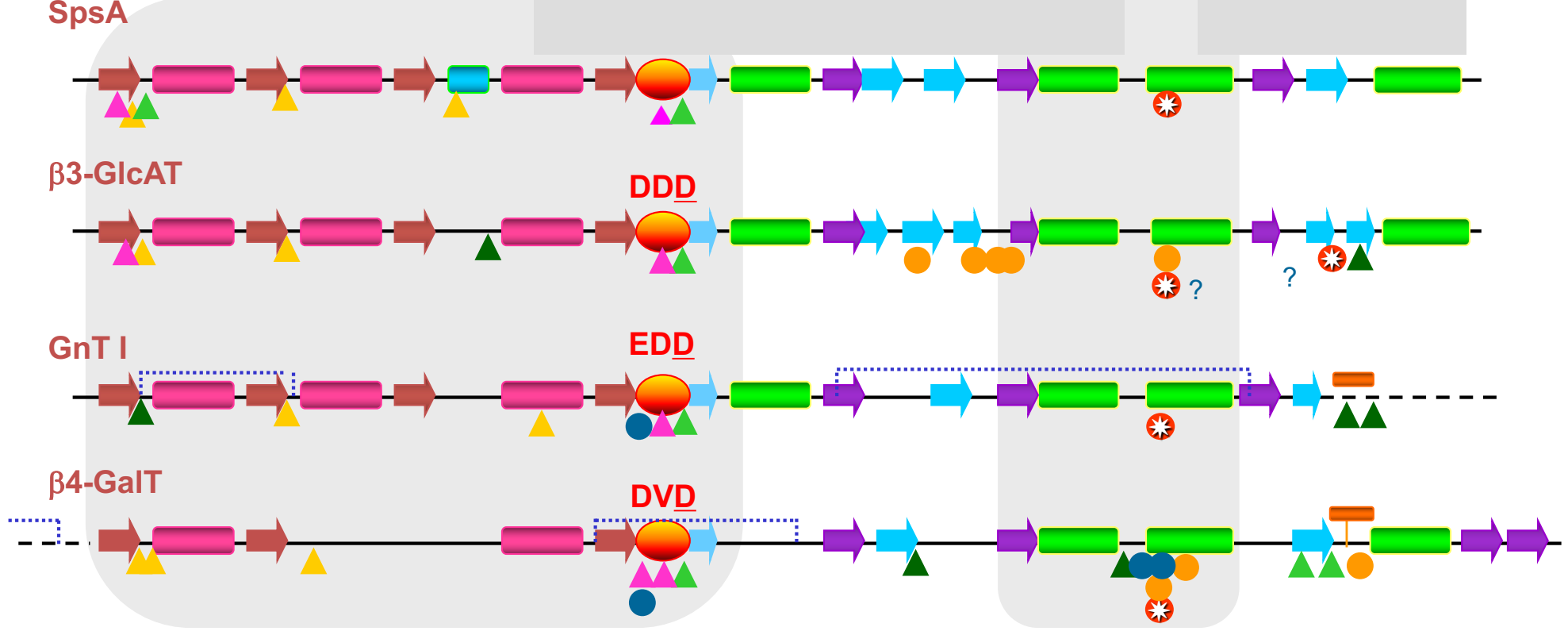
„klasické“ metody: vícenásobné aminokyselinové přiložení  
 pozitivní alignment pouze mezi sekvencemi stejné rodiny

Analýza 2D struktury  
 identifikuje některé  
 «Rossmann»  
 (10)

Gal $\alpha$ 1,4-Gal $\beta$ -R  
 Gal $\alpha$ 1,3-Glc $\alpha$ -R  
 Gal $\alpha$ 1,3-Glc $\alpha$ -R  
 Glc $\alpha$ 1,2-Glc $\alpha$ -R  
 Gal $\alpha$ 1,6-Man $\alpha$ -R  
 Glc $\alpha$ 1,3-Man $\alpha$ -R  
 SpsA

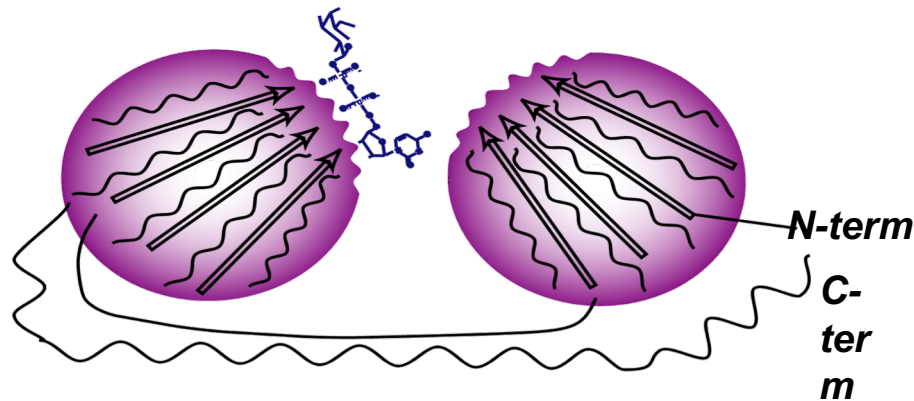
LgtC *N. men*  
 RfaI *E. coli*  
 RfaI *S. typh*  
 RfaJ *E. coli*  
 LpcA *R. leg*  
 DUGT *D. mel*

CDKVLVYLDIDVLVRDSLTP	LWDTDLGDNWLGACID	...	YFNAGVLLINLKKWR
APKVLVYLDADIICQGTIEPLINFSFPDDKVAMVVT	...	...	YFNSGFLINTAQWA
QIKVLVYLDADIACKGSIQELIDLNF	AEENEIAAVVA	...	YFNAGFILIXIPLWT
LDRLLYLDADVCKGDISQLLHLGLN-GAVAAVVK	...	...	YFNSGVVYLDLKKWA
IERLLYLDADVLAVSPVDELFTTRNFQKGALAAVDD	...	...	YFNAGVLLFDWSACR
VRKIIFVDADAIVRTDIKELYDMDLGGAPYAYTPF	...	...	YHISALYVVDLKRFR



# Dvě pozorované topologie 3D struktur glykosyltransferas

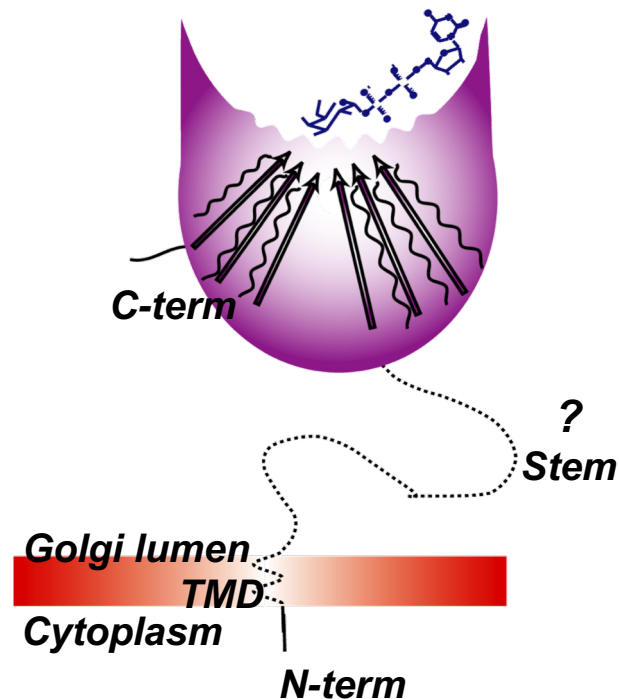
## BGT-fold



### (Prokaryotes/Phage)

$\beta$ -GlcT ( <b>BGT</b> , phage T4)	<b>n.c.</b>	<b>inv</b>
$\beta$ 4-GlcNAcT ( <b>MurG</b> , <i>E.coli</i> )	<b>GT28</b>	<b>inv</b>
$\beta$ -GlcT ( <b>GtfB</b> , <i>M. orientalis</i> )	<b>GT1</b>	<b>inv</b>

## SpsA-fold



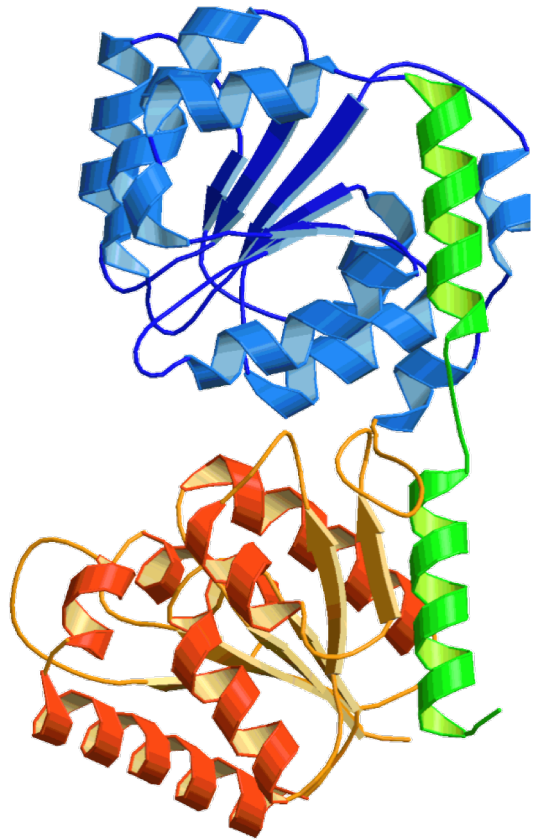
### (Prokaryotes)

<b>SpsA</b> ( <i>B. subtilis</i> )	<b>GT2</b>	<b>inv</b>
$\alpha$ 4-GalT ( <b>LgtC</b> , <i>N.meningitis</i> )	<b>GT8</b>	<b>ret</b>

### (Eucaryotes)

$\beta$ 4-GalT1 (bovine)	<b>GT7</b>	<b>inv</b>
$\beta$ 2-GlcNAcT ( <b>GnT I</b> , rabbit)	<b>GT13</b>	<b>inv</b>
$\beta$ 3-GlcAT I (human)	<b>GT43</b>	<b>inv</b>
$\alpha$ 3-GalT (bovine)	<b>GT6</b>	<b>ret</b>
Glycogenin (rabbit)	<b>GT8</b>	<b>ret</b>
$\alpha$ 3-GalNacT ( <b>GTA</b> , human)	<b>GT6</b>	<b>ret</b>
$\alpha$ 3-GalT ( <b>GTB</b> , human)	<b>GT6</b>	<b>ret</b>

