

Predikce struktury proteinů

1

Struktura proteinů

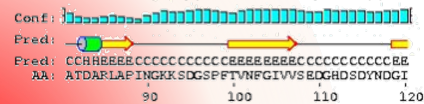
1D

ADSQTSSNRAGEFSIPPNTDFRAIFFANAAE
QQHIKLFIGDSQEPAAAYHKLTTTRDGPREATL
NSGNGKIRFEVSVNGKPSATDARLAPINGK
KSDGSPFTVNFIVVSEDDGHSDYNDGIVV
LQWPIG

primární
(sekvence)

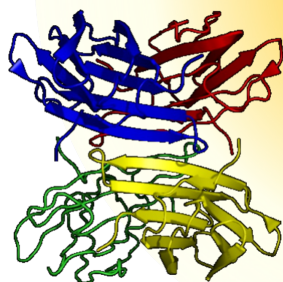


2D

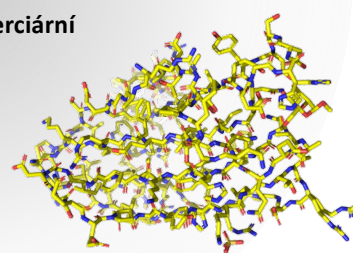


sekundární

4D



kvartérní



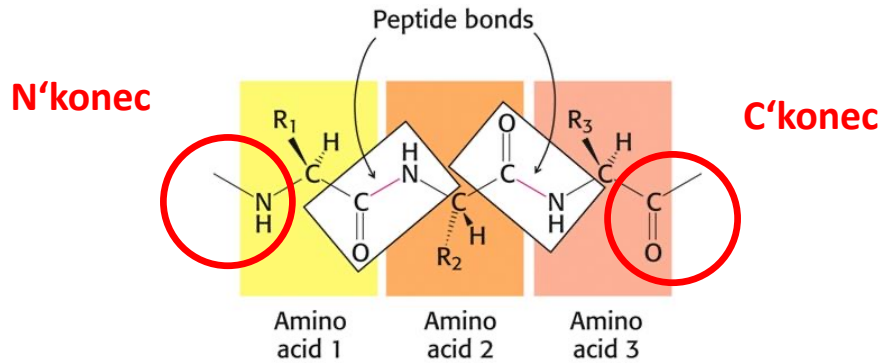
terciární

3D

2

Primární struktura

- Sekvence aminokyselin zapsaná od N' konce k C' konci



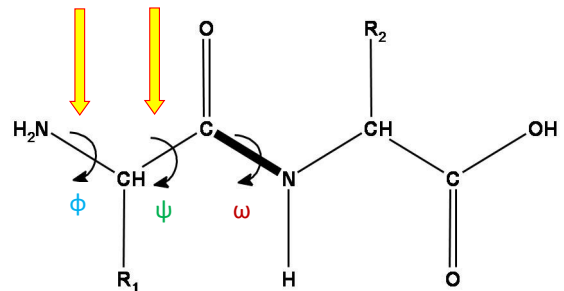
3

Sekundární struktura

2D

- Definována pomocí **torzních úhlů** peptidové páteře
- Pro každou aminokyselinu lze definovat tři úhly:
 - ϕ – úhel kolem vazby N-C α
 - ψ – úhel kolem vazby C α -C(karb.)
 - ω – úhel kolem peptidové vazby (180°, výjimečně 0°)

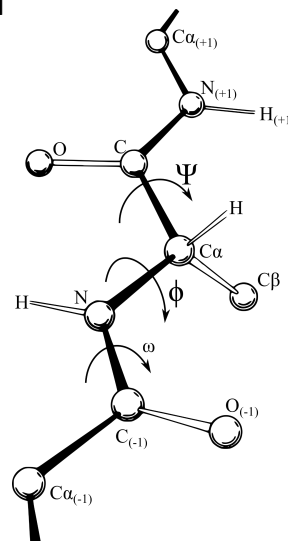
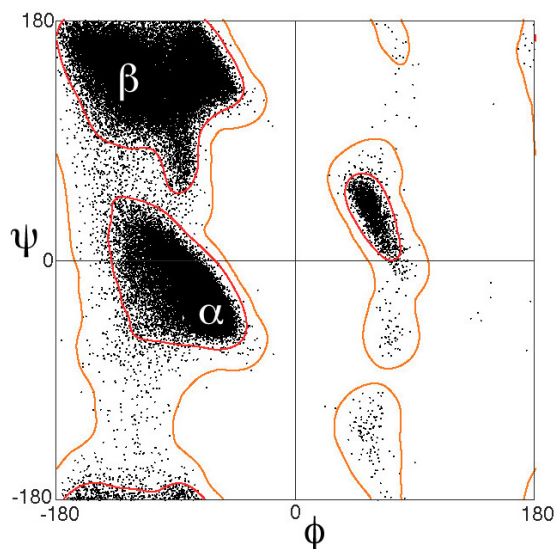
- Stabilizována pomocí vodíkových můstků mezi atomy peptidové kostry



4

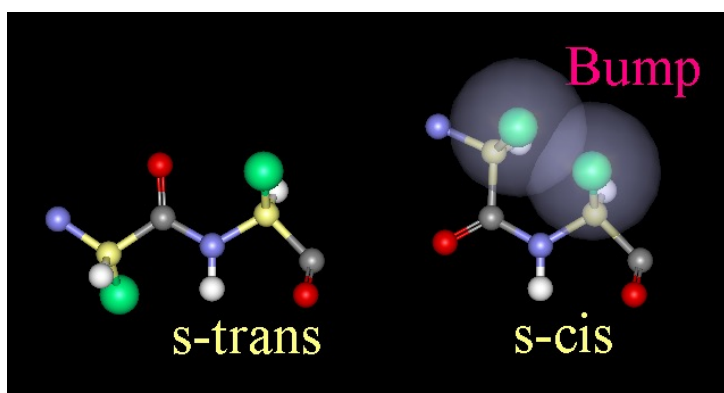
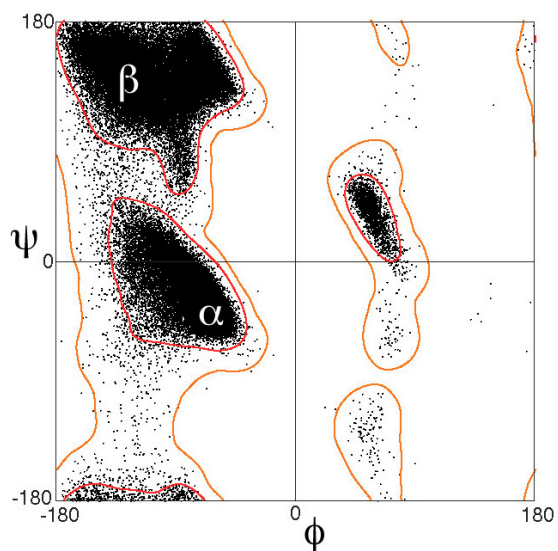
Ramachandranův diagram

➤ Každé aminokyselině odpovídá jeden bod v diagramu



Ramachandranův diagram

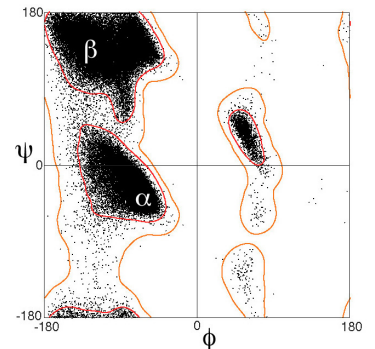
➤ Každé aminokyselině odpovídá jeden bod v diagramu



Sekundární struktura

2D

- Stabilní konformace polypeptidového řetězce
- Důležité pro udržení 3D struktury
- α -šroubovice (helix), β -skládaný list (sheet), otáčky, smyčky
- Cca 50 % aminokyselin je součástí α a β struktur

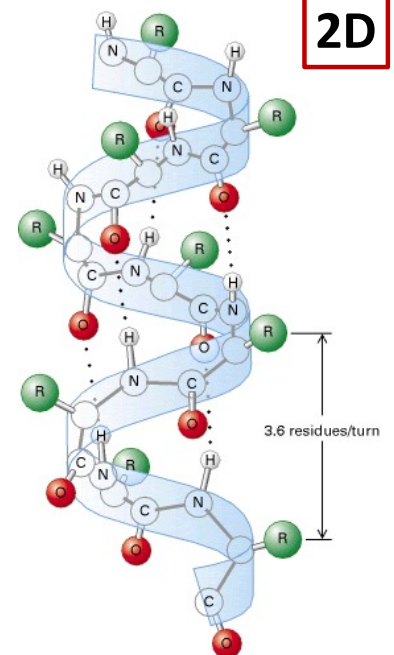


7

Šroubovice (helix)

2D

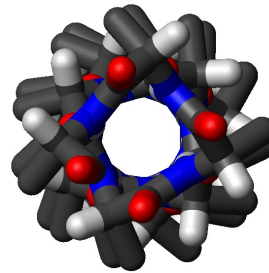
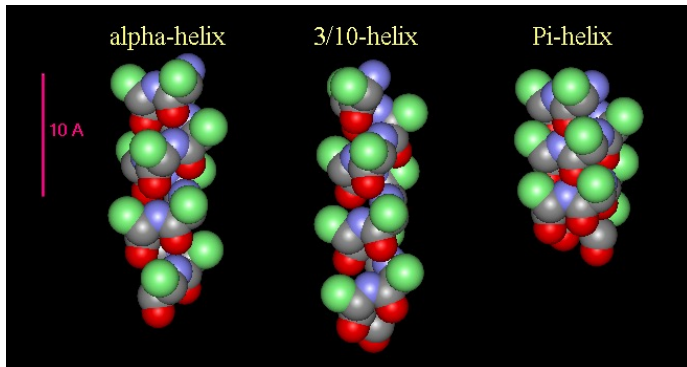
- α -helix – nejčastější
- 3_{10} -helix – obvykle na začátku nebo na konci α -helixu
- π -helix – málo stabilní, málo častý



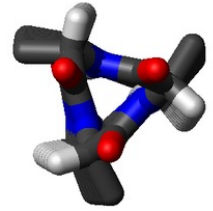
8

Šroubovice (helix)

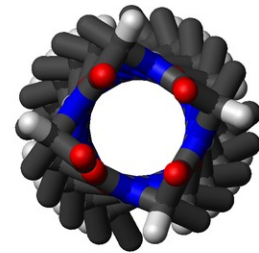
2D



α-helix



3₁₀-helix



π-helix

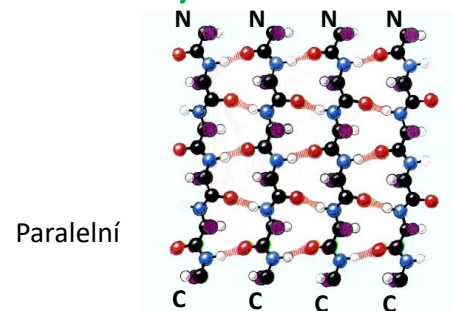
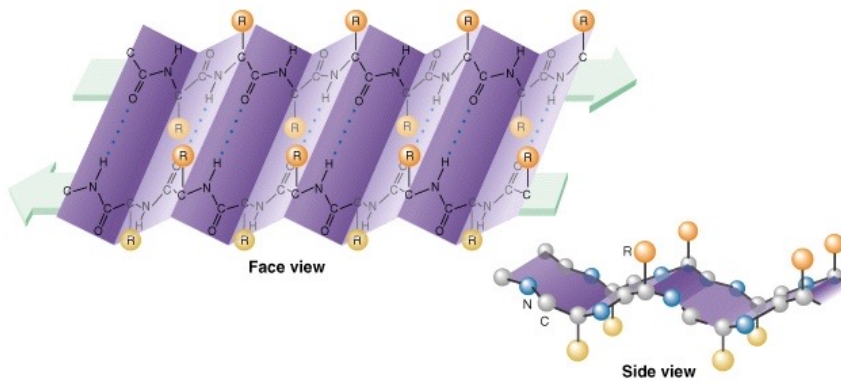
9