# Nadrodina s BGT foldem

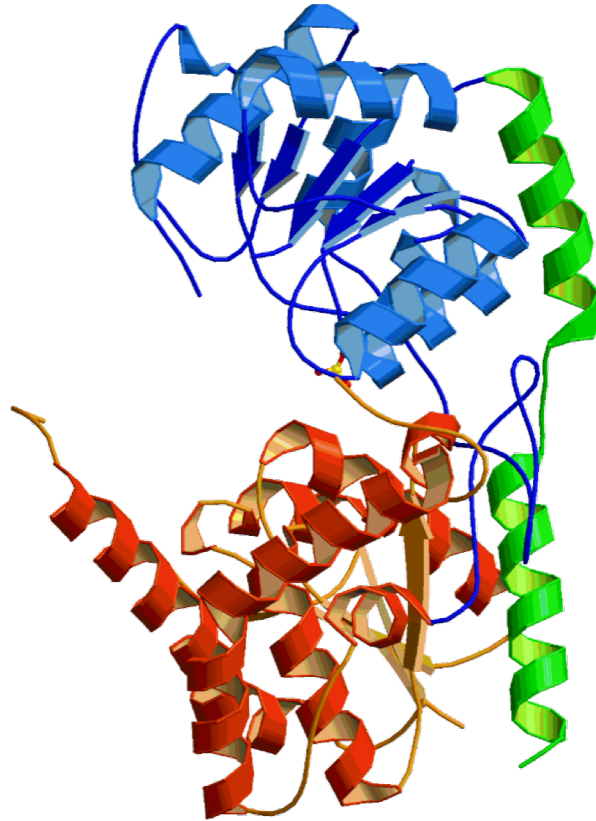


**MurG ( $\beta$ -GlcNAcT)**

**GT28**

*E. coli*

Ha *et al.*, 2000

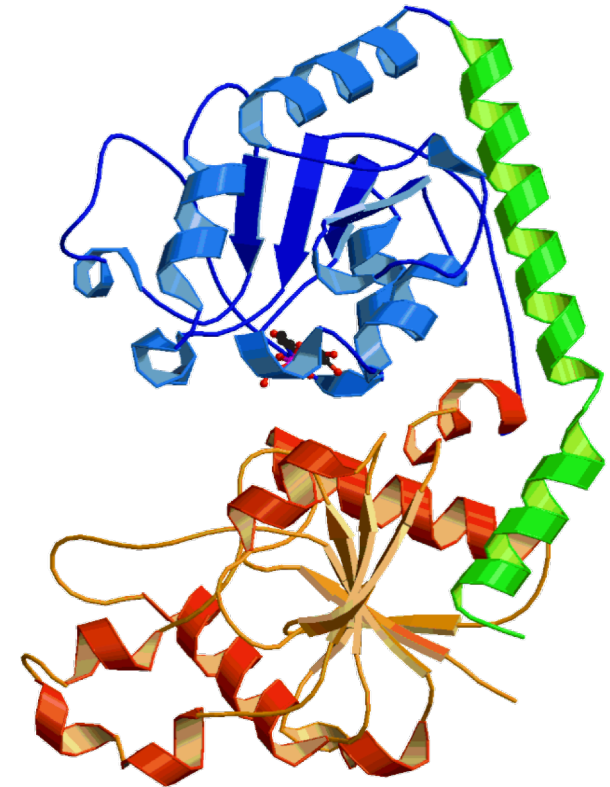


**GtfB ( $\beta$ -GlcT)**

**GT1**

*A. orientalis*

Mulichak *et al.*, 2001



**BGT ( $\beta$ -GlcT)**

**n.c.**

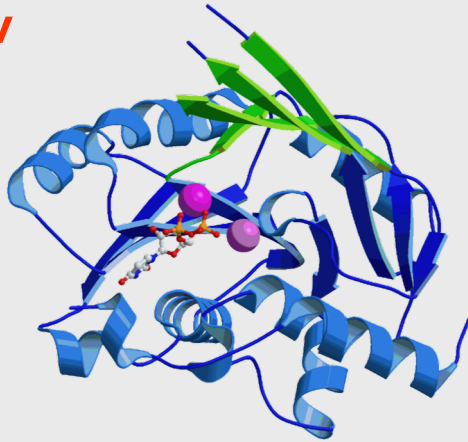
Phage T4

Vrielink *et al.*, 1994

# Nadrodina s SpsA foldem

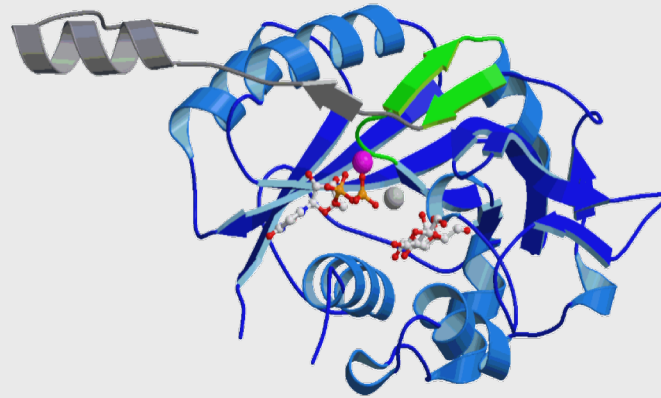
Společná NBD

Inv



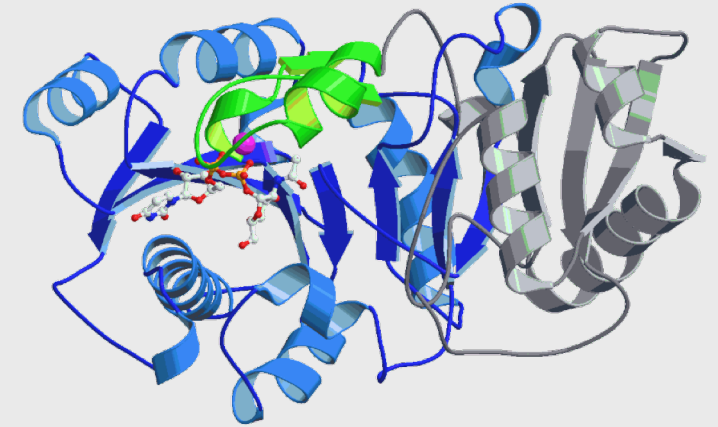
**SpsA [GT2]**

Charnok *et al*, 1999, 2001



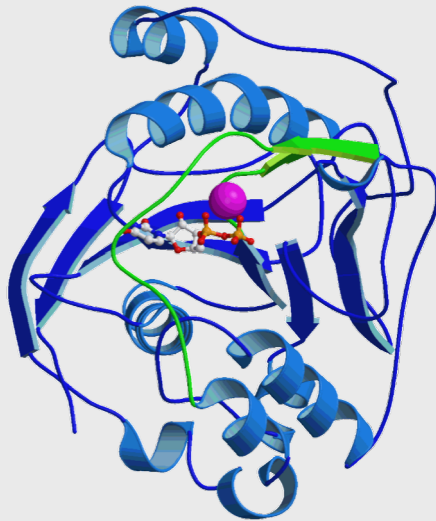
**Hum  $\beta$ 3-GlcAT [GT43]**

Pedersen *et al*, 2000



**Rabbit GnT I [GT13]**

Ünlügil *et al*, 2000

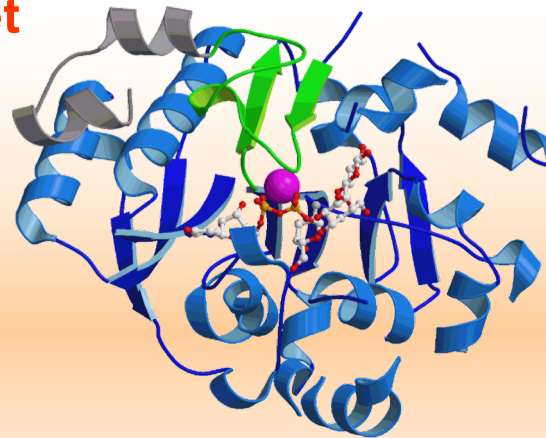


**Bovine  $\beta$ 4-GalT [GT7]**

Gastinel *et al*, 1999

Ramakrishnan *et al*, 2001, 2002

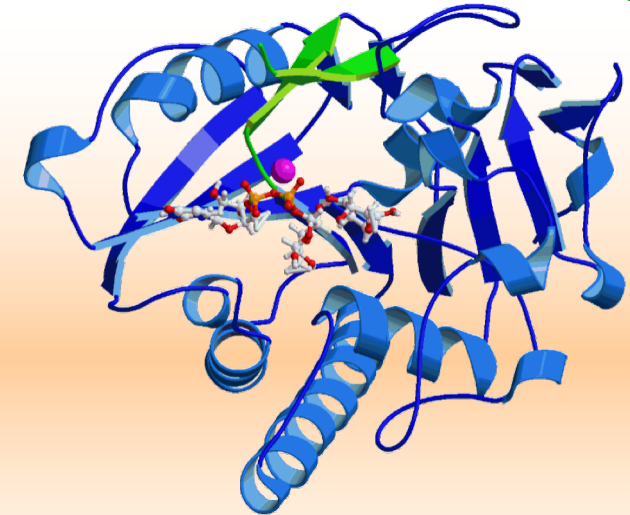
Ret



**LgtC ( $\alpha$ 4-GalT) [GT8]**

*Neisseria meningitidis*

Persson *et al*, 2001



**Bovine  $\alpha$ 3-GalT [GT6]**

Gastinel *et al*, 2001

Boix *et al*, 2001, 2002

# Předpověď 3D-struktury/ foldu

- Klasifikace proteinů
- Předpověď funkce
- Vytvoření modelu pro další studium
  
- Threading - „navlékání“
- Homology modeling
- *Ab initio* 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 multidoménných struktur je potřeba aminokyselinovou sekvenci rozdělit na jednotlivé domény a analyzovat je separátně



## **PHYRE2 (3D-PSSM)**

<http://www.sbg.bio.ic.ac.uk/phyre2>

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

## **The PSIPRED Protein Sequence Analysis Workbench**

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

GenTHREADER            Rapid fold recognition, matching your sequence against a library of whole PDB chains.

pGenTHREADER            Highly sensitive fold recognition using profile-profile comparison (whole chain library).

pDomTHREADER            Highly sensitive homologous domain recognition using profile-profile comparison (domain library).

## **I-TASSER**

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

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.

# Threading

## Protein Homology/analogY Recognition Engine

*(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 porovnávané struktury společně se shodou jejich sekundárních struktur.
- Jediné zásahy do aminokyselinové páteře templátu jsou při modelování inzercí a delecí v sekvenci oproti porovnávané struktuře.

# Phyre2

ARLDVIPMIYCGHGY



Homologous  
sequences

User sequence

Search the 10 million known  
sequences for homologues  
using PSI-Blast.

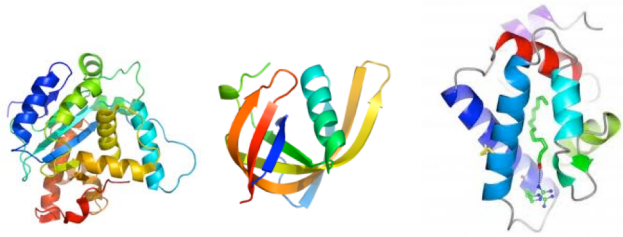
# Phyre2



Capture the mutational propensities at each position in the protein

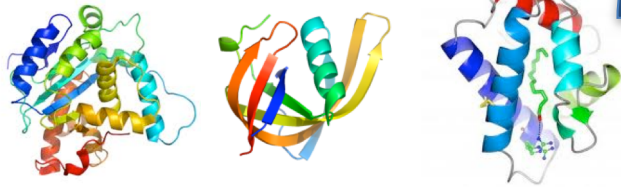
## An evolutionary fingerprint

# Phyre2



~ 65,000 known 3D structures

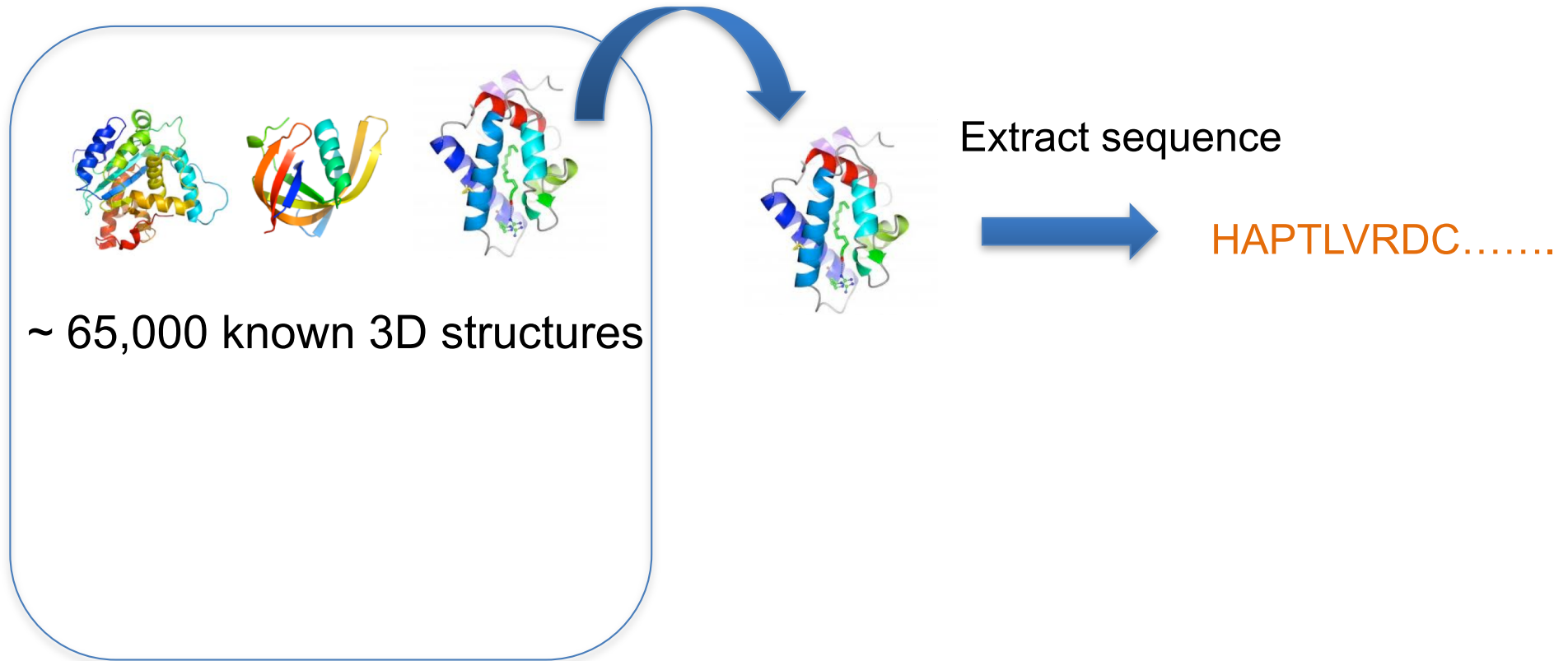
# Phyre2



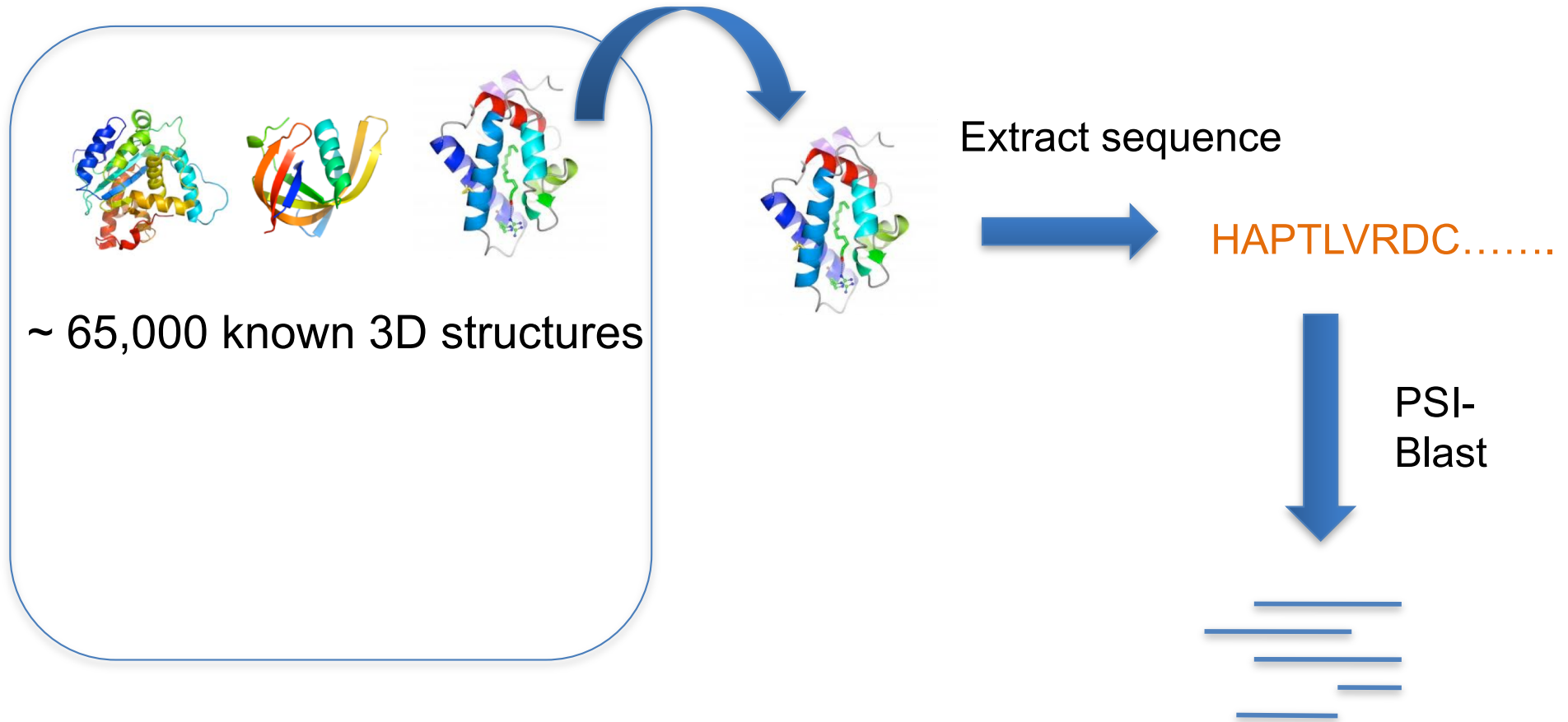
~ 65,000 known 3D structures



# Phyre2

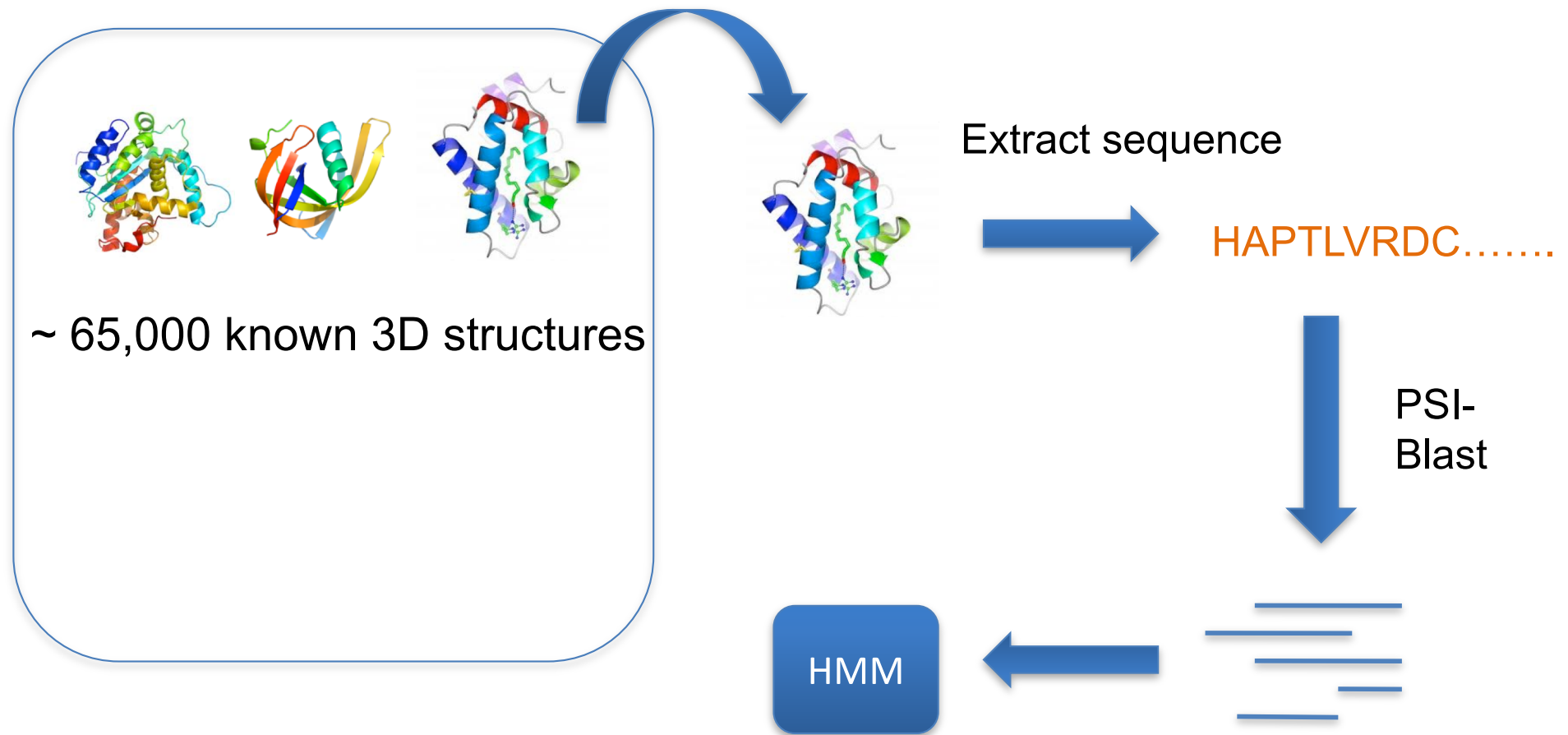


# Phyre2



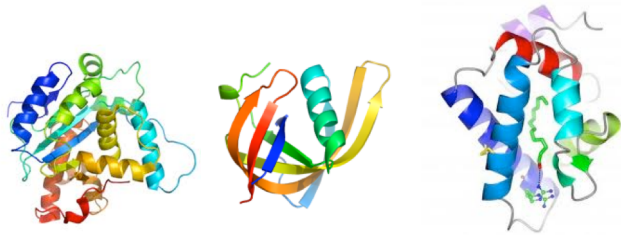


# Phyre2



Hidden Markov model  
for sequence of KNOWN structure

# Phyre2



~ 65,000 known 3D structures



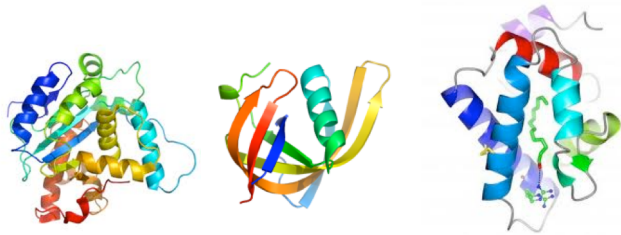
HMM

HMM

HMM

~ 65,000 hidden Markov models

# Phyre2



~ 65,000 known 3D structures



Hidden Markov Model  
Database of  
**KNOWN  
STRUCTURES**

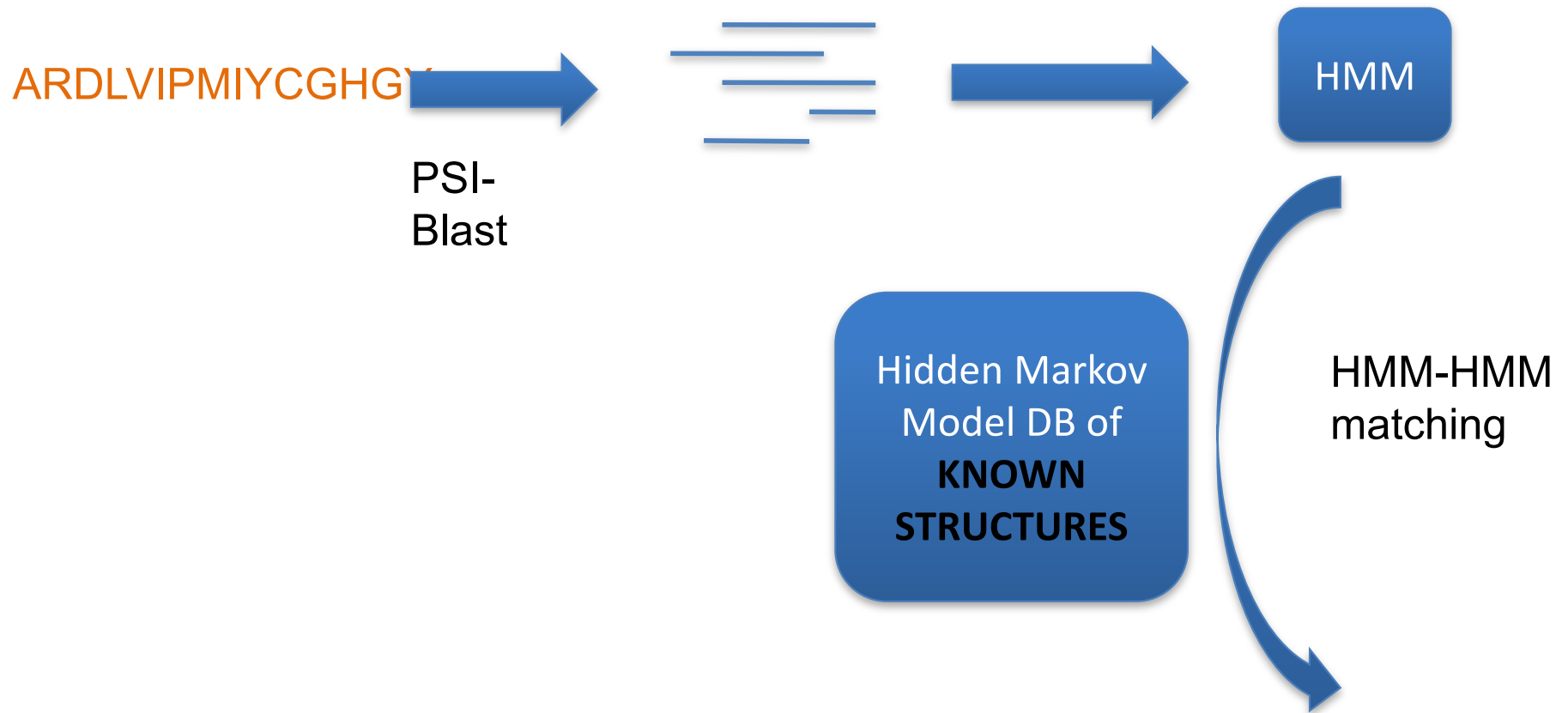
# Phyre2



Capture the mutational propensities at each position in the protein

## An evolutionary fingerprint

# Phyre2

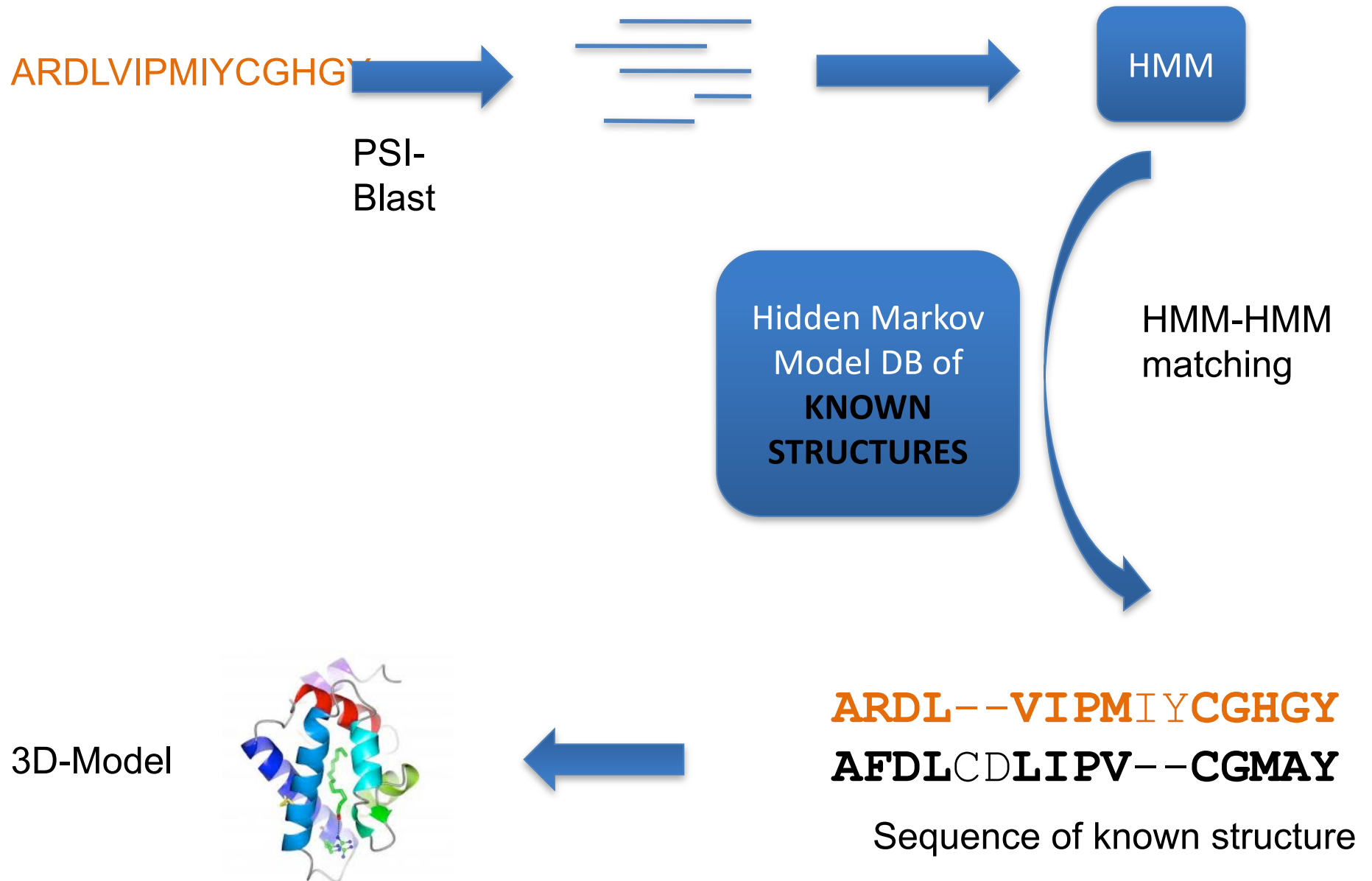


Alignments of user sequence to known structures ranked by confidence.

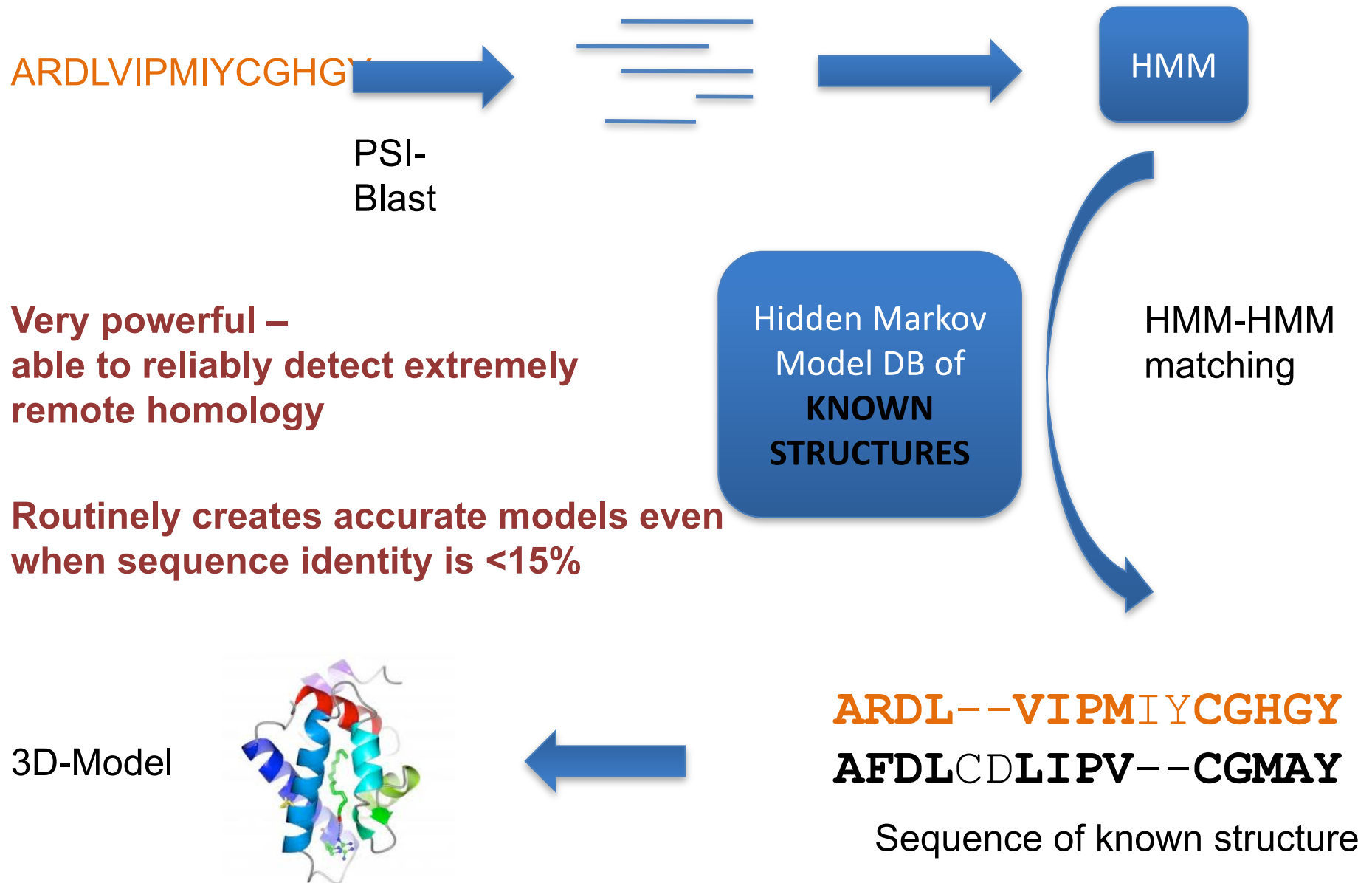
**ARLD--VIPMIYCGHGY**  
**AFDLCDLIPV--CGMAY**

Sequence of known structure

# Phyre2



# Phyre2



**Very powerful – able to reliably detect extremely remote homology**

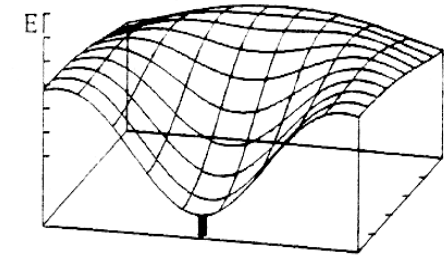
**Routinely creates accurate models even when sequence identity is <15%**

# Knowledge-based Potentials used by Fold Recognition methods

**Calculation of  
Mean Force Potentials**



**Databank of  
3D structures**



**Mean force potential  
derived from the  
databank**

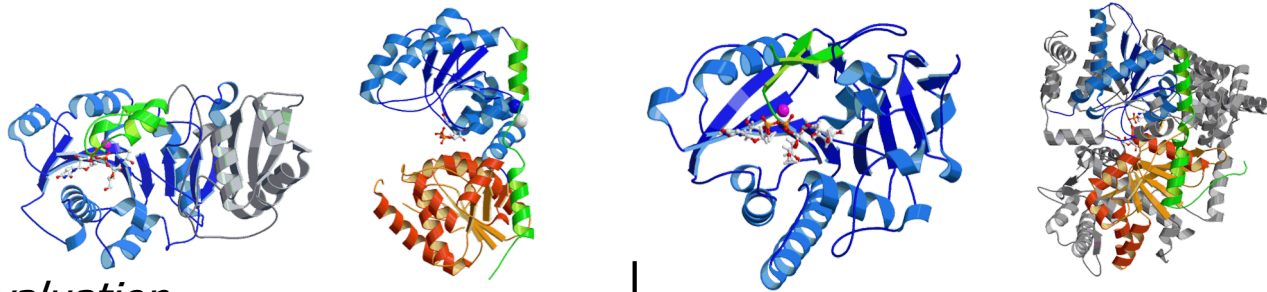


**Building up a proposed model  
from amino acid sequence  