	α-helix	3 ₁₀ -helix	π-helix
Vodíkové můstky	$O_i \dots N_{i+4}$	$O_i \dots N_{i+3}$	$O_i \dots N_{i+5}$
Residua na otáčku	3,6	3	4,4
Vinutí (Å na 1 AK)	1,5	2	1,15

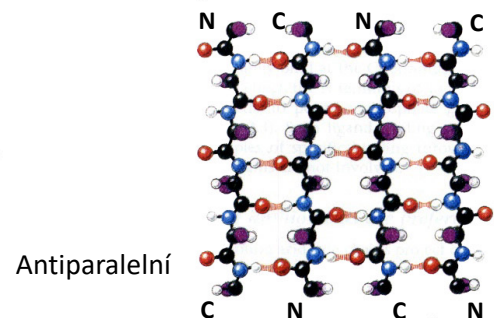
Skládaný list (extended β-sheet)

2D

➤ Paralelní, antiparalelní, mix



Paralelní



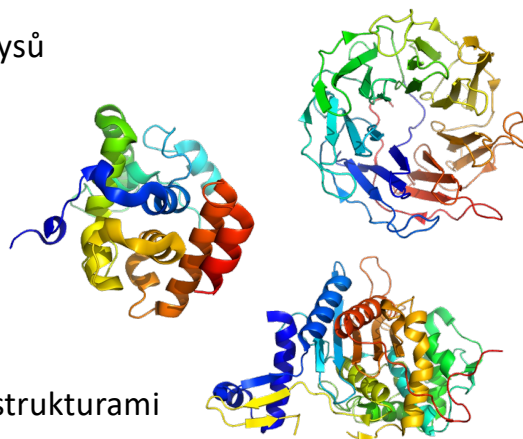
Antiparalelní

Dělení proteinů dle 2D struktury

Zejména pro účely klasifikace, hledání společných rysů

Každý protein obsahuje mj. smyčky a ohyby

- Jen α struktury
- Jen β struktury
- α/β
 - Motivy kombinující α i β struktury
- $\alpha + \beta$
 - Oddělené domény tvořené jen α nebo jen β strukturami
- **Malé proteiny**
 - Speciální případy, např. obsahující ionty kovů, stabilizované disulfidickými můstky



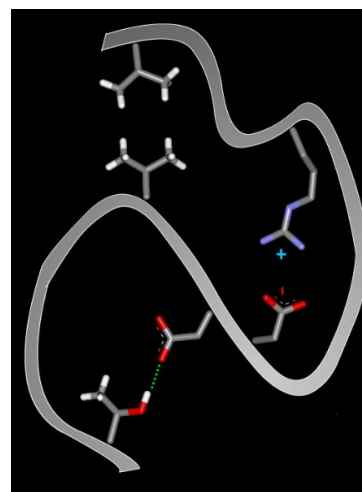
13

Terciární struktura

- Konkrétní umístění jednotlivých atomů polypeptidového řetězce v prostoru

- Stabilizována pomocí:

- **Vodíková vazba** (H-můstek)
mezi polárními AK, mezi hlavním řetězcem
- **Iontová** interakce – nabité AK
- **Hydrofobní** interakce – nepolární AK
- „**Stacking**“ (π - π , CH- π interakce) – aromatické AK
- Kovalentní vazba **síra-síra** – cystein / cystin
- Vazba **iontů kovů**



14

Od 2D ke 3D



➤ Motivy

- 2-3 prvky sekundární struktury

➤ Foldy

- Kombinace jednoduchých motivů

➤ Domény

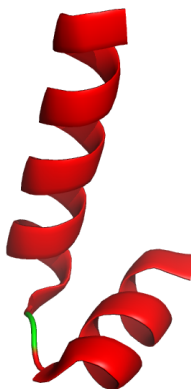
- Tvořeny motivy/foldy
- Část struktury s vlastní funkcí (nejmenší funkční jednotka)
- Nezávislá jednotka (alespoň částečně nezávislá)

17

Jednoduché motivy



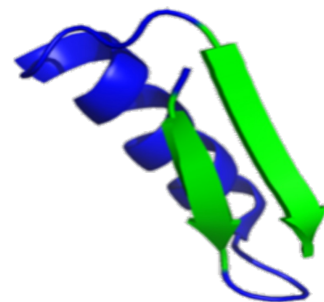
Helix-turn-helix



β -vlásenka



β - α - β

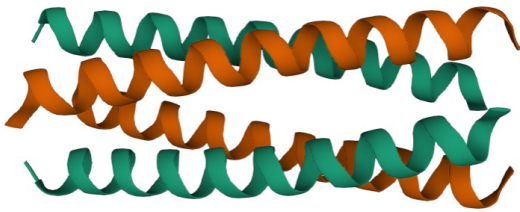


18

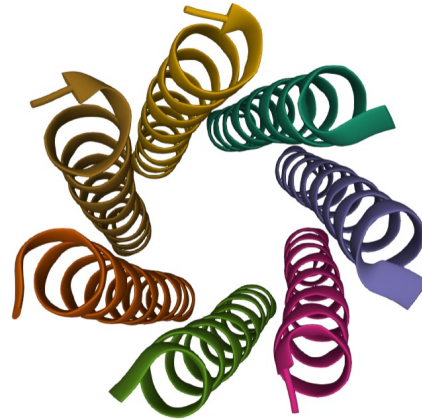
Složené α -motivy/foldy



4-helix bundle



7-helix barrel

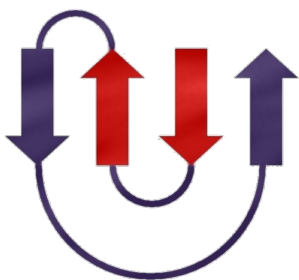


19

Složené β -motivy/foldy



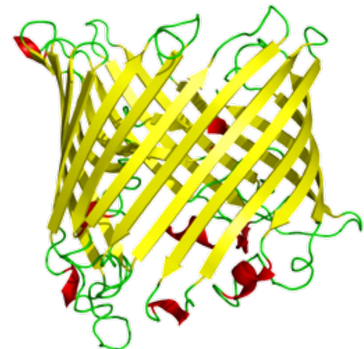
Řecký klíč



β -meandr



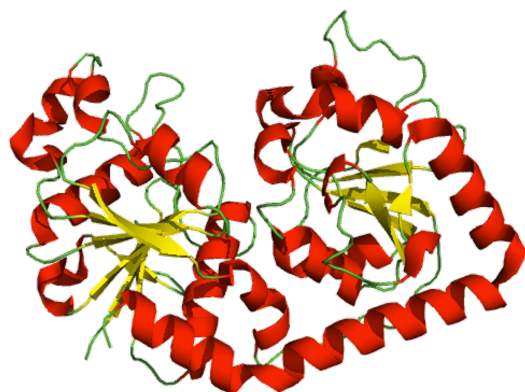
β -barel



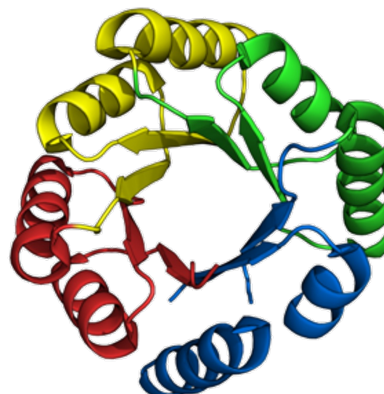
20

Složené α/β -motivy/foldy

Rossmannův fold



TIM-barel



21

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)

AlphaFold Protein Structure Database (<https://alphafold.ebi.ac.uk/>)

Structural classification of proteins (SCOP)

2D

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

The screenshot shows the SCOP 2 website interface. At the top, there is a navigation bar with 'About', 'Contact', and 'Download' links, and a search box containing 'dmr1'. Below this, a banner states: 'The legacy SCOP websites can be accessed at SCOP 1.75 and SCOP2 prototype'. The main heading is 'SCOP 2' with a 'Learn More' button. The text below describes SCOP as a database for structural classification of proteins, mentioning its creation by manual inspection and automated methods. It notes that the latest update on 2020-03-31 includes 44,218 non-redundant domains representing 532,428 protein structures. Below the text are two search options: 'Keyword and ID search' and 'Sequence search'. The 'Browse by structural class' section lists: All alpha proteins, All beta proteins, Alpha and beta proteins(a/b), Alpha and beta proteins(a+b), and Small proteins. The 'Browse by protein type' section lists: Globular proteins, Membrane proteins, Fibrous proteins, and Non-globular/Intrinsically unstructured proteins.

23

CATH – Protein structure classification database

2D

- Domény jsou klasifikovány podle CATH hierarchie
 - Třída (Class)
 - Podle sekundární struktury
 - Jen α , jen β , α i β , minimum sekundární struktury
 - Architektura
 - 3D uspořádání sekundární struktury
 - Topologie/fold
 - Jak jsou prvky sekundární struktury uspořádané za sebou
 - Homologní nadrodina
 - V případě, že jsou domény evolučně příbuzné (homologní proteiny)

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

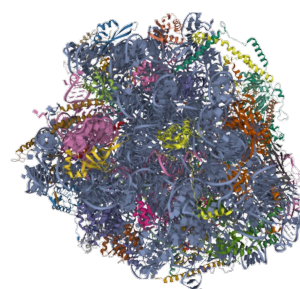
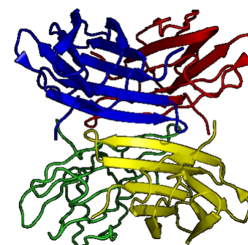
The screenshot displays the CATH database interface for the protein 4agi. It shows three levels of classification: 1. Matching CATH Superfamilies: 2.120.10.70, Fucose-specific lectin. 2. Matching CATH Domains: 4agiA00, FDB code 4agi, chain A, domain 00. 3. Matching PDB Structures: 4agi, FDB code 4agi. Each level includes a 3D ribbon diagram of the protein structure and a 'View all entries' button.

24

Kvartérní struktura



- Vzájemná kombinace více řetězců (monomerů)
- Podle typu podjednotek:
 - **Homooligomery** (identické jednotky)
 - **Heterooligomery** (alespoň dva různé typy jednotek)
- **Komplexy** proteinů s dalšími makromolekulami
 - Ribosom, proteasom, replikační komplex,...
- Nadmolekulární komplexy
 - Virové částice, buněčná membrána, organely,...



25

Predikce struktury

- Predikce struktury znamená přiřazení strukturních atributů jednotlivým aminokyselinám (2D struktura, koordináty – tvorba 3D modelu)
- Struktura 2D a 3D je konzervovaná více než samotná sekvence
- **Vstupní informace:**
 - Sekvence
 - Fyzikálně-chemické parametry
 - Informace v databázích
- **Výstup:**
 - Model struktury (2D, 3D, 4D)

33

Proč predikovat strukturu?

- **Klasifikace** proteinů
- Vytvoření modelu struktury pro další studium
- **Předpověď funkce** proteinu
 - Homologní struktury
 - Vazebná místa
- **Analýza povrchu**
 - Přístupnost solventu, tunely, kavity

34

Predikce sekundární struktury

2D

- Predikce 3 základních typů: H (helix), E (β -list), C/- (smyčka/vše ostatní)
- 1. GENERACE
 - *ab-initio*
 - Vychází z fyzikálně-chemických vlastností a ze statistik pro jednotlivé aminokyseliny

35

1. Generace – *ab initio*

Relative Amino acid Propensity Values for Secondary Structure Elements Used in the Chou-Fasman Methods

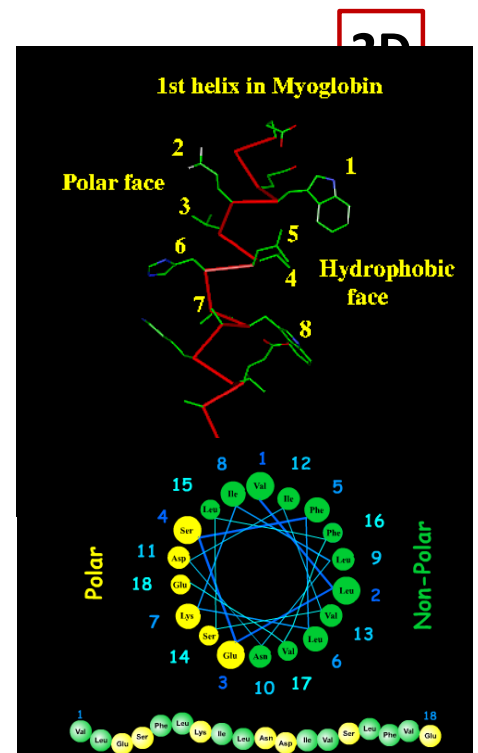
$$\frac{R_i(SS)}{R_i(SS)}$$

$$\frac{\sum R_i}{\sum R_i}$$

Amino Acid	(α -Helix)	P (β -Strand)	P (Turn)
Alanine	1.42	0.83	0.66
Arginine	0.98	0.93	0.95
Asparagine	0.67	0.89	1.56
Aspartic acid	1.01	0.54	1.46
Cysteine	0.70	1.19	1.19
Glutamic acid	1.51	0.37	0.74
Glutamine	1.11	1.11	0.98
Glycine	0.57	0.75	1.56
Histidine	1.00	0.87	0.95
Isoleucine	1.08	1.60	0.47
Leucine	1.21	1.30	0.59
Lysine	1.14	0.74	1.01
Methionine	1.45	1.05	0.60
Phenylalanine	1.13	1.38	0.60
Proline	0.57	0.55	1.52
Serine	0.77	0.75	1.43
Threonine	0.83	1.19	0.96
Tryptophan	0.83	1.19	0.96
Tyrosine	0.69	1.47	1.14
Valine	1.06	1.70	0.50

Typické znaky α -šroubovice

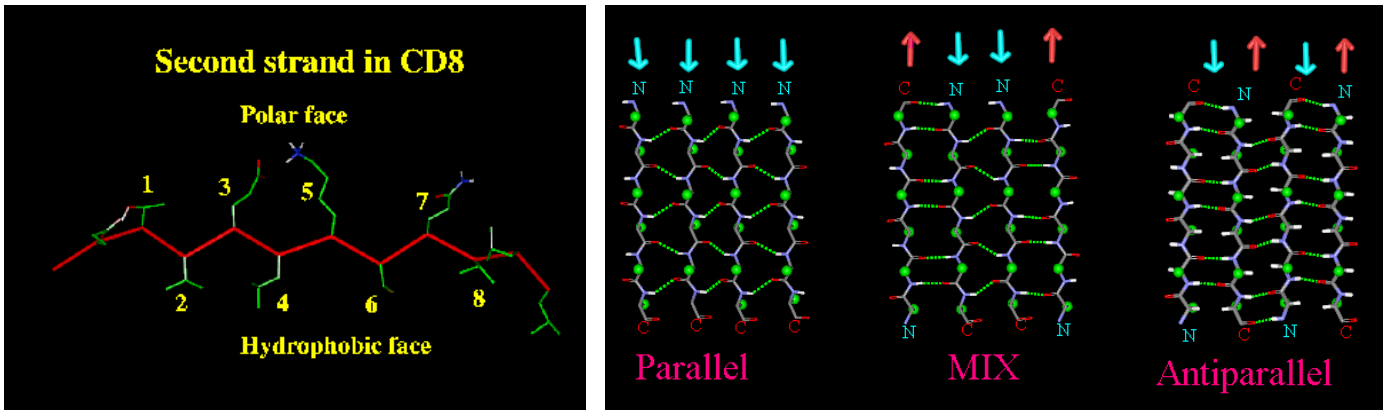
- Často je částečně exponovaná
 - Jedna strana je otočená dovnitř proteinu (hydrofobní) a druhá ven (hydrofilní)
 - Residuum (aminokyselina) n , $n+3$, $n+4$, $n+7$ míří na stejnou stranu
- Transmembránový helix
 - Všechny aminokyseliny hydrofobní



Typické znaky β -list (musí být stabilizován jinou částí polypeptidového řetězce!)

U β -listu se střídají boční řetězce po 180°

pro částečně zanořený β -list platí, že každé liché reziduum je polární, každé sudé nepolární, u plně zanořeného jsou všechna nepolární... tj. residua směřující na stejnou stranu by měla mít stejný charakter



α -šroubovice nebo β -list?

2D

ELKAHIRVDLTQ

α

ELKAHIRVDLTQ

ELKAHIRVDLTQ

β

Polární

Nepolární

2D

α-šroubovice nebo β-list?

ELKAHIRVDLTQ

α ✗

ELKAHIRVDLTQ

ELKAHIRVDLTQ

β ✓

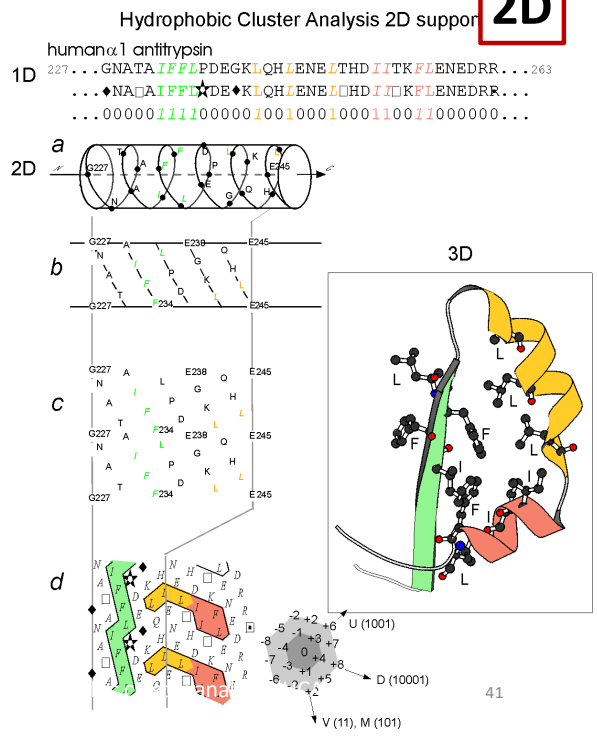
Polární

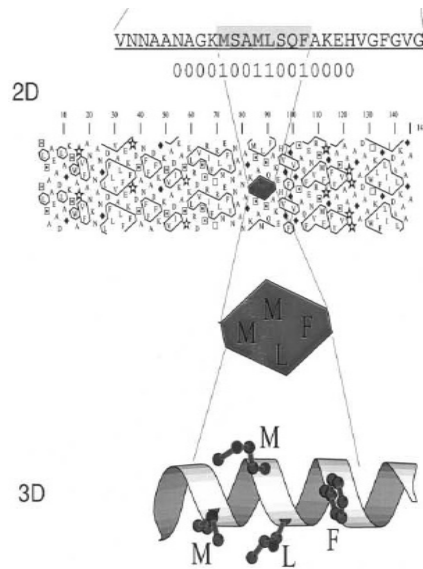
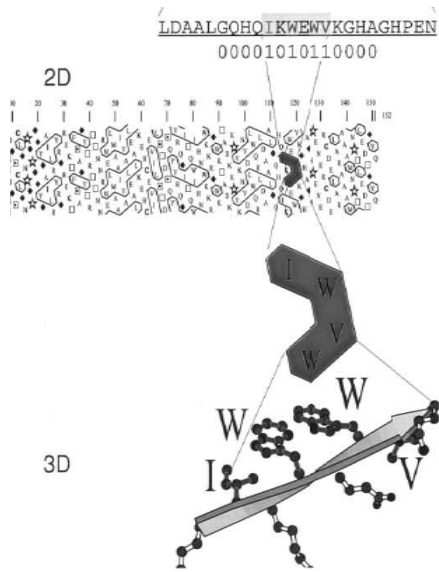
Nepolární

Analýza hydrofobních klastrů (HCA)

2D

- Sekvence „se namotá“ na válec (α-helix)
- HCA graf je zobrazení válce v rovině
- Hydrofobní aminokyseliny jsou ohraničeny a tvoří specifické tvary pro α-helixy a β-listy





RPBS Web Portal – HCA

2D

<https://mobyte.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py?form=HCA#forms::HCA>

RPBS Web Portal

(guest)
set email sign-in sign-out
refresh workspace

Search [more] Welcome Forms Data Bookmarks Jobs Tutorials

HCA x

HCA 1.0.2
Hydrophobic Cluster Analysis.