that is based on a real protein  
structure**

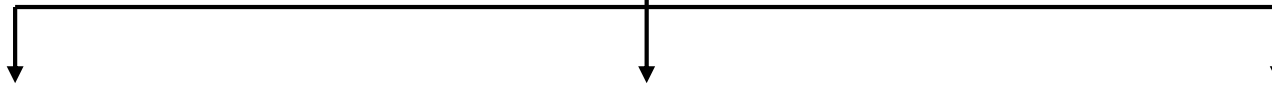


SDVDIEAGQTLVQVNVNISNGETWVAIQLPAYRSFDLVFENVSPSTSGSVLVAQMA  
PQSGGVYGSNYSGSGWGNDLGGGGFYGYSEAKWMCLWPANRSGPNSKTGIYG  
TCKLMNLNQSNAVPSVTSNLFAPTAYKNEPGYANVGGCCQKIRGLASSIQFAFALH  
GGNVPQNTDTFSGGTIKVYGWN

*3D-fold calculation based  
on known structures*



*Model quality evaluation*



**pair**  
residue-residue  
interactions

**surface**  
residue-solvent  
interactions

**pair/surface**  
residue-residue and  
residue-solvent interaction

**“Quality” scores**

# Phyre<sup>2</sup>

Subscribe to Phyre at Google Groups

Email:

[Visit Phyre at Google Groups](#)

Protein Homology/analogY Recognition Engine V 2.0



## [What's New in Phyre2](#)

E-mail Address	<input type="text"/>
Optional Job description	<input type="text"/>
Amino Acid Sequence	<input type="text"/>
Modelling Mode	Normal <input checked="" type="radio"/> Intensive <input type="radio"/>
	<input type="button" value="Phyre Search"/> <input type="button" value="Reset"/>

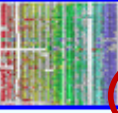
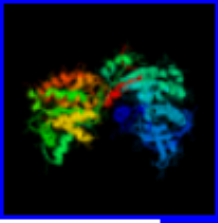

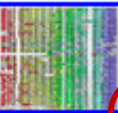
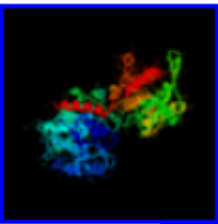

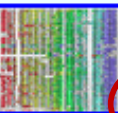
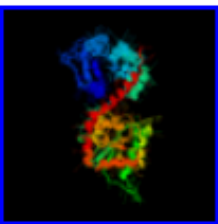

553492 submissions since Feb 14 2011

Glykogensynthasa – rodina GT3 (v rodině v době analýzy nebyla vyřešena 3D-struktura)

[http://www.sbg.bio.ic.ac.uk/phyre/qphyre\\_output/95cbaa7600a9bfff/summary.html](http://www.sbg.bio.ic.ac.uk/phyre/qphyre_output/95cbaa7600a9bfff/summary.html)

To predict functional residues and GO classification, try [ConFunc](#)