Run Reset Help pages

Input Data

query.data.seq Drawn by Luc Canard

10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310

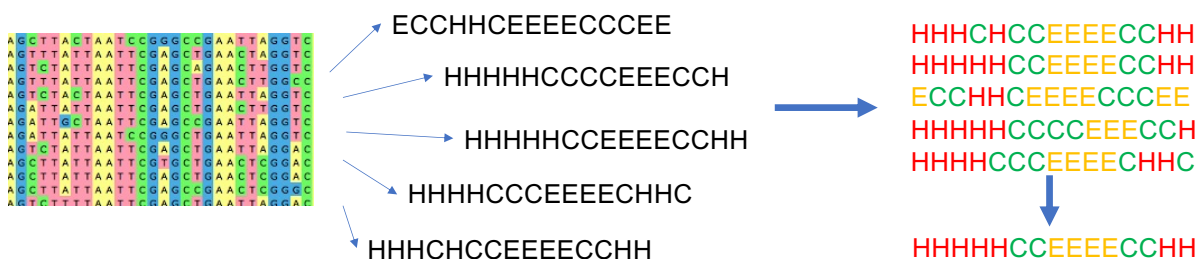
BCSearch FAF-Drugs4 fpocket Frog2 HAlign-Kbest InterEvDock2 MTAutoDock/MTIOpenScreen PatchSearch

Predikce sekundární struktury

- Predikce 3 základních typů: H (helix), E (β-list), C/- (smyčka/vše ostatní)
- 1. GENERACE
 - *ab-initio*
 - Vycházela z fyzikálně-chemických vlastností a ze statistik pro jednotlivé aminokyseliny
- 2. GENERACE
 - Zahrnuje i vliv okolních aminokyselin
- 3. GENERACE
 - *Homology-based models*
 - Metody strojového učení
 - Využívá multiple sequence alignmentu a toho, že 2D struktura je více konzervovaná než sekvence

Metody založené na homologii (*Homology-based*)

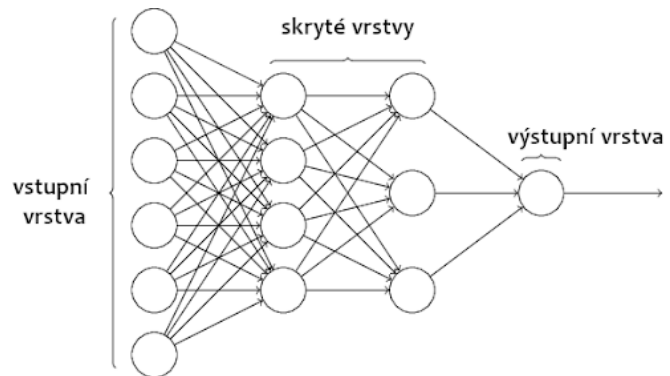
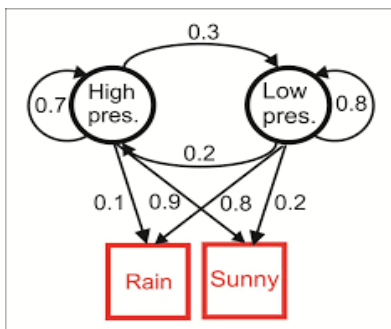
- Vychází z předpokladu, že 2D struktura je více konzervovaná než sekvence
- 1. Multiple sequence alignment
- 2. Predikce sekundárních struktur pro každou sekvenci zvlášť
- 3. Porovnání predikovaných sekundárních struktur s alignmentem
- 4. Konsenzus sekundární struktury



Metody strojového učení (*Machine learning*)

2D

- Model, který je natrénovaný na známé sadě dat
- Neuronové sítě
- Skryté Markovovy modely



46

PSIPRED

2D

- Predikce sekundární struktury pomocí 2 neuronových sítí
- Časově náročnější
- Ve srovnání s většinou programů na predikci sekundární struktury má lepší výsledky

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

Choose prediction methods

Popular Analyses

- PSIPRED 4.0 (Predict Secondary Structure)
- MEMSAT-SVM (Membrane Helix Prediction)
- DISOPRED3 (Disopred Prediction)
- pGenTHREADER (Profile Based Fold Recognition)

Contact Analysis

- DeepMetaPSPICOV 1.0 (Structural Contact Prediction)
- MEMPACK (TM Topology and Helix Packing)

Fold Recognition

- GenTHREADER (Rapid Fold Recognition)
- pDomTHREADER (Protein Domain Fold Recognition)

Structure Modelling

- Bioserf 2.0 (Automated Homology Modelling)
- DMPfold 1.0 Fast Mode (Protein Structure Prediction)
- Domserf 2.1 (Automated Domain Homology Modelling)

Domain Prediction

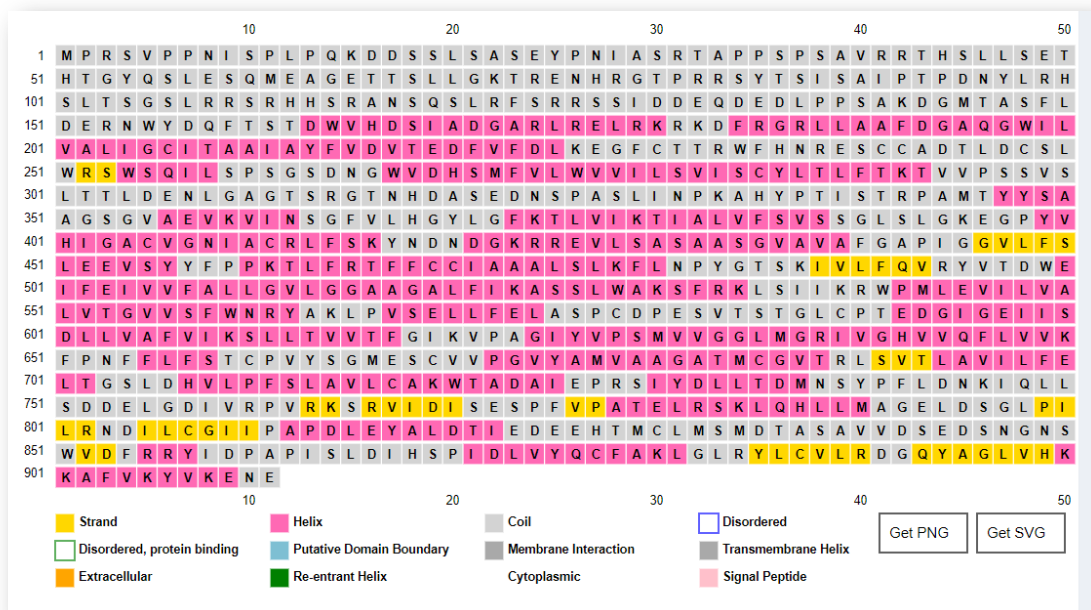
- DomPred (Protein Domain Prediction)

Function Prediction

- FFPred 3 (Eukaryotic Function Prediction)

PSIPRED

2D



48

Rozšíření predikce 2D struktury

2D

➤ Predikce **více typů** 2D struktury (dle DSSP – Database of Secondary Structure Assignments)

- α -helix (H)
- 3_{10} -helix (G)
- π -helix (I)
- β -řetězec, extended strand (E)
- β -bridge (B)
- turn (T)
- bend (S)
- ostatní, coil (C)

➤ Predikce **přístupnosti solventu**

➤ Predikce **transmembránových helixů**

49

Predikce terciární struktury

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

- **Ab initio**
- **Homologní modelování**
- **Threading** („navlékání“)



Predikce terciární struktury

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

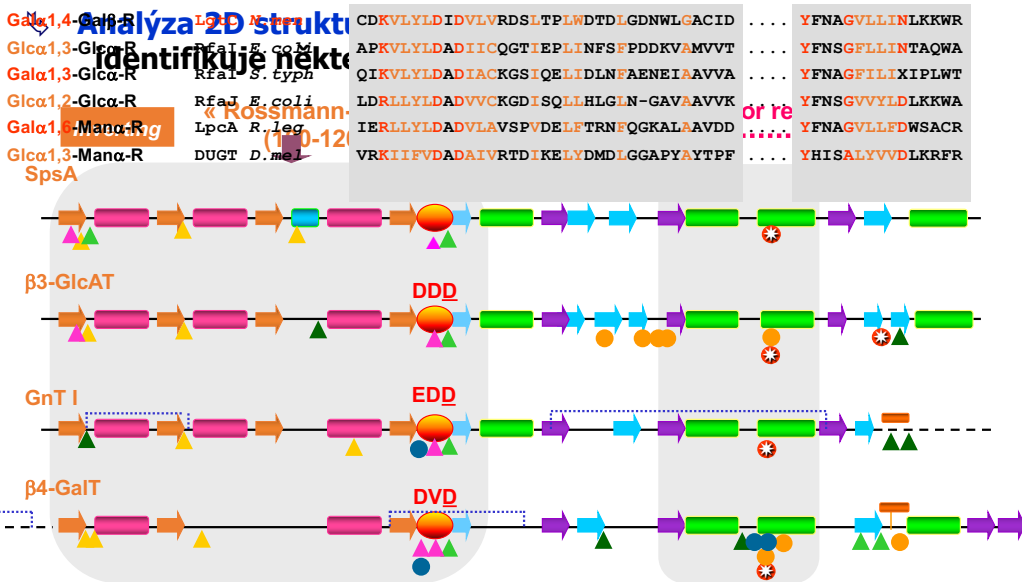
- **Ab initio**
- **Homologní modelování**
- **Threading** („navlékání“)



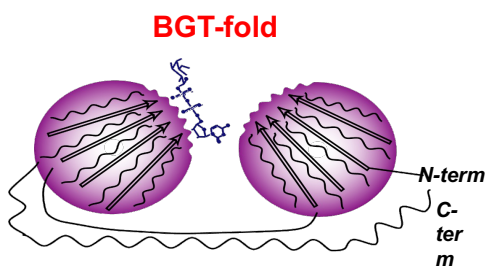
Před Alfafoldem ...

Metody pro predikci funkce

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

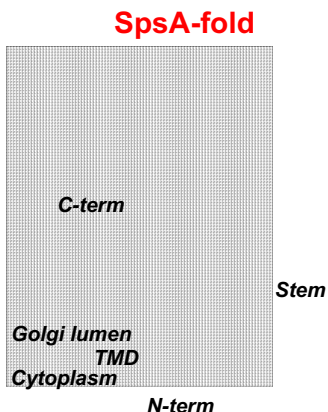


Dvě pozorované topologie 3D struktur glykosyltransferas



(Prokaryotes/Phage)

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



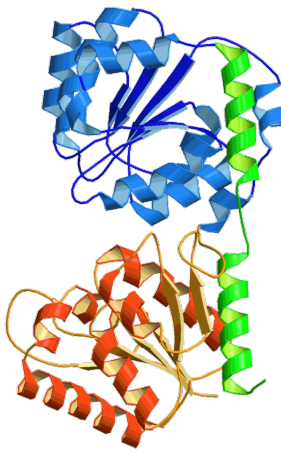
(Prokaryotes)

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

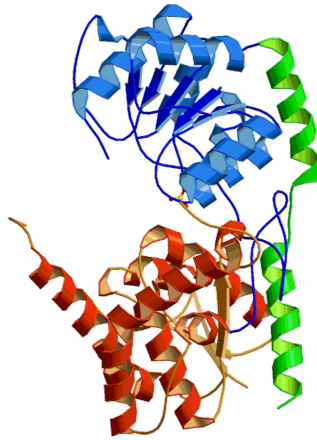
(Eucaryotes)

β 4-GalT1 (bovine)	GT7	inv
β 2-GlcNAcT (GnT I, rabbit)	GT13	inv
β 3-GlcAT I (human)	GT43	inv
α 3-GalT (bovine)	GT6	ret
Glycogenin (rabbit)	GT8	ret
α 3-GalNAcT (GTA, human)	GT6	ret
α 3-GalT (GTB, human)	GT6	ret

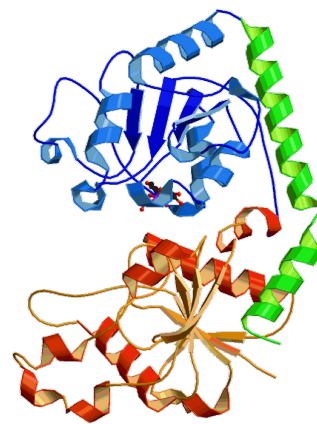
Nadrodina s BGT foldem



MurG (β-GlcNAcT)
GT28
E. coli
 Ha *et al.*, 2000



GtfB (β-GlcT)
GT1
A. orientalis
 Mulichak *et al.*, 2001

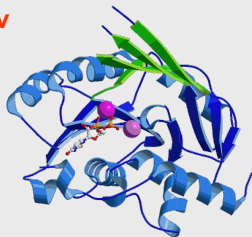


BGT (β-GlcT)
n.c.
 Phage T4
 Vrieling *et al.*, 1994

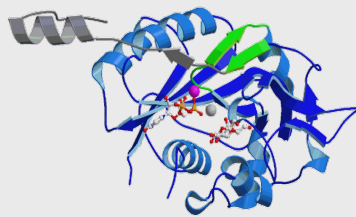
Nadrodina s SpsA foldem

Společná NBD

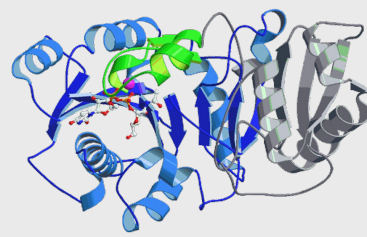
Inv



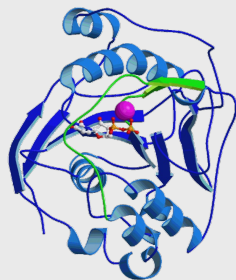
SpsA [GT2]
 Charnok *et al.*, 1999, 2001



Hum β3-GlcAT [GT43]
 Pedersen *et al.*, 2000

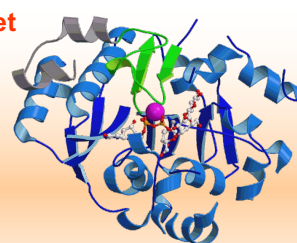


Rabbit GnT I [GT13]
 Ünligil *et al.*, 2000

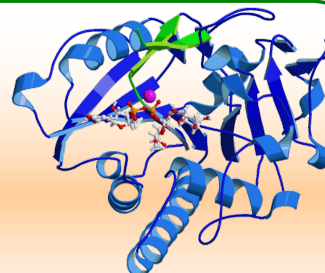


Bovine β4-GalT [GT7]
 Gastinel *et al.*, 1999
 Ramakrishnan *et al.*, 2001, 2002

Ret



LgtC (α4-GalT) [GT8]
Neisseria meningitidis
 Persson *et al.*, 2001



Bovine α3-GalT [GT6]
 Gastinel *et al.*, 2001
 Boix *et al.*, 2001, 2002

Predikce terciární struktury

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

- **Ab initio**
- **Homologní modelování**
- **Threading („navlékání“)**



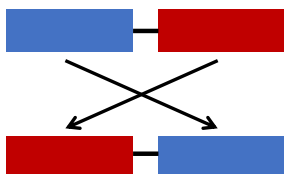
Threading



- Porovnává možnost přiložení sekvence na proteiny známých **foldů**
- „navlékání“ = rozpoznání a přiřazení proteinového foldu aminokyselinové sekvenci
- S využitím strukturních databází (PDB, SCOP, CATH) je vytvořena databáze existujících foldů - sekvence je porovnávána s touto databází (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**

PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLVILNVPTPYATGNNFPGIYFAIATNQGVDGCGFTYSSKV
 PESTGRMPFTLVATIDVGSVTFVKQWKSVRGSAMHIDSYASLSAIWGTAAAPSSQSGNQGAETGGTGAGNIG
 GGGERDGTFLNPPHIKFGVTALHAANDQTIDIYIDDDPKPAATFKGAGAQQNLGKVLDSGNGRVRVIVMANGR
 PSRLGSRQVDIFKKSYPFGIIGSEDGADDDYNDGIVFLNWPLG

ERDGTFLNPPHIKFGVTALHAANDQTIDIYIDDDPKPAATFKGAGAQQNLGKVLDSGNGRVRVIVMANGRPSR
 LGSRQVDIFKKSYPFGIIGSEDGADDDYNDGIVFLNWPLGPLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLV
 IILNVPTPYATGNNFPGIYFAIATNQGVDGCGFTYSSKVPESTGRMPFTLVATIDVGSVTFVKQWKSVRGSAM
 HIDSYASLSAIWGTAAAPSSQSGNQGAETGGTGAGNIGGGGKLAALAEIKRASQPELAPEDPEDVEHHHHHHH



```

#
#=====
EMBOSS_001      1  ----- 0
EMBOSS_001      1  ERDGTFLNPPHIKFGVTALHAANDQTIDIYIDDDPKPAATFKGAGAQQ  50
EMBOSS_001      1  ----- 0
EMBOSS_001     51  NLGTRVLDGNGRVRVIVMANGRPSRLGSRQVDIFKKSYPFGIIGSEDGAD  100
EMBOSS_001      1  -----PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRD  35
EMBOSS_001     101  DDYNDGIVFLNWPLGPLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRD  150
EMBOSS_001     36  GRLQVILNVPTPYATGNNFPGIYFAIATNQGVDGCGFTYSSKVPESTGR  85
EMBOSS_001    151  GRLQVILNVPTPYATGNNFPGIYFAIATNQGVDGCGFTYSSKVPESTGR  200
EMBOSS_001     86  MPFTLVATIDVGSVTFVKQWKSVRGSAMHIDSYASLSAIWGTAAAPSSQ  135
EMBOSS_001    201  MPFTLVATIDVGSVTFVKQWKSVRGSAMHIDSYASLSAIWGTAAAPSSQ  250
EMBOSS_001    136  GSGNQGAETGGTGAGNIGGGGERDGTFLNPPHIKFGVTALHAANDQTID  185
EMBOSS_001    251  GSGNQGAETGGTGAGNIGGGG-----  271
EMBOSS_001    186  IYIDDDPKPAATFKGAGAQQNLGKVLDSGNGRVRVIVMANGRPSRLGS  235
EMBOSS_001    272  -----KLAALAEIKRASQPELAPEDPEDVEHHHHHHH  283
EMBOSS_001    236  RQVDIFKKSYPFGIIGSEDGADDDYNDGIVFLNWPLG  271
EMBOSS_001    284  -QFE-----LAPEDPEDVEHHH-----HHH  302
  
```

Threading

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.

Phyre2



- Server pro 3D predikci struktur pomocí **threadingu**
- Vysoce výkonný – poměrně spolehlivá detekce foldu i při **nízké homologii** (i pod 15%)

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

Phyre2

#	Template	Alignment Coverage	3D Model	Confidence	% I.D.	Template Information
1	c1i1z8			100.0	31	PDB header: sugar binding protein Chain: B: PDB Molecule: ergic-53 protein; PDBTitle: the crystal structure of the carbohydrate recognition2 domain of the glycoprotein sorting receptor p58/ergic-53 reveals a novel metal binding site and conformational4 changes associated with calcium ion binding
2	c2a6yA			100.0	21	PDB header: sugar binding protein Chain: A: PDB Molecule: emp47p (form1); PDBTitle: crystal structure of emp47p carbohydrate recognition domain2 (crd), tetragonal crystal form
3	d2a6za1			100.0	20	Fold: Concanavalin A-like lectins/glucanases Superfamily: Concanavalin A-like lectins/glucanases Family: Lectin leg-like
4	c2dug8			100.0	42	PDB header: protein transport Chain: B: PDB Molecule: vesicular integral-membrane protein vip36; PDBTitle: crystal structure of vip36 exoplasmic/luminal domain, metal-free form

Phyre2

ARDLVIPMIYCGHG



Homologous sequences

User sequence

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

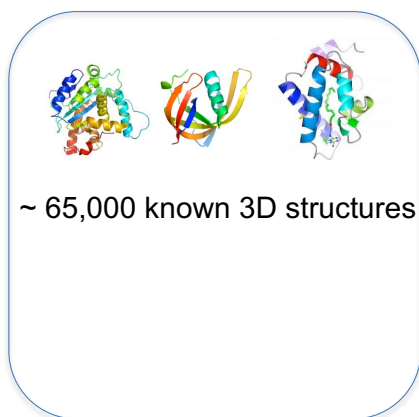
Phyre2



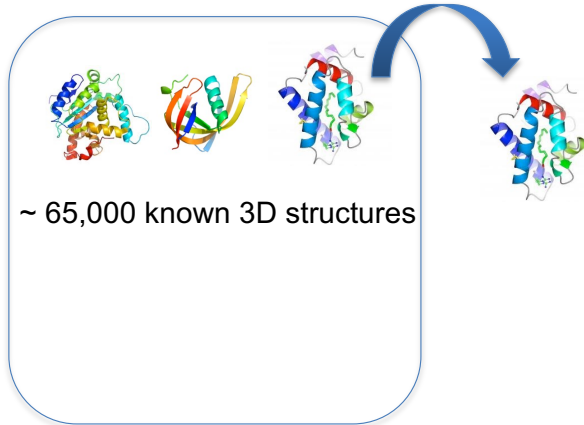
Capture the mutational propensities at each position in the protein