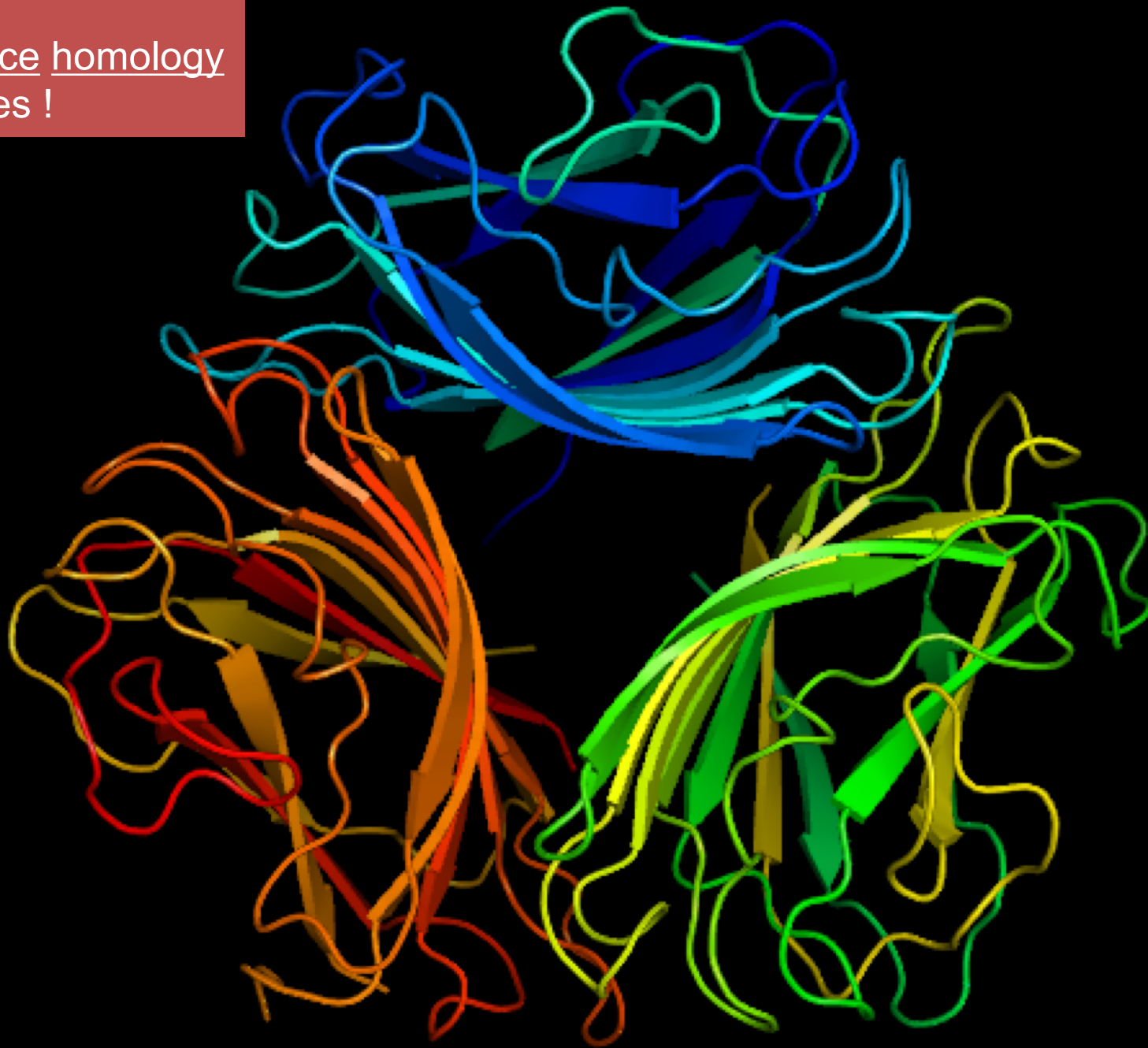
Recognition

Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor	Superfamily
	<a href="#">d2bisa1</a> (length:437) <b>18% i.d.</b>	 	3.9e-36	100 %	n/a	UDP-Glycosyltransferase/glycogen phosphorylase	UDP-Glycosyltransferase/g phosphorylase
	<a href="#">d1rzua</a> (length:477) <b>14% i.d.</b>	 	6.1e-36	100 %	n/a	UDP-Glycosyltransferase/glycogen phosphorylase	UDP-Glycosyltransferase/g phosphorylase
	<a href="#">c3c48A</a> (length:438) <b>11% i.d.</b>	 	6.1e-31	100 %	n/a	<b>PDB header:</b> transferase	<b>Chain: A: PDB Molecule:</b> predicted glycosyltransferases;

A co protein, který nemá v sekvenčních databázích žádný homolog?

RS-20L

No sequence homology  
in databases !





## Fold Recognition

View Alignments	SCOP Code	View Model	E-value	Estimated Precision	Bio Text	Fold/PDB descriptor	Superfamily	Fa
	<a href="#">d1eh9a2</a> (length:67) 24% i.d.	 	50	0 %	n/a	Glycosyl hydrolase domain	Glycosyl hydrolase domain	alpha- Amylas C-termi beta-sh domain
	<a href="#">c2fsdA</a> (length:142) 19% i.d.	 	50	0 %	n/a	<b>PDB header:</b> virus/viral protein	<b>Chain: A: PDB Molecule:</b> putative baseplate protein;	<b>PDBTi</b> commo the rece binding domain lactoco phages crystal s of the h domain phage t
	<a href="#">c2ct4A</a> (length:70) 11% i.d.	 	56	0 %	n/a	<b>PDB header:</b> signaling protein	<b>Chain: A: PDB Molecule:</b> cdc42- interacting protein 4;	<b>PDBTi</b> solution strucur sh3 dor the cdc- interact protein

# 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

- Start SMR-Pipeline in automated mode on BC2-cluster at Thu May 2 08:51:47 2013
- Start BLAST for highly similar template structure identification
- No suitable templates found!
- Run HHSearch to detect remotely related template structures
- Unfortunately, we could not identify useful template structures
- For troubleshooting, please see our article in Nature Protocols:
  - Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J. and Schwede, T. (2009). Protein structure homology modelling using SWISS-MODEL Workspace. Nature Protocols, 4, 1.

## What are protein domains?

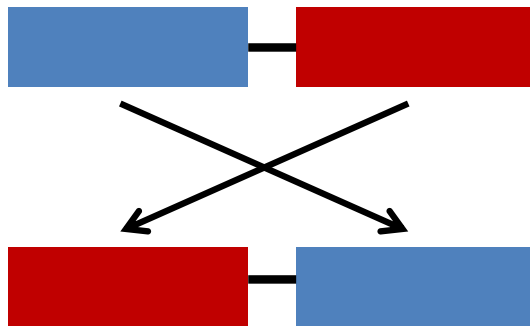
Since the first protein structures were solved, it was apparent that the polypeptide chain **could often fold into one or more distinct regions of structure**. Such substructures, or domains, are considered as the basic units of folding, function and evolution and often have **similar chain topologies** (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.

PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGCVVADGCFTYSSKV  
 PESTGRMPFTLVATIDVGSVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAAPSSQGSNGNQAETGGTGAGNIG  
 GGERDGTFLNPPHIKFGVTALHAANDQTIDIYIDDDPKPAATFKGAGAQQNLGTVLDSGNGRVRVIVMANGR  
 PSRLGSRQVDIFKKSYPFGIIGSEDGADDDYNDGIVFLNWPLG

ERDGTFLNPPHIKFGVTALHAANDQTIDIYIDDDPKPAATFKGAGAQQNLGTVLDSGNGRVRVIVMANGRPSR  
 LGSRQVDIFKKSYPFGIIGSEDGADDDYNDGIVFLNWPLGPLLSASIVSAPVVTSTQTYVDIPGLYLDVAKAGIRDGKLQ  
 VILNVPTPYATGNNFPGIYFAIATNQGCVVADGCFTYSSKVPESTGRMPFTLVATIDVGSVTFVKGQWKSVRGSAM  
 HIDSYASLSAIWGTAAAPSSQGSNGNQAETGGTGAGNIGGGGKLAALAEIKRASQPELAPEDPEDVEHHHHHH



```

#
#=====
EMBOSS_001      1 ----- 0
EMBOSS_001      1 ERDGTFLNPPHIKFGVTALHAANDQTIDIYIDDDPKPAATFKGAGAQQ 50
EMBOSS_001      1 ----- 0
EMBOSS_001     51 NLGTVLDSGNGRVRVIVMANGRPSRLGSRQVDIFKKSYPFGIIGSEDGAD 100
EMBOSS_001      1 -----PLLASIVSAPVVTSETYVDIPGLYLDVAKAGIRD 35
EMBOSS_001     101 DDYNDGIVFLNWPLGPLLSASIVSAPVVTSTQTYVDIPGLYLDVAKAGIRD 150
EMBOSS_001     36 GKLVILNVPTPYATGNNFPGIYFAIATNQGCVVADGCFTYSSKVPESTGR 85
EMBOSS_001     151 GKLVILNVPTPYATGNNFPGIYFAIATNQGCVVADGCFTYSSKVPESTGR 200
EMBOSS_001     86 MPFTLVATIDVGSVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAAPSSQ 135
EMBOSS_001     201 MPFTLVATIDVGSVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAAPSSQ 250
EMBOSS_001     136 GSGNQAETGGTGAGNIGGGGERDGTFLNPPHIKFGVTALHAANDQTID 185
EMBOSS_001     251 GSGNQAETGGTGAGNIGGGG----- 271
EMBOSS_001     186 IYIDDDPKPAATFKGAGAQQNLGTVLDSGNGRVRVIVMANGRPSRLGS 235
EMBOSS_001     272 -----KLAALAEIKRASQPELAPEDPEDVEHHHHHH 283
EMBOSS_001     236 RQVDIFKKSYPFGIIGSEDGADDDYNDGIVFLNWPLG 271
EMBOSS_001     284 -QPE-----LAPEDPEDVEHHHH-----HHH 302
  
```

# 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“  $S$  ( $f_i^{\text{linker}}$  a  $f_i^{\text{domain}}$  je frekvence zastoupení aminokyseliny  $i$  v úsecích linkeru a domény)
- :

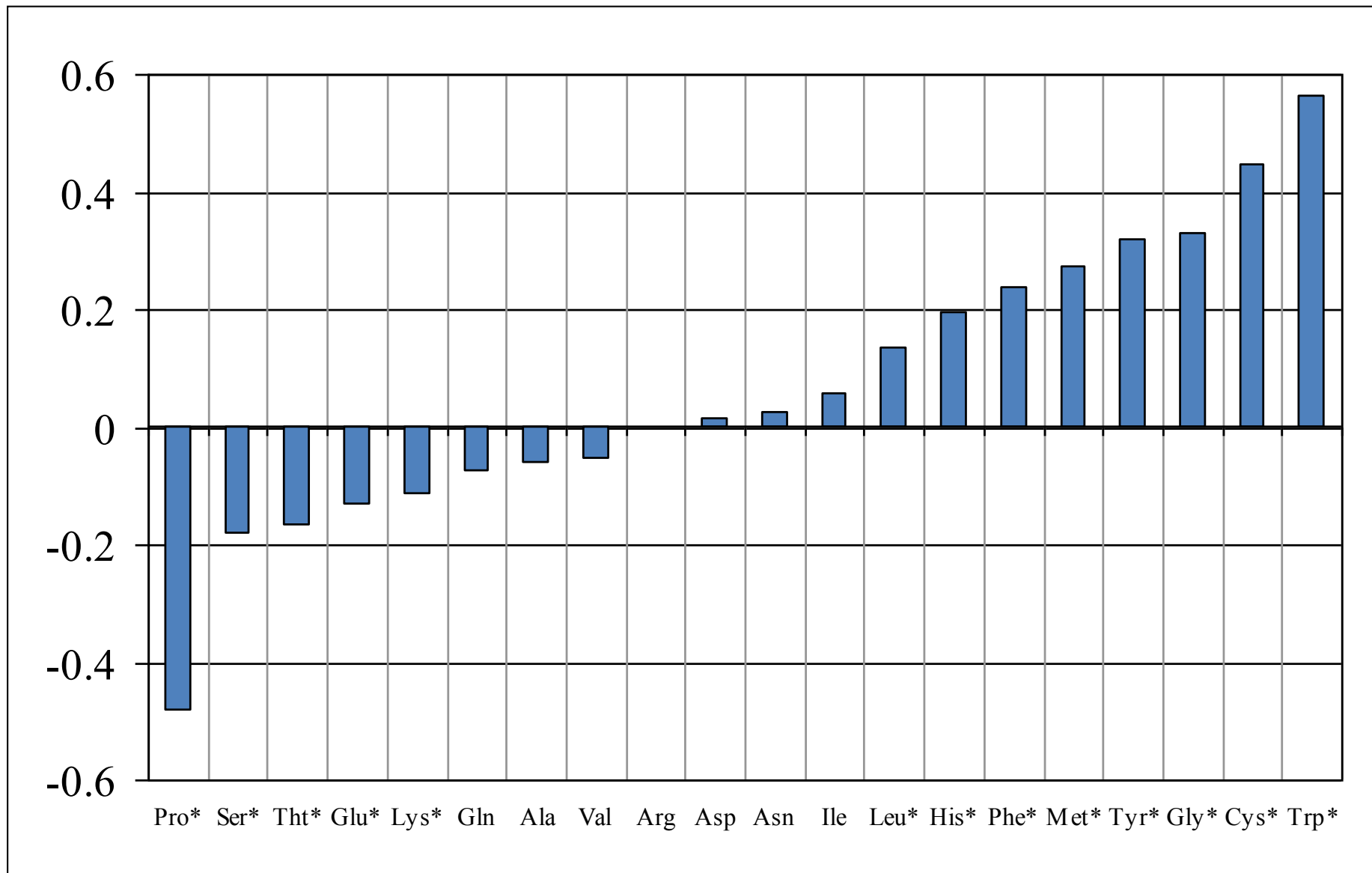
$$S_i = - \ln \frac{f_i^{\text{linker}}}{f_i^{\text{domain}}}$$