An evolutionary fingerprint

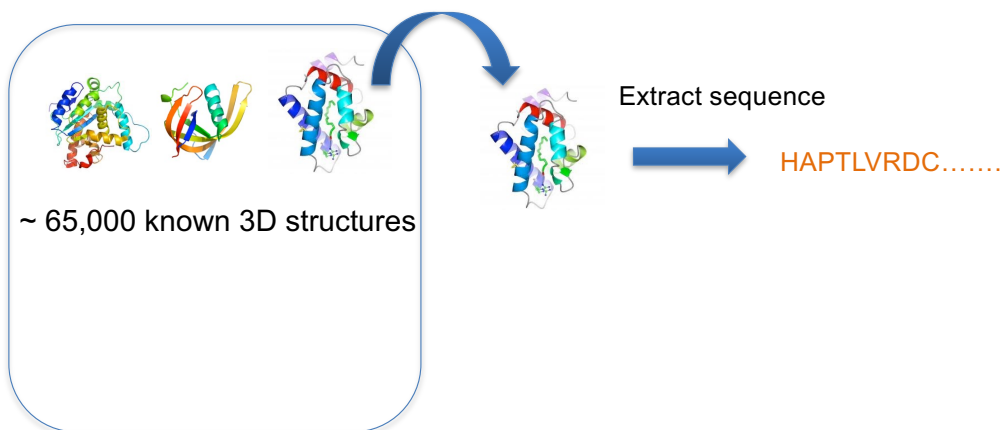
Phyre2



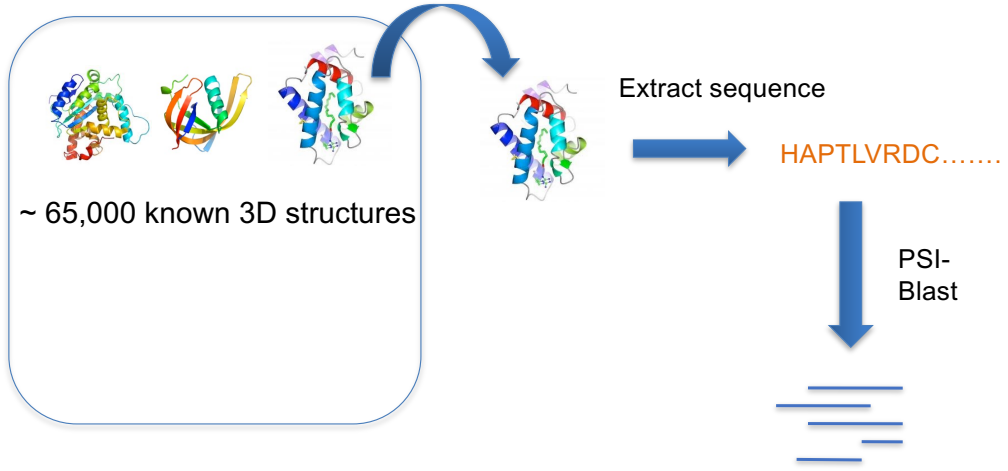
Phyre2



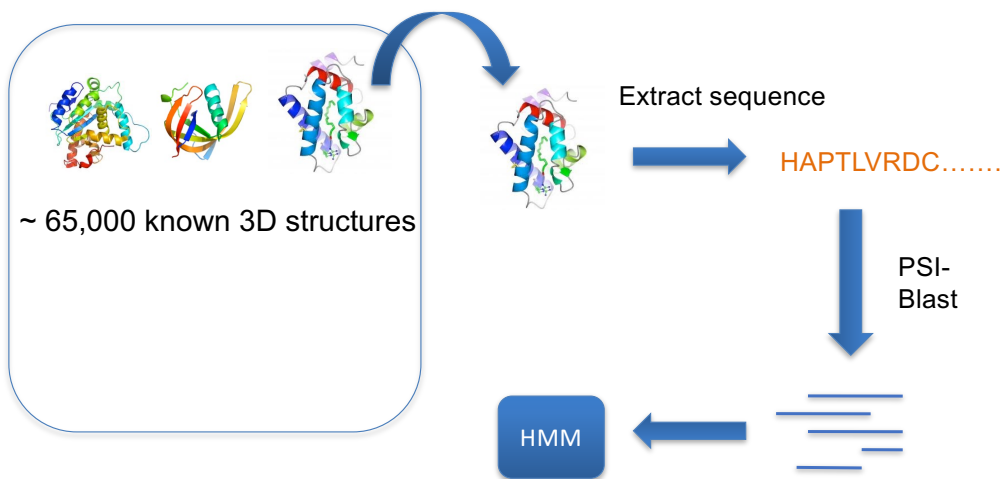
Phyre2



Phyre2

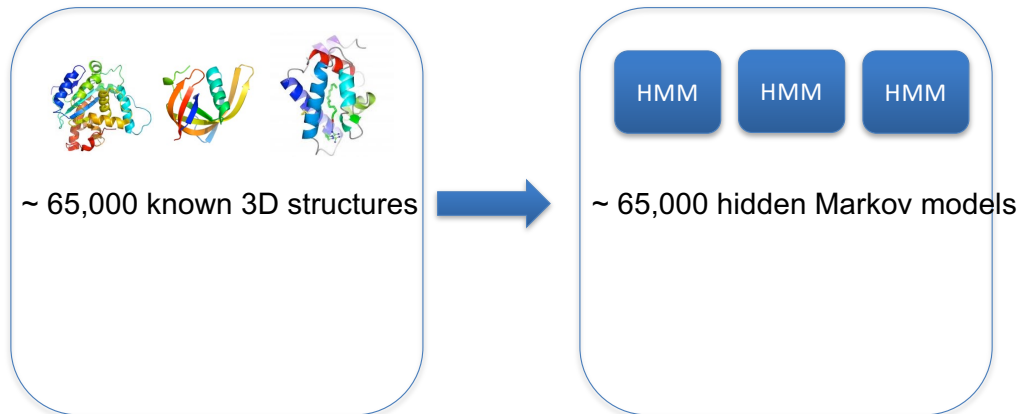


Phyre2

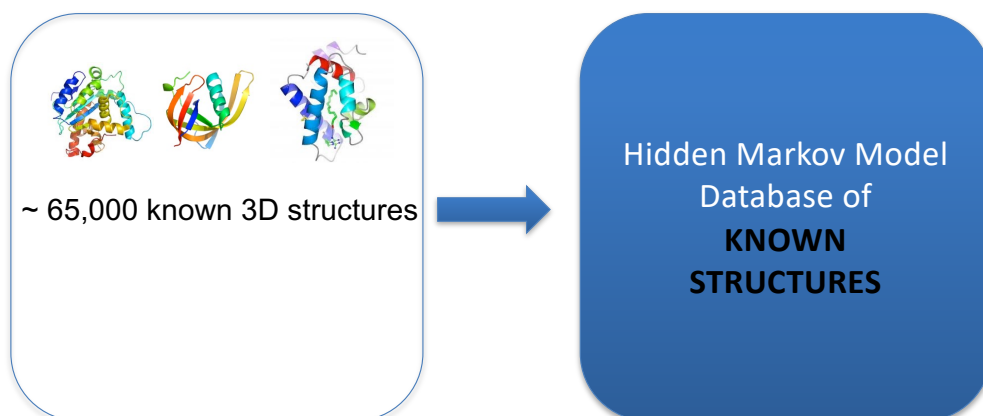


Hidden Markov model
for sequence of KNOWN structure

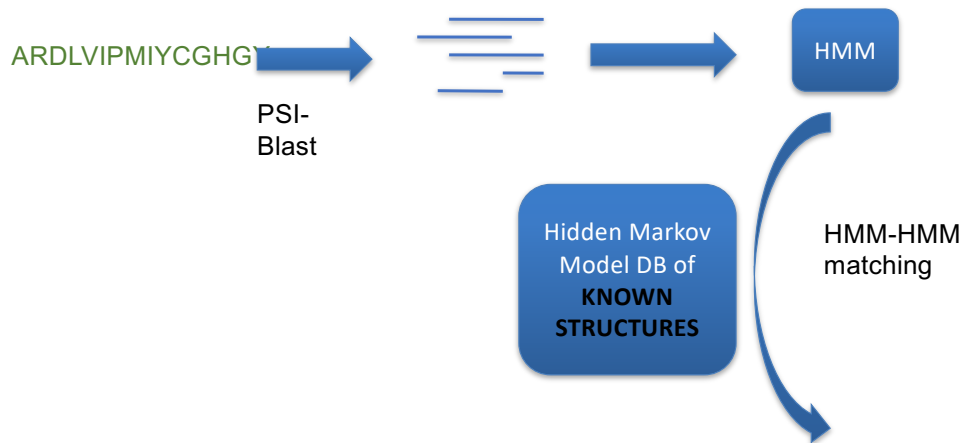
Phyre2



Phyre2



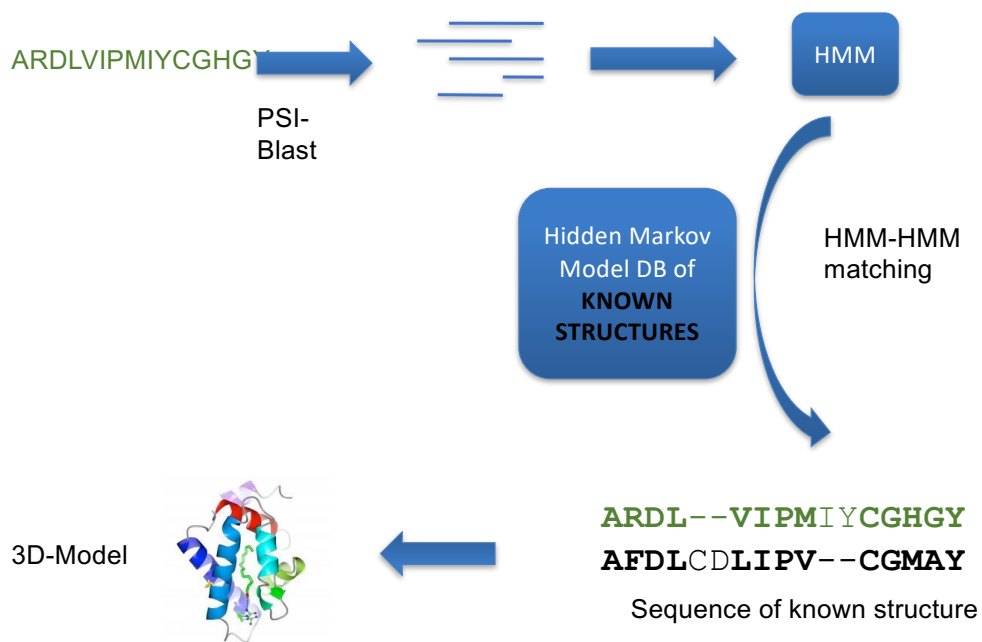
Phyre2



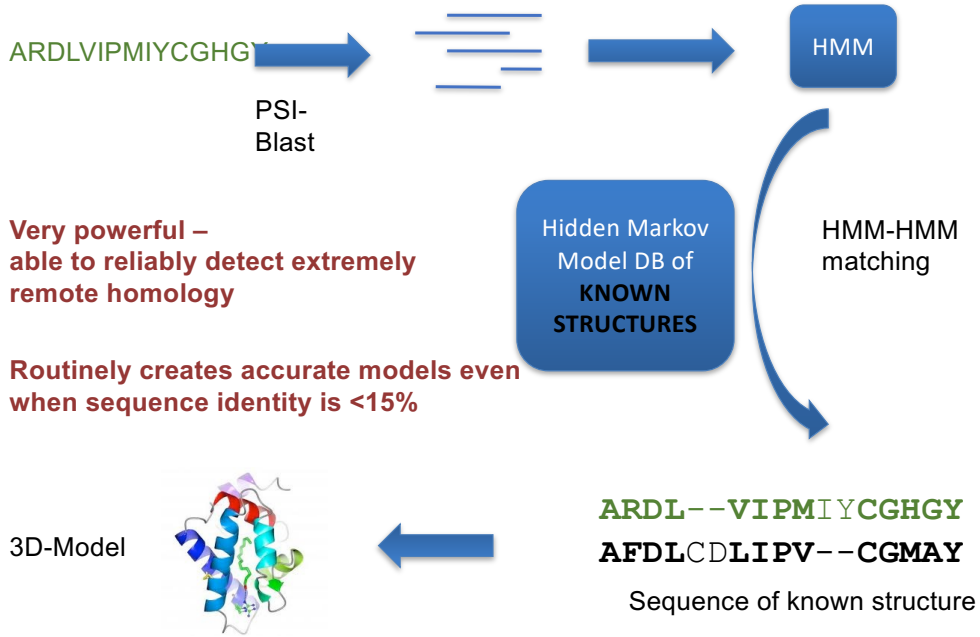
Alignments of user sequence to known structures ranked by confidence.

ARDL--VIPMIYCGHG
AFDLCDLIPV--CGMAY
Sequence of known structure

Phyre2

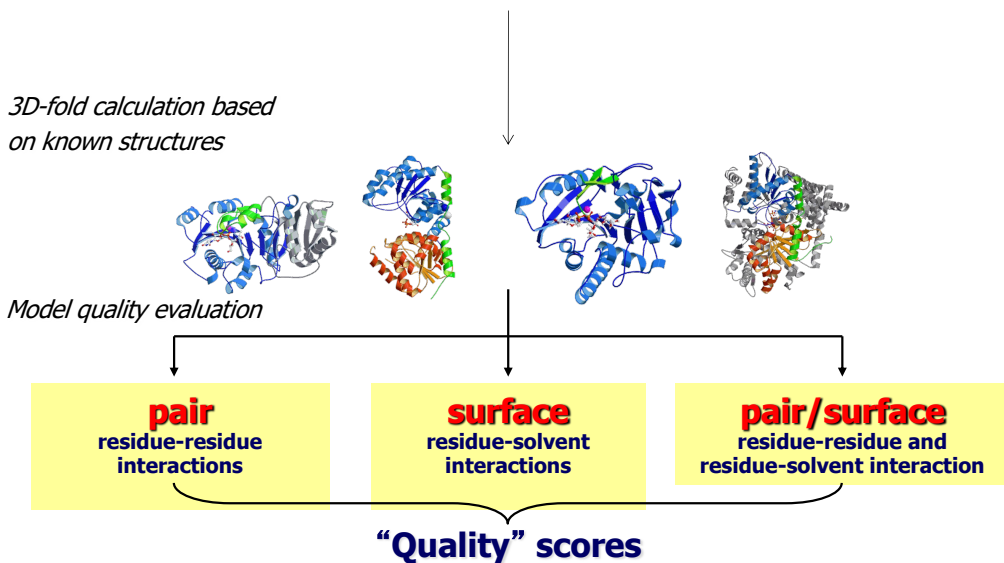


Phyre2



Fold library last updated: 20 Apr 2019 | UNIREF50 protein sequence database updated: 7 Feb 2017 | SCOP version 1.75

SDVDIEAGQTLVQVVISNGETWVAIQLPAQYRSFDLVFENVSPSTSGSVLVAQMAPQSGGVYGSNYS
 GSGWGNDLGGGGFYGYSEAKWMCLWPANRSGPNSKTGIYGTCKLMNLNQSNVPSVTSNLFAPTAY
 KNEPGYANVGGCCQKIRGLASSIQFALHGGNVPQNTDTFSGGTIKVYGWN



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

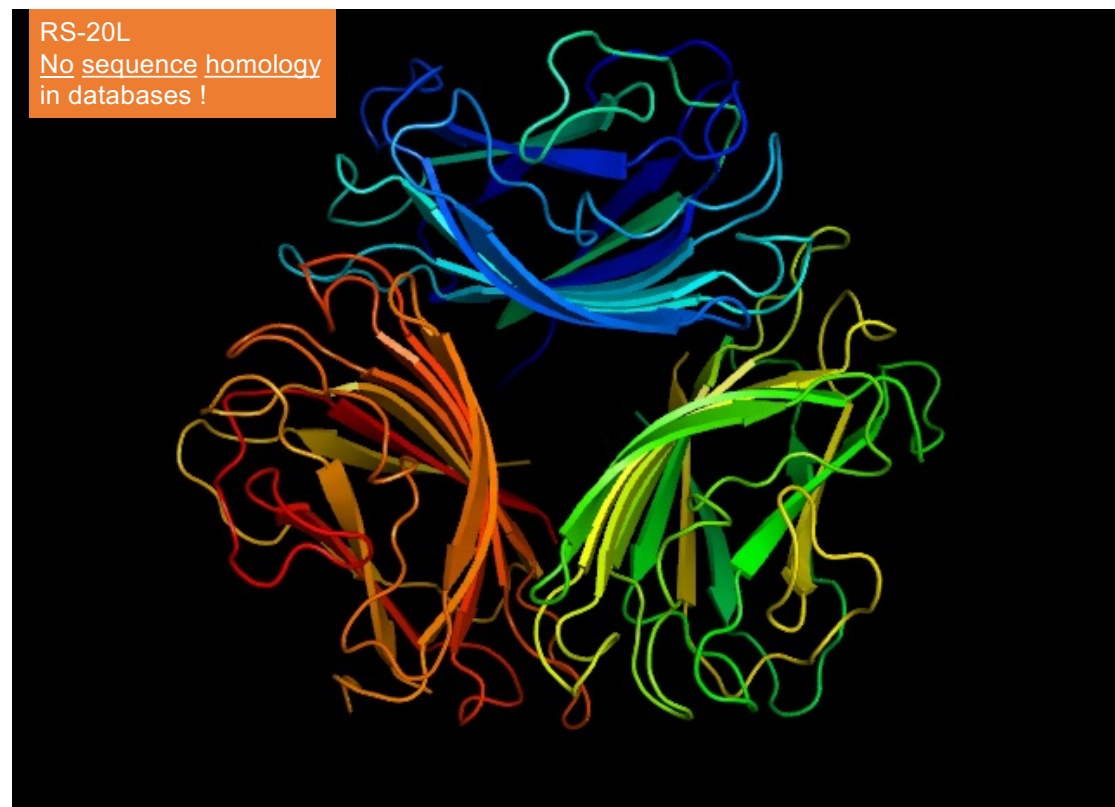
Quickphyre results for job synt... - Mozilla Firefox

http://www.sbg.bio.ic.ac.uk/phyre/qphyre_output/95cbaa7600a9bfff/summary.html

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

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

A co protein, který nemá v sekvenčních databázích žádný homolog



Quickphyre results for job rs20 - Mozilla Firefox

Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda

http://www.sbg.bio.ic.ac.uk/phyre/qphyre_output/964f0704319f5953/summary.html

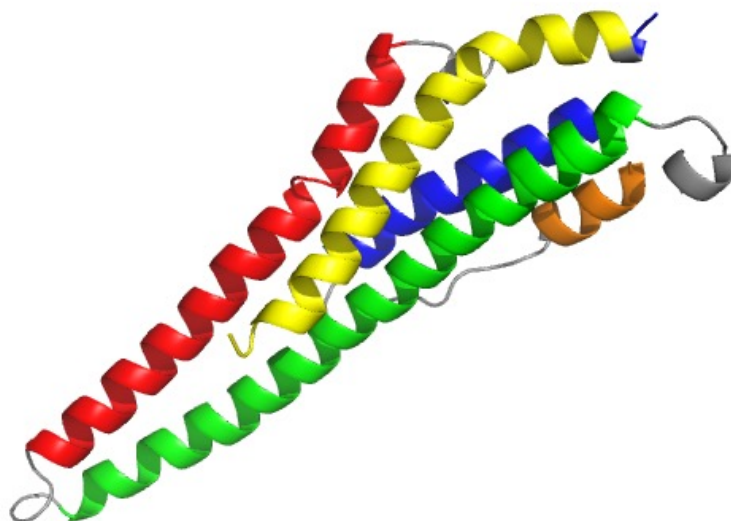
Nejnavštěvovanější Jak začít Přehled zpráv http://www.ncbi.nlm... http://www.glycoscie... CHMI Radar Departme...

Fold Recognition

View Alignments	SCOP Code	View Model	E-value	Estimated Precision	Bio Text	Fold/PDB descriptor	Superfamily	Family	(beta-test)
	d1eh9a2 (length:67) 24% i.d.	 	50	0 %	n/a	Glycosyl hydrolase domain	Glycosyl hydrolase domain	alpha-Amylases, C-terminal beta-sheet domain	n/a
	c2fsdA (length:142) 19% i.d.	 	50	0 %	n/a	PDB header: virus/viral protein	Chain: A: PDB Molecule: putative baseplate protein;	PDBTitle: a common fold for the receptor binding domains of 2 lactococcal phages? the crystal structure of the head3 domain of phage bil170	n/a
	c2ct4A (length:70) 11% i.d.	 	56	0 %	n/a	PDB header: signaling protein	Chain: A: PDB Molecule: cdc42-interacting protein 4;	PDBTitle: solution structure of the sh3 domain of the cdc42-2 interacting protein 4	n/a

AB2L structure overview

Structure: 4 helical bundle



Prozkoumání možností a principů fungování I-TASSERu bude domácím úkolem

Homologní modelování



- Je založeno na existenci blízkého **strukturního homology** (typicky 50 % sekvenční podobnosti a více, minimálně 30%)
- Využívá skutečnosti, že dva proteiny ze stejné rodiny a s podobnou sekvencí mají i podobnou 3D strukturu
- Kromě sekvence našeho proteinu potřebujeme znát strukturu homologního proteinu = **templát**
- Pro vysoce homologní sekvence je spolehlivost velmi vysoká

MODELLER

Mostly used program in academic environment for serious homology modeling

SWISS-MODEL

An automated knowledge-based protein modelling server



Homologní modelování

1. **Alignment** zadané sekvenční a sekvenční templátu
2. Extrakce proteinové **páteře** ze struktury templátu a umístění **postranních řetězců**
3. **Modelování** očípek a smyček
4. **Minimalizace** energie
5. **Validace** namodelované struktury

85

Swiss-Model



- Výběr modelu (manuální, automatický)
- Podle vybraného modelu pak predikuje strukturu zadané sekvenční
- Součástí výstupu je sada parametrů hodnotících **kvalitu** modelu. Při využití více templátů je tak možno porovnat jednotlivé modely

<http://swissmodel.expasy.org/>

The screenshot displays the SWISS-MODEL web interface. On the left, the 'Start a New Modelling Project' section is visible, including a 'Search For Templates' button. The central part shows 'Template Results' with a table of template options. The right part shows 'Model Results' for a selected template, including quality estimates (Global and Local Quality Estimate), a comparison graph, and a 3D ribbon diagram of the protein structure.

Template ID	Template Name	Coverage	OMEGA	CS	AS	TS	Model
2x62 1.A	BARRIER-TO-AUTOREGULATION FACTOR	0.96	0.90	100.00	NMR	homodimer	CS
2x62 1.B	BARRIER-TO-AUTOREGULATION FACTOR	0.96	0.90	100.00	NMR	homodimer	AS
2x62 1.A	BARRIER-TO-AUTOREGULATION FACTOR	0.96	0.90	100.00	NMR	homodimer	TS
2x62 1.A	BARRIER-TO-AUTOREGULATION FACTOR	0.96	0.90	100.00	NMR	homodimer	AS
2x62 1.A	BARRIER-TO-AUTOREGULATION FACTOR	0.96	0.90	100.00	X-ray 2.0 Å	homodimer	AS
5xh1 1.B	Barrier to autoregulation factor	0.96	0.90	95.45	X-ray 2.1 Å	heterodimer	AS
6x6r 1.B	barrier to autoregulation factor (BAG)	0.96	0.87	95.45	X-ray 2.3 Å	heterodimer	AS
2x62 1.B	DNA repair and recombination protein HHR23	0.40	-	22.03	X-ray 3.2 Å	homodimer	AS
2x64 1.A	DNA REPAIR AND RECOMBINATION PROTEIN HHR23	0.40	-	22.03	X-ray 3.2 Å	homodimer	AS



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.
```

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.

Ab initio



- Nejuniverzálnější – vychází pouze ze sekvence
- Výpočetně **nejnáročnější**
- Zahrnuje řadu kroků:
 - Predikce 2D struktury
 - Modelování jednotlivých fragmentů
 - Kombinace fragmentů navzájem
 - Doplnění smyček a flexibilních úseků
- **Nízká spolehlivost** zejm. pro větší proteiny

Ab initio



- Quark
- RaptorX
- Rossetta

User Input

```
>1ci4A (87 residues)
TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKLEERGFDKAYVVLGQFLVLKKDEDLFREW
LKDTCGANAKQSRDCFGLREWCDAFL
```

Predicted Secondary Structure

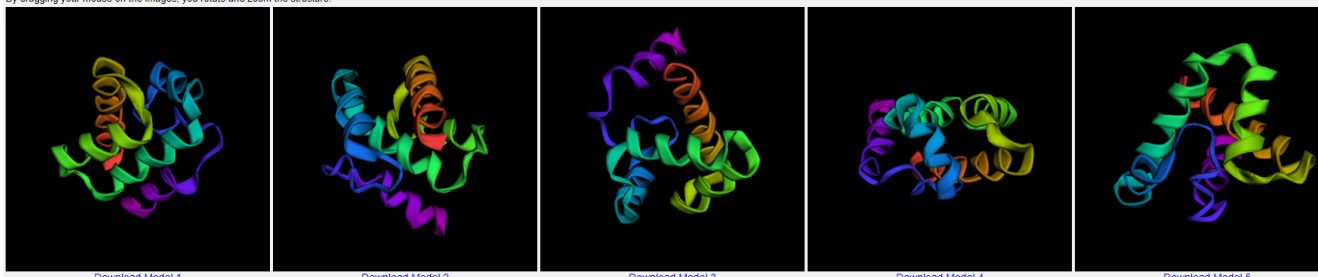
```
                20          40          60          80
Sequence  TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKLEERGFDKAYVVLGQFLVLKKDEDLFREMLKDTCGANAKQSRDCFGLREWCDAFL
Prediction CCC#####C CCCCCCCCC#####C C#####C C#####C
Conf. Score 98889999987999998744789889999999796599999999588899999999688999999999999859
           H:Helix; S:Strand; C:Coil
```

Predicted Solvent Accessibility