# 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

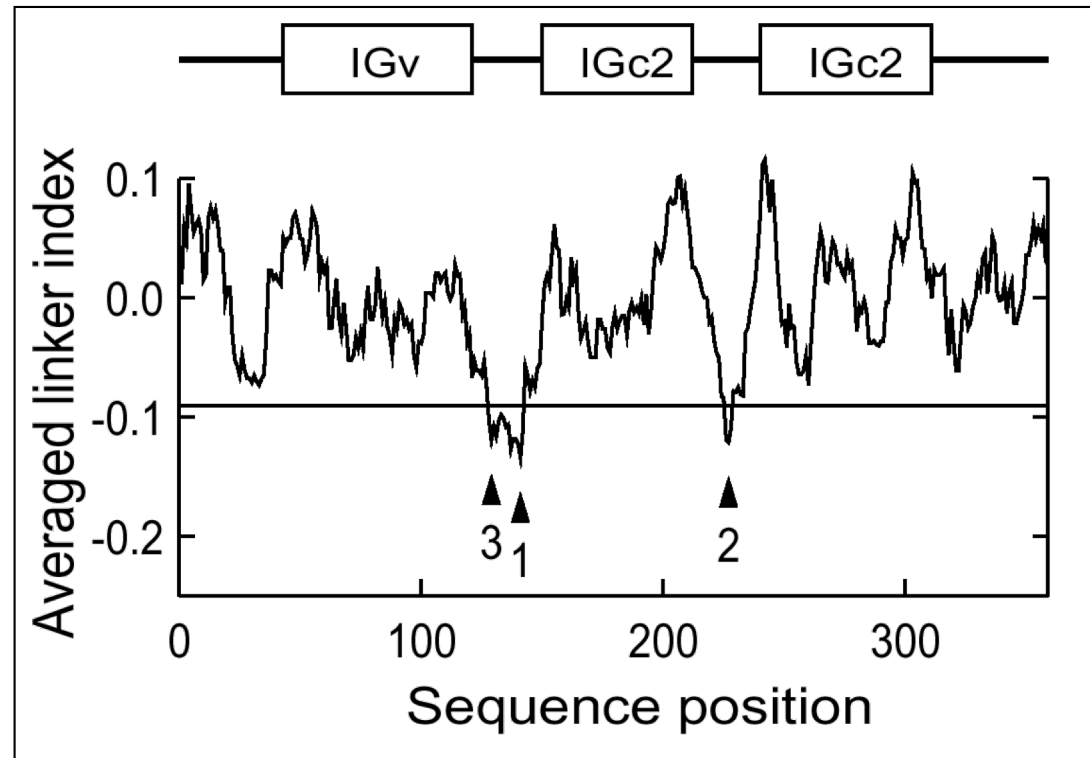
Aminokyselinová sekvence  
Q24372

- 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.*

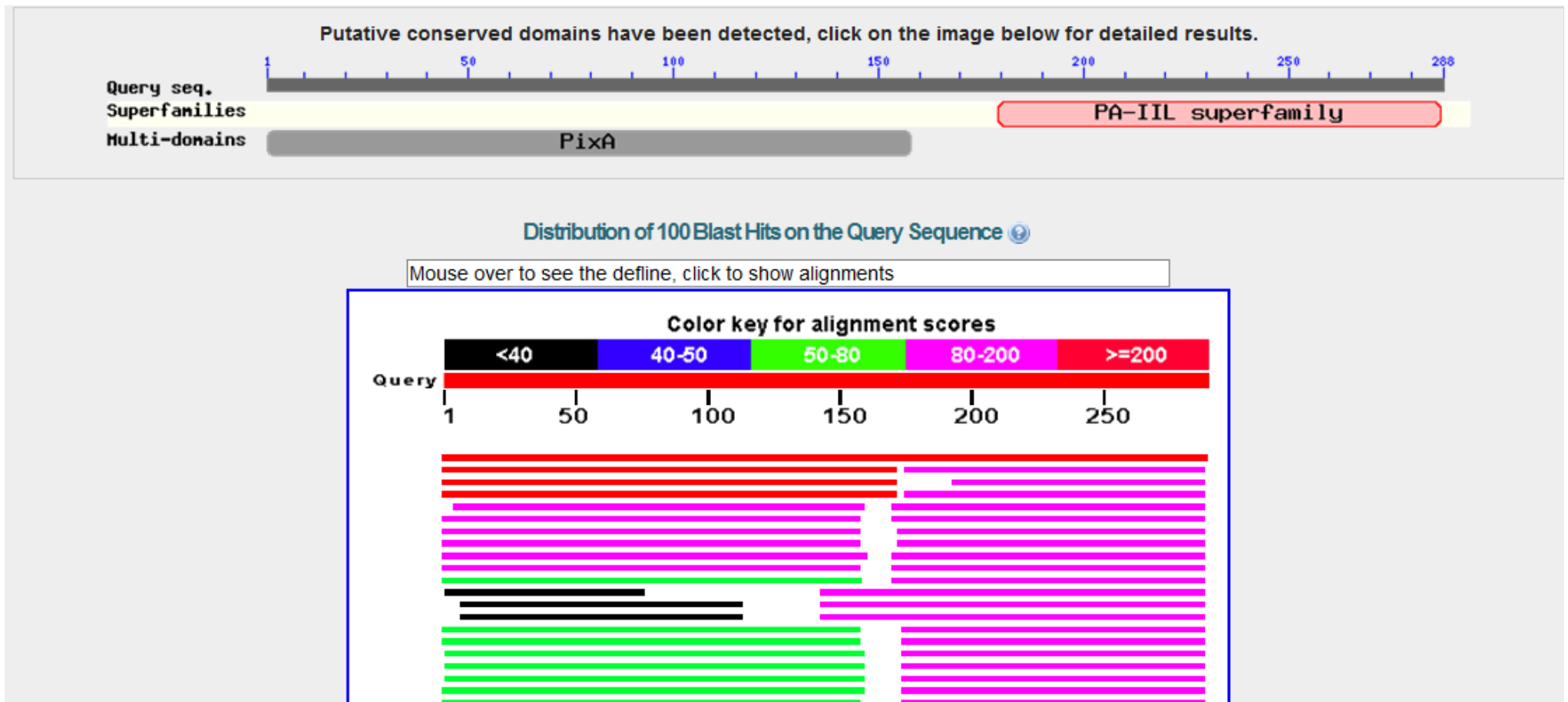


Domény předpovídají i programy používané primárně pro jiné účely na základě podobnosti s dosud identifikovanými doménami/funkčními jednotkami

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

Prohledávání databází známých aminokyselinových sekvencí

➤ celý protein



# NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

➤ celý protein

**Conserved domains on** [1cl|15110] [View concise result](#) ?

Local query sequence

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

Query seq. 1 50 100 150 200 250 288

Non-specific hits PA-IIL

Superfamilies PA-IIL superfamily

Multi-domains PixA

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

**List of domain hits** ?

	Description	PssmId	Multi-dom	E-value
[+]PA-IIL[ <a href="#">pfam07472</a> ], Fucose-binding lectin II (PA-IIL); In <i>Pseudomonas aeruginosa</i> the fucose-binding lectin II (PA-IIL) contributes to the ...		203639	no	3.60e-45
[+]PixA[ <a href="#">pfam12306</a> ], Inclusion body protein; This family of proteins is found in bacteria. Proteins in this family are typically ...		204875	yes	4.88e-43

# NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí


➤ celý protein



NCBI

Home Search Site Map Entrez CDD Structure Protein Help

**pfam07472: PA-IIL** 7



**Fucose-binding lectin II (PA-IIL)**  
In *Pseudomonas aeruginosa* the fucose-binding lectin II (PA-IIL) contributes to the pathogenic virulence of the bacterium. PA-IIL functions as a tetramer when binding fucose. Each monomer is comprised of a nine-stranded, antiparallel beta-sandwich arrangement and contains two calcium cations that mediate the binding of fucose in a recognition mode unique among carbohydrate-protein interactions.

- Links
- Statistics
- Structure

### PubMed References

Structural basis for oligosaccharide-mediated adhesion of *Pseudomonas aeruginosa* in the lungs of cystic fibrosis patients. *Nat Struct Biol*. 2002 Dec; 9(12):915-921

pfam07472 is a member of the superfamily cl06486.

Sequence Alignment 7

Reformat: Format: Compact Hypertext Row Display: up to 10 Color Bits: 2.0 bit Type Selection: the most diverse members

1UGX_A	8	FILPANTFSGVIAFANAAHQTIQVIVDS	VVK	AIFGSGTSDK	[1].LGS	[2].LSSGS	GAIK	63
qi 81656026	7	FILPARIHFGVTVLVMSAATQSEVLIIVDS	KFR	AAFSGVGIGD	[1].LGT	[2].LSSGS	GRVR	64
qi 79468912	234	FQLPSEIKLSLSAYGRTTHGQTIKVIYID	QLV	DILISQGVNSV	LGF	[2].YSSST	GRVC	290
qi 123466640	14	FSIPFHTDFRAIFFAQAAGQSEIKLFIQD	SQK	[2].AYEKLTTRDGP	[1].EAT	LSSGS	GKIR	71
qi 123570089	187	FSLPFTAFKALFYAQAADRQDLKLIYID	APK	[2].AIFVGSXEDGV	[1].LFT	LSSGS	GKIR	246
qi 123569196	174	FSLPFRKIFGVIALTSAANHQTIIDIVDD	MPK	[2].AIFKAGVQDQ	[1].LGT	[2].LSSGS	GRVR	233
ZXR4_A	7	FSLPFRKIFGVIALTSAANHQTIIDIVDD	DPK	[2].AIFKAGVQDQ	[1].LGT	[2].LSSGS	GRVR	66
ZBO1_A	8	FILPARIHFGVTVLVMSAATQSEVLIIVDS	KFR	AAFSGVGIGD	[1].LGT	[2].LSSGS	GRVR	65
qi 107102893	2	FILPANTFSGVIAFANSSGQTVIVVIVDS	KTA	AIFGSGTSDK	[1].LGT	[2].LSSGS	[1].GRVQ	60
ZVRV_A	14	FSIPFHTDFRAIFFAQAAGQSEIKLFIQD	[2].EPA	AYEKLTTRDGP	[1].EAT	LSSGS	GKIR	71

# NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

➤ celý protein

**pfam12306: PtxA**

**Inclusion body protein**  
This family of proteins is found in bacteria. Proteins in this family are typically between 173 and 191 amino acids in length. PtxA is thought to be specifically produced in *Xenorhabdus nematophila*. It is an inclusion body protein.

**Links** ➤

**Statistics** ➤

**Structure** ➤

**PubMed References** 🔗

[Analysis of the PtxA inclusion body protein of \*Xenorhabdus nematophila\*. J. Bacteriol. 2006 Apr; 188\(7\):2706-2710](#)

pfam12306 is classified as a model that may span more than one domain.  
pfam12306 is not assigned to any domain superfamily.

**Sequence Alignment**

Reformat:  Row Display:  Color Bits:  Type Selection:

gi 123655921	2	-[2]-NIVDILVTEIDVDI	ILE	[17]-S	-[2]-PTQL	[4]-SNG	[7]-VHVARED	[7]-GSELAVNLRQGD	84
gi 123464695	13	-[2]-QSIQILAVIDIDY	ENK	[10]-N	PTGI	[1]-SIA	LPHLNGEI	[8]-TGNLGLKLNFGD	77
gi 123180777	10	-[2]-QDINILAVIDIEH	VVK	[10]-A	PTGI	[1]-ENG	QFLICIGA	[7]-IADLEITAYPGD	73
gi 53717990	9	-[2]-QKIRVLFVIDIAY	IRS	[10]-Q	PTGI	[1]-ENS	QILLCIGS	[8]-TGOLEFRANFGD	73
gi 254248506	27	-[2]-QQIDILAVIDIEY	EKL	[10]-L	PIAV	[1]-ERA	VRLLYTGA	[8]-VADPVLILYPGD	81
gi 170734880	2	-[2]-VRCDALAVDAVI	LLS	[10]-A	PTVI	[1]-GRS	IYVLSFGD	[7]-DGRLEFAGLSPGD	85
gi 83748592	18	-[2]-LTIINVTINQVDA	ILA	[10]-M	PTAI	[1]-EAY	IKHVSDDP	[8]-PGNITLDAHVED	82
gi 134279425	20	-[2]-SRVOLLVVIDSDY	VKE	[10]-I	PTPV	[1]-SRA	LFVICAGS	[8]-SGEAICTAAYGD	82
gi 170702239	10	-[2]-QKITLLAVINAEK	[1]-ENK	[10]-R	PTGI	[1]-ENS	QILLCHDP	[8]-ANIKFYAKQFD	78
gi 258424079	20	-[2]-QIVVWVFLVDIAY	IYA	[11]-K	EMPI	[1]-ENS	EVMACSFV	[7]-IADLSYVPRQIS	84

# 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/>**

The screenshot displays the InterProScan Results page. At the top, there is a navigation bar with links for Research, Training, Industry, About Us, Help, Site Index, and RSS. Below this, the breadcrumb trail reads: EBI > Tools > Protein Functional Analysis > InterProScan Sequence Search. The main heading is "InterProScan Results", with tabs for Summary Table, Tool Output, Visual Output (selected), Submission Details, and Submit Another Job. A "Download in SVG format" button is visible. The main content area shows the InterProScan (version: 4.8) results for a sequence named "Sequence\_1" with a length of 288 and CRC64: 3FAE4C40C2498B64. The job was launched on Wed, May 16, 2012 at 17:31:03 and finished at 17:35:39. The results are presented as a horizontal bar chart comparing the "Query Sequence" (length 288) with "InterPro Match" (length 1). Two matches are shown: IPR010907 (Calcium-mediated lectin) and IPR021087 (Uncharacterised protein family PixA/AidA). The IPR010907 match is further detailed with sub-domain annotations: G3DSA:2.60.120.400 (no description), PF07472 (PA-III), and SSF82026 (Calcium-mediated lectin). The IPR021087 match is annotated with PF12306 (PixA). A legend at the bottom identifies various database sources: PRODOM, HAMAP, PRINTS, PROSITE, PIR, SUPERFAMILY, PFAM, SIGNALP, SMART, TMHMM, TIGRFAMs, PANTHER, PROFILE, and GENE3D. The footer contains the copyright notice: © European Bioinformatics Institute 2006-2012. EBI is an Outstation of the European Molecular Biology Laboratory.

# Proč potřebujeme predikci domén?

- Prohledávání sekvenčních databází bez predikce domén může být neúspěšné
- Automatická predikce struktury se zaměří jen na nejlépe „definovanou“ část
- .....

# Phyre – whole protein [http://www.sbg.bio.ic.ac.uk/phyre2/phyre2\\_output/a132b051273537c4/summary.htm](http://www.sbg.bio.ic.ac.uk/phyre2/phyre2_output/a132b051273537c4/summary.htm)

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	<a href="#">c2vnvC</a> <input type="radio"/> <input type="checkbox"/>	 <input type="button" value="Alignment"/>		100.0	60	<b>PDB header:</b> sugar-binding protein <b>Chain:</b> C: <b>PDB Molecule:</b> bcla; <b>PDBTitle:</b> crystal structure of bcla lectin from burkholderia2 cenocepacia in complex with alpha-methyl-mannoside at 1.73 angstrom resolution
2	<a href="#">c2xr4A</a> <input type="radio"/> <input type="checkbox"/>	 <input type="button" value="Alignment"/>		100.0	43	<b>PDB header:</b> sugar binding protein <b>Chain:</b> A: <b>PDB Molecule:</b> lectin; <b>PDBTitle:</b> c-terminal domain of bc2l-c lectin from burkholderia cenocepacia
3	<a href="#">d2chha1</a> <input type="radio"/> <input type="checkbox"/>	 <input type="button" value="Alignment"/>		100.0	37	<b>Fold:</b> Calcium-mediated lectin <b>Superfamily:</b> Calcium-mediated lectin <b>Family:</b> Calcium-mediated lectin



# NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

➤ celý protein

**Conserved domains on** [1c1|15110] [View concise result](#) ?

Local query sequence

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

Query seq. 1 50 100 150 200 250 288

Non-specific hits

Superfamilies


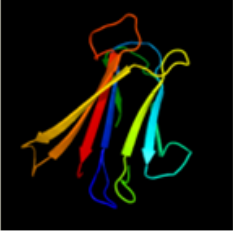

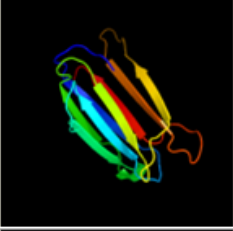

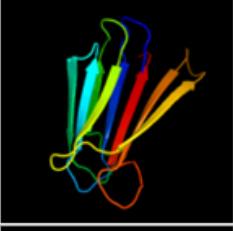
Multi-domains PixA

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

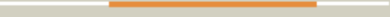
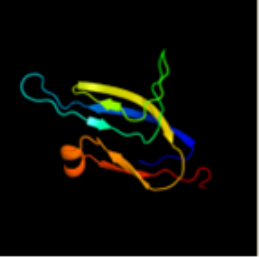
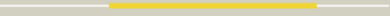
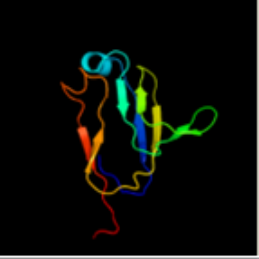

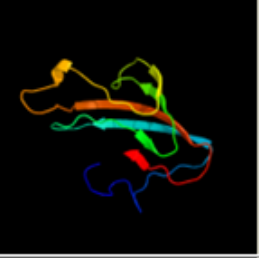
**List of domain hits** ?

	Description	PssmId	Multi-dom	E-value
<a href="#">+</a> PA-IIL[ <a href="#">pfam07472</a> ], Fucose-binding lectin II (PA-IIL); In Pseudomonas aeruginosa the fucose-binding lectin II (PA-IIL) contributes to the ...		203639	no	3.60e-45
<a href="#">+</a> PixA[ <a href="#">pfam12306</a> ], Inclusion body protein; This family of proteins is found in bacteria. Proteins in this family are typically ...		204875	yes	4.88e-43



# Phyre – C-term [http://www.sbg.bio.ic.ac.uk/phyre2/phyre2\\_output/e332b1ecabb8d0a6/summary.html](http://www.sbg.bio.ic.ac.uk/phyre2/phyre2_output/e332b1ecabb8d0a6/summary.html)

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	<a href="#">c2xr4A</a> <input type="radio"/> <input type="checkbox"/>	 <input type="button" value="Alignment"/>		100.0	44	<b>PDB header:</b> sugar binding protein <b>Chain:</b> A: <b>PDB Molecule:</b> lectin; <b>PDBTitle:</b> c-terminal domain of bc2l-c lectin from burkholderia cenocepacia
2	<a href="#">c2vnvC</a> <input type="radio"/> <input type="checkbox"/>	 <input type="button" value="Alignment"/>		100.0	62	<b>PDB header:</b> sugar-binding protein <b>Chain:</b> C: <b>PDB Molecule:</b> bclA; <b>PDBTitle:</b> crystal structure of bclA lectin from burkholderia2 cenocepacia in complex with alpha-methyl-mannoside at 1.73 angstrom resolution
3	<a href="#">d1uzva</a> <input type="radio"/> <input type="checkbox"/>	 <input type="button" value="Alignment"/>		100.0	30	<b>Fold:</b> Calcium-mediated lectin <b>Superfamily:</b> Calcium-mediated lectin <b>Family:</b> Calcium-mediated lectin


# Phyre – n-term [http://www.sbg.bio.ic.ac.uk/phyre2/phyre2\\_output/e332b1ecabb8d0a6/summary.html](http://www.sbg.bio.ic.ac.uk/phyre2/phyre2_output/e332b1ecabb8d0a6/summary.html)

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	<a href="#">c1sddB</a> <input type="radio"/> <input type="checkbox"/>	 <input type="button" value="Alignment"/>		83.7	9	<b>PDB header:</b> blood clotting <b>Chain:</b> B; <b>PDB Molecule:</b> coagulation factor v; <b>PDBTitle:</b> crystal structure of bovine factor vai
2	<a href="#">c3cdzB</a> <input type="radio"/> <input type="checkbox"/>	 <input type="button" value="Alignment"/>		76.1	6	<b>PDB header:</b> blood clotting <b>Chain:</b> B; <b>PDB Molecule:</b> coagulation factor viii light chair <b>PDBTitle:</b> crystal structure of human factor viii
3	<a href="#">d1kbva2</a> <input type="radio"/> <input type="checkbox"/>	 <input type="button" value="Alignment"/>		68.0	13	<input type="button" value="Info"/> <b>Fold:</b> Cupredoxin-like <b>Superfamily:</b> Cupredoxins <b>Family:</b> Multidomain cupredoxins

# Swissprot – whole protein



Universität Basel  
The Center for Molecular Life Sciences





## SWISS-MODEL Workspace


Modelling Tools Repository Documentation


[ myWorkspace ] [ login ]


**Workunit: P000007 - Overview**



Print/Save this page as 

**Model Summary** 



<b>Model information:</b>	
Modelled residue range:	169 to 288
Based on template:	[2vnnD] (1.7 Å)
Sequence Identity [%]:	56.35
Evalue:	0.00e-1
<b>Quality information:</b>	[details]
QMEAN Z-Score: -0.71	

**Quaternary structure information:** [details]

Template (2vnn): DIMER  
Model built: SINGLE CHAIN


**Ligand information:** [details]

Ligands in the template: CA: 3, MMA: 1, SO4: 1.  
Ligands in the model: CA: 2

logs: [Templates] [Alignment] [Modelling]

display model: as [pdb] - as [DeepView project] - in [AstexViewer]

download model: as [pdb] - as [Deepview project] - as [text]

**Global Model Quality Estimation**  [ +/- ]

# Swissprot - only N terminal part

## Computation of this workunit has stopped.

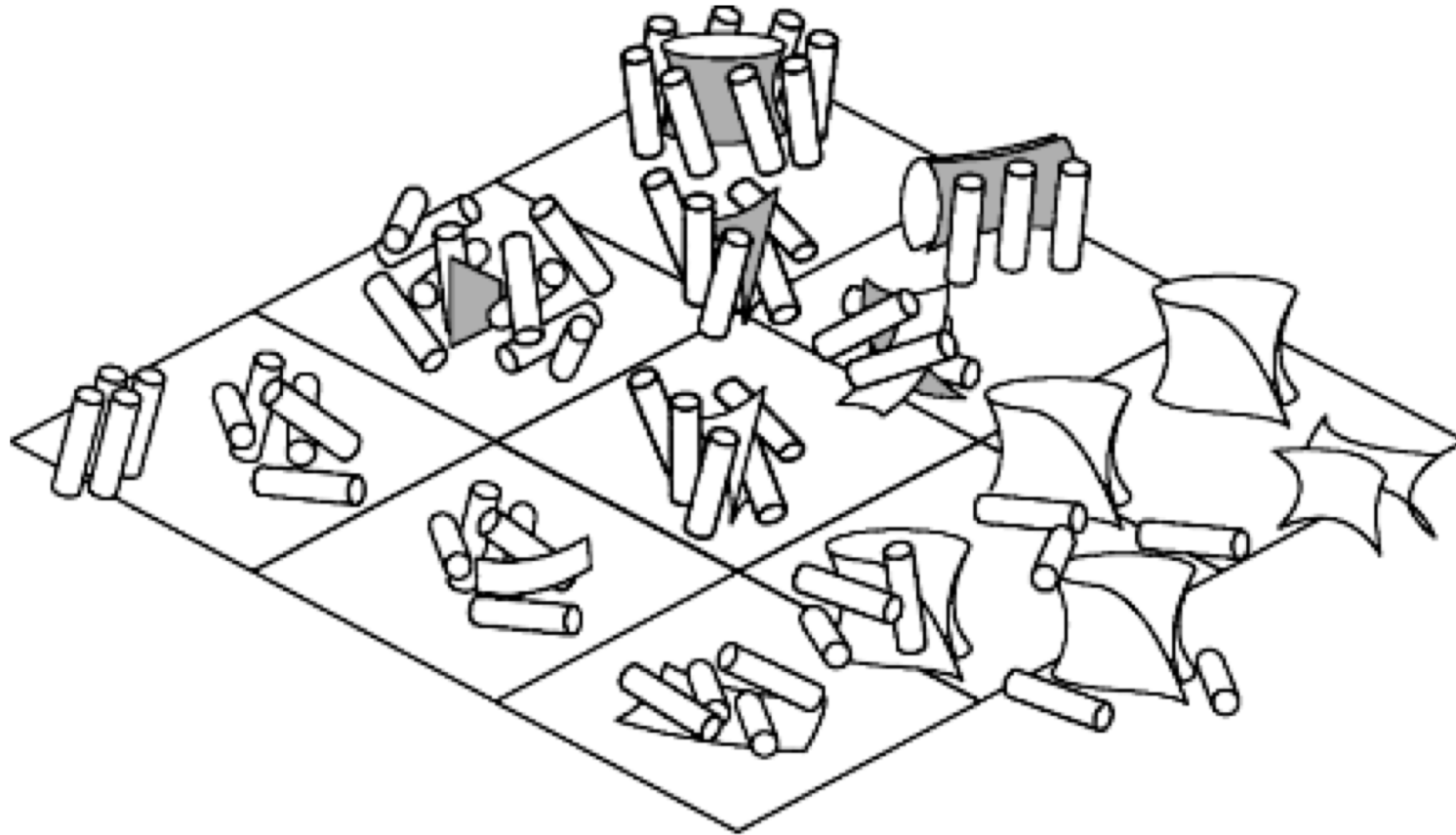
Please see the following log report for details:

Started: Thu May 17 15:21:24 2012 (sms\_automode\_2011)

Reading user input sequence

- Start SMR-Pipeline in automated mode on BC2-cluster at Thu May 17 13:21:24 2012
  
- Start BLAST for highly similar template structure identification
- No suitable templates found!
  
- Run HHSearch to detect remotely related template structures
- Unfortunately, we could not identify useful template structures
  
- For troubleshooting, please see our article in Nature Protocols:  
  
- Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J. and Schwede, T. (2009). Protein structure homology modelling using SWISS-MODEL Workspace. Nature Protocols, 4, 1.
  
- Workspace Pipeline parameter  
Cut-off parameters to model the target based on a BLAST target-template alignment  
Value : 0.0001  
Minimum Template size (aa) for ranking : 25  
Minimum Sequence identity : 60  
Cut-off parameters to model the target based on a HHSearch target-template alignment  
Value : 0.0001  
Probability : 50  
MAC : 0.3  
Parameters for model selection  
Minimal number of uncovered target  
residues after BLAST to run HHSEARCH : 50  
Minimal number of uncovered target  
residues to model an additional template : 25
- Finish SMR-Pipeline in automated mode on BC2-cluster at Thu May 17 13:35:44 2012

# Structural classes of proteins



Others:

Multi-domain, membrane and cell surface, small proteins, peptides and fragments, designed proteins,

..

# Databases of Protein Folds

**SCOP** (<http://scop.berkeley.edu/>) - **known** domain structure

- Structural Classification of Proteins
- Class-Fold-Superfamily-Family
- Manual assembly by inspection

**Superfamily** (<http://supfam.org/SUPERFAMILY/>) - **predicted** domain structures

- HMM models for each SCOP fold
- Fold assignments to all genome ORFs
- Assessment of specificity/sensitivity of structure prediction
- Search by sequence, genome and keywords

**CATH + Gene3D** (<http://www.biochem.ucl.ac.uk/bsm/cath/>) - **both**

- Class - Architecture - Topology - Homologous Superfamily
- Manual classification at Architecture level
- Automated topology classification using SSAP (Orengo & Taylor)

**PDB eFold** (<http://www.ebi.ac.uk/msd-srv/ssm/>)

- Fully automated using the DALI algorithm (Holm & Sander)

**Pfam** (<http://pfam.xfam.org>)- domain sequences (MSA, HMM)

# SCOP Structural Classification of Proteins (<http://scop.mrc-lmb.cam.ac.uk/scop>)



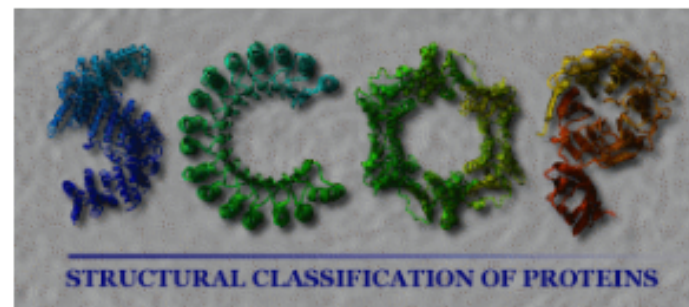
Welcome to **SCOP**: Structural Classification of Proteins.  
**1.75 release** (June 2009)

38221 PDB Entries. 1 Literature Reference. 110800 Domains. (excluding nucleic acids and theoretical models).

Folds, superfamilies, and families [statistics here](#).

[New folds](#) [superfamilies](#) [families](#).

[List of obsolete entries and their replacements](#).



**Authors.** Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. [scop@mrc-lmb.cam.ac.uk](mailto:scop@mrc-lmb.cam.ac.uk)

**Reference:** Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [\[PDF\]](#)

**Recent changes** are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [\[PDF\]](#),

Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [\[PDF\]](#), and

Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucl. Acid Res.* advance access, doi:10.1093/nar/gkm993. [\[PDF\]](#).

## Access methods

- Enter SCOP at the [top of the hierarchy](#)
- [Keyword search of SCOP entries](#)
- [SCOP parseable files](#) (MRC site)
- [All SCOP releases and reclassified entry history](#) (MRC site)
- [pre-SCOP - preview of the next release](#)
- SCOP domain sequences and pdb-style coordinate files ([ASTRAL](#))
- [Hidden Markov Model library for SCOP superfamilies \(SUPERFAMILY\)](#)



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

**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 hexokinase 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*

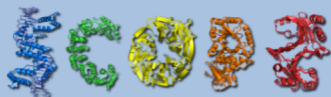
*placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.*



# Root: scop

## Classes:

1. [All alpha proteins](#) [46456] (258)
2. [All beta proteins](#) [48724] (165)
3. [Alpha and beta proteins \(a/b\)](#) [51349] (141)   
*Mainly parallel beta sheets (beta-alpha-beta units)*
4. [Alpha and beta proteins \(a+b\)](#) [53931] (334)   
*Mainly antiparallel beta sheets (segregated alpha and beta regions)*
5. [Multi-domain proteins \(alpha and beta\)](#) [56572] (53)   
*Folds consisting of two or more domains belonging to different classes*
6. [Membrane and cell surface proteins and peptides](#) [56835] (50)   
*Does not include proteins in the immune system*
7. [Small proteins](#) [56992] (85)   
*Usually dominated by metal ligand, heme, and/or disulfide bridges*
8. [Coiled coil proteins](#) [57942] (7)   
*Not a true class*
9. [Low resolution protein structures](#) [58117] (26)   
*Not a true class*
10. [Peptides](#) [58231] (120)   
*Peptides and fragments. Not a true class*
11. [Designed proteins](#) [58788] (44)   
*Experimental structures of proteins with essentially non-natural sequences. Not a true class*



## News

### November, 2013

During the development of SCOP2, we have identified a new, previously unrecognised type of alpha-alpha superhelix. Unlike other alpha-alpha superhelices..  
[More...](#)

### January, 2014

SCOP2 article in NAR is published  
[More...](#)

### January, 2014

The structure of the month  
[More...](#)

## Welcome to SCOP2!

### Citation

Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, Alexey Murzin, SCOP2 prototype: a new approach to protein structure mining (2014) Nucl. Acid Res., 42 (D1): D310-D314. [\[PDF\]](#)

### Description of the SCOP2 database

SCOP2 is a successor of Structural classification of proteins ([SCOP](#)). Similarly to SCOP, the main focus of SCOP2 is on proteins that are structurally characterized and deposited in the PDB. Proteins are organized according to their structural and evolutionary relationships, but, in contrast to SCOP, instead of a simple tree-like hierarchy these relationships form a complex network of nodes. Each node represents a relationship of a particular type and is exemplified by a region of protein structure and sequence.

In SCOP2, we try to put in use the knowledge we acquired over the past years and the lessons we have learned during the classification of protein structures. We believe that there are many peculiarities of proteins and their structures that have been missed due to the constraints of the original SCOP hierarchical schema. We hope that our users will find the new resource useful and that it could open new avenues for protein analysis and research.

### Quick introduction on how to browse, search and download

SCOP2 offers two different ways for accessing data: [SCOP2-browser](#), that allows navigation through the SCOP2 classification in a traditional way by browsing pages displaying the node information, and [SCOP2-graph](#), which is a graph-based web tool for display and navigation through the SCOP2 classification. Both tools provide search of

## Search Browser

Add an asterisk to search free text (e.g. serine\*)

## Search Graph

Add an asterisk to search free text (e.g. protein\*domain)

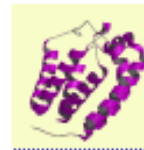
## CATH Protein Structure Classification ([http:// www.cathdb.info](http://www.cathdb.info) )

**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

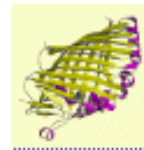
[Class \(C\)](#), [Architecture \(A\)](#) - the overall shape of the domain structure as determined by the orientations of the secondary structures but ignores the connectivity between the secondary structures., [Topology \(T\)](#) - the same overall shape and connectivity of the secondary structures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

### CATH Classification Browser

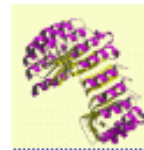
#### Main Classification Levels



[Class 1: Mainly Alpha](#)



[Class 2: Mainly Beta](#)



[Class 3: Mixed Alpha-Beta](#)



[Class 4: Few Secondary Structures](#)

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

**CATH** Home Search ▾ Browse Download About Support

16 million protein domains classified into 2,626 superfamilies

[Browse »](#) [Search »](#) [Download »](#) [Take the Tour](#)

## What's New?

The CATH website has recently undergone a big overhaul. We really hope you find the new pages more useful, easier to use and quicker to load. Please [get in touch](#) and let us know what you think.

## Searching CATH


- [Search by ID / keyword](#)
- [Search by FASTA sequence](#)
- [Search by PDB structure](#)

## Example pages

- [PDB "2bop"](#)
- [Domain "1cukA01"](#)
- [Relatives of "1cukA01"](#)
- [Superfamily "HUPs"](#)
- [Functional Family](#)
- [FunFam Alignment](#)
- [Search for "enolase"](#)
- [Superfamily Comparison](#)

## Latest News

**CATH @ VIZBI 2013**  
*March 21, 2013*



**"Evolution of Protein Architecture" - Dr Sillitoe presents the CATH resource at VIZBI (Harvard/MIT)**

## Latest Release

**CATH v3.5** based on PDB dated September

173,536	<a href="#">CATH Domains</a>
2,626	<a href="#">CATH Superfamilies</a>
51,334	<a href="#">PDBs</a>

**Gene3D v11** released March 18, 2012

1,639	<a href="#">Cellular Genomes</a>
1,016	<a href="#">Viral Genomes</a>
14,963,305	<a href="#">Protein Sequences</a>
16,297,076	<a href="#">CATH Domain Predictions</a>

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

The banner features a background of various protein structures rendered as red and green ribbons. The text is overlaid on this background. At the top left, there are four colored squares (C, A, T, H) corresponding to the letters in the title. A navigation menu is located at the top, and a search bar is on the right. The main title and a key statistic are prominently displayed in the center. Below the main title is a search input field with a green search button.

C A T H Home Search Browse Download About Support Search CATH by keywords or ID

# CATH / Gene3D v4.2

95 million protein domains classified into 6,119 superfamilies

Search by keywords, PDB code, GO term, etc Search

**Core classification files for the latest version of CATH-Plus (v4.2) are [now available to download](#). [Daily updates](#) of our very latest classifications [are also available](#).**

We are currently working on generating the [CATH-Plus](#) database for v4.2 which comprises all the extra derived data from the classification data. This includes: incorporation of the latest [Gene3D](#) sequence and functional annotation data; updating the [Functional Families \(FunFams\)](#); creating new [superfamily superpositions](#); producing [structural clusters](#) for each superfamily. We will update the web pages when this data is ready.

## **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](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](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://rosetta.bakerlab.org/>)

Swiss-Model (<http://www.expasy.org/swissmod/SWISS-MODEL.html>)