```
                20          40          60          80
Sequence  TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKLEERGFDKAYVVLGQFLVLKKDEDLFREMLKDTCGANAKQSRDCFGLREWCDAFL
Prediction 553330221123223321120110032002102421132002000200113232310220022102031310310010022003324
           Values range from 0 (buried residue) to 9 (highly exposed residue)
```

Top 5 Final Structure Model

By dragging your mouse on the images, you rotate and zoom the structure.



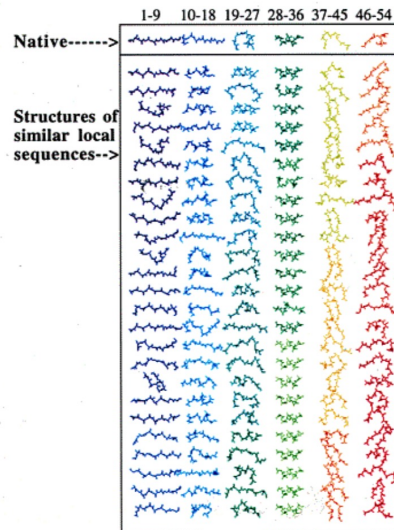
De novo modelling with Rossetta

(David Baker lab, Univ. of Washington)

- In contrast to threading, Rosetta does *de novo* prediction – doesn't use templates/homologous structures
- instead performs Monte Carlo search through space of conformations to find minimal energy conformation

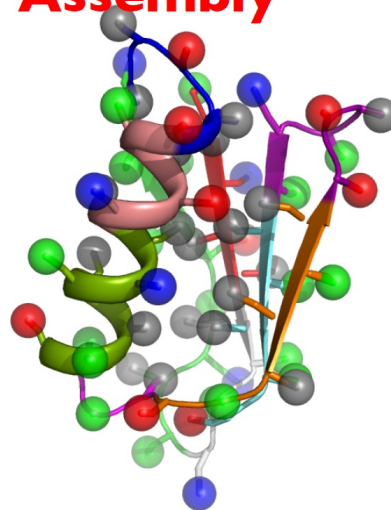
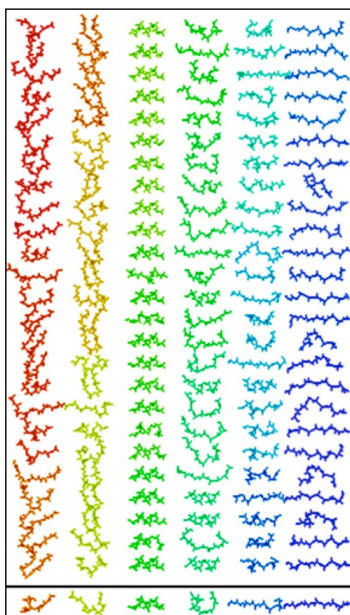
De novo modelling with Rosetta

- fragments are selected from known structures
- the window-fragment matches are calculated using
 - PSI-BLAST to build a profile model of the sequence
 - the predicted secondary structure of the sequence



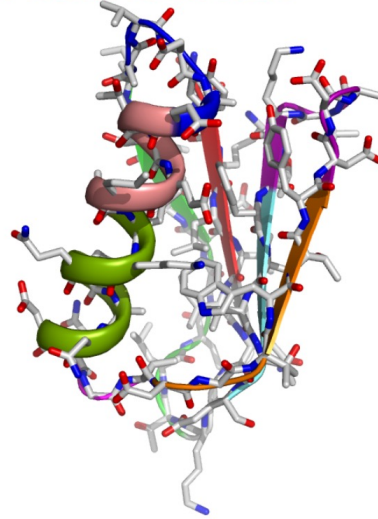
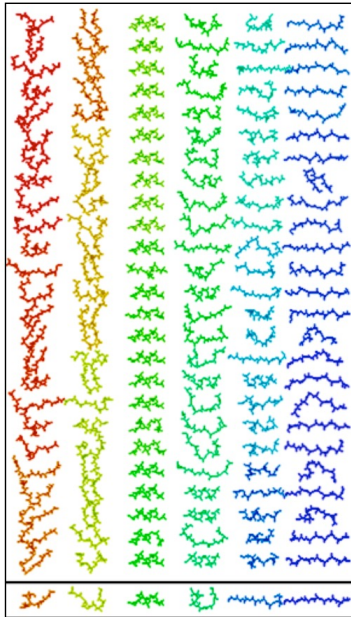
De novo Modeling with Rosetta

Stage I. Fragment Assembly



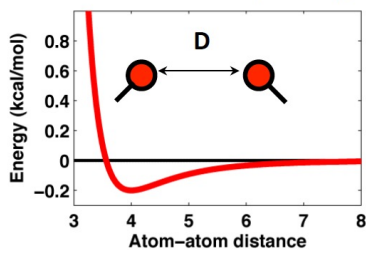
De novo Modeling with Rosetta

Stage II. All-atom refinement

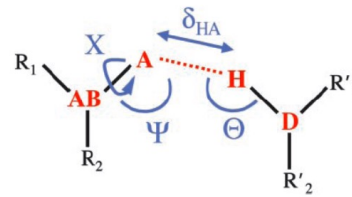


Ingredients of a high resolution potential

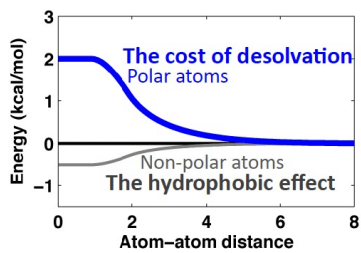
1. Van der waals packing



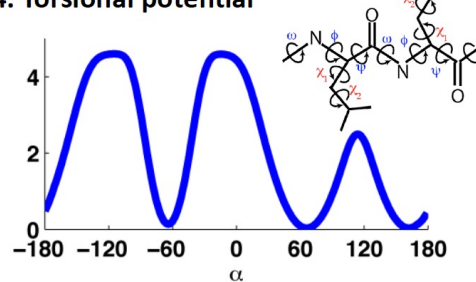
2. Hydrogen bonds



3. Manifestations of water



4. Torsional potential



Scoring Function Takes Into Account

- residue environment (solvation)
- residue pair interactions (electrostatics, disulfides)
- strand pairing (hydrogen bonding)
- strand arrangement into sheets
- helix-strand packing
- steric repulsion
- etc.
- scoring function search progressively adds terms during search
 - initially the steric overlap term is used
 - then all but “compactness” terms are used
 - etc.
- search is initiated from different random seeds

WEB server - Robetta

<http://robetta.bakerlab.org>

Response Times

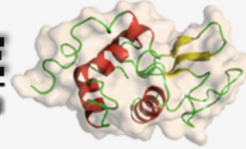
To prevent unnecessary usage we require two manual steps for full structure predictions. The first step is to submit your sequence for domain and template detection. The second step is to continue for 3-D models. You may only select one domain at a time for structure predictions. The second step is computationally expensive so please continue with this step only if necessary. You may help increase computing resources for this service by joining our distributed computing project [Rosetta@HOME](#) and spreading the word out to friends and colleagues.

- ~10 minutes - hours for domain and template detection.
- ~1 day - weeks for high accuracy homology models (templates detected with high confidence > 0.8 and sequence identity > 40%).
- ~1 week - months for difficult targets.

Zhang Lab - QUARK



QUARK ONLINE
Ab Initio Protein Structure Prediction



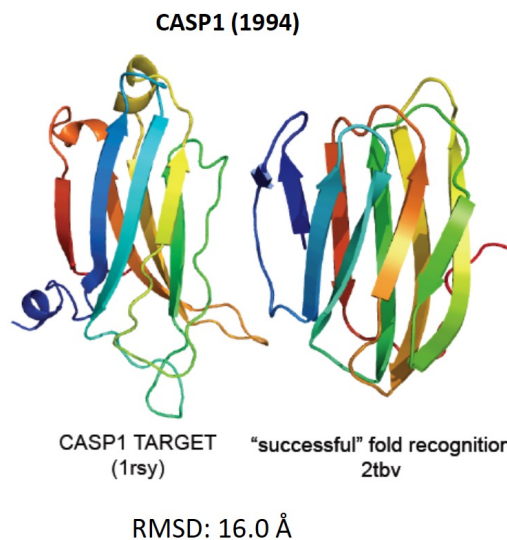
QUARK is a computer algorithm for ab initio protein structure prediction and protein peptide folding, which aims to construct the correct protein 3D model from amino acid sequence only. QUARK models are built from small fragments (1-20 residues long) by replica-exchange Monte Carlo simulation under the guide of an atomic-level knowledge-based force field. QUARK was ranked as the No 1 server in Free-modeling (FM) in CASP9 and CASP10 experiments. Since no global template information is used in QUARK simulation, the server is suitable for proteins that do not have homologous templates in the PDB library. Go to [example](#) to view an example of QUARK output. The server is only for non-commercial use. Questions about the QUARK server can be posted at the [Service System Discussion Board](#).

Cut and paste your sequence (in [FASTA format](#), less than 200 AA. [Example input](#))

Driving innovation in
protein structure
prediction:
“CASP”

Critical Assessment of
Structure Prediction

**Five *blind*
predictions per
target**



CASP 11 (2014)

CASP11 in numbers

Number of groups registered	208
including: expert groups	123
prediction servers	85
Number of regular targets released	100
including all-group (human) targets	55
Targets canceled for all/manual prediction	7 / 10
Number of refinement targets released	37
Number of assisted prediction targets released	71
Number of targets received from	
Joint Center for Structural Genomics (JCSG):	32
Structural Genomics Consortium (SGC):	4
Midwest Center for Structural Genomics (MCSG):	8
Northeast Structural Genomics Consortium (NESG):	5
New York Structural Genomics Research Center (NYSGRC):	6
Non-SGI research Centers and others (Others):	40
Seattle Structural Genomics Center for Infectious Disease (SSGCID):	4
NatPro PSI:Biological (NatPro):	1

<http://predictioncenter.org/casp11/results.cgi>

CA

12th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction



CASP12 in numbers

Number of groups registered	192
including: <i>expert groups</i>	112
<i>prediction servers</i>	80
Number of regular targets released	82
including <i>all-group (human) targets</i>	56
Targets canceled and not re-released for all/manual prediction	11 / 11
Number of refinement targets released	42
Number of assisted prediction targets released	14

Prediction category	Number of groups/servers contributing	Number of models designated as 1	Total number of models
Tertiary structure predictions	128 / 43	8362	37672
Data assisted predictions	16 / 1	109	528
Residue-residue contacts	38 / 30	3077	3077
Accuracy estimation	47 / 32	3700	7400
Interface accuracy	3 / 0	65	66
Refinement	39 / 5	1457	6227
All (unique):	188 / 80	16770	54970

<http://predictioncenter.org/casp12/results.cgi>

CASP13 in numbers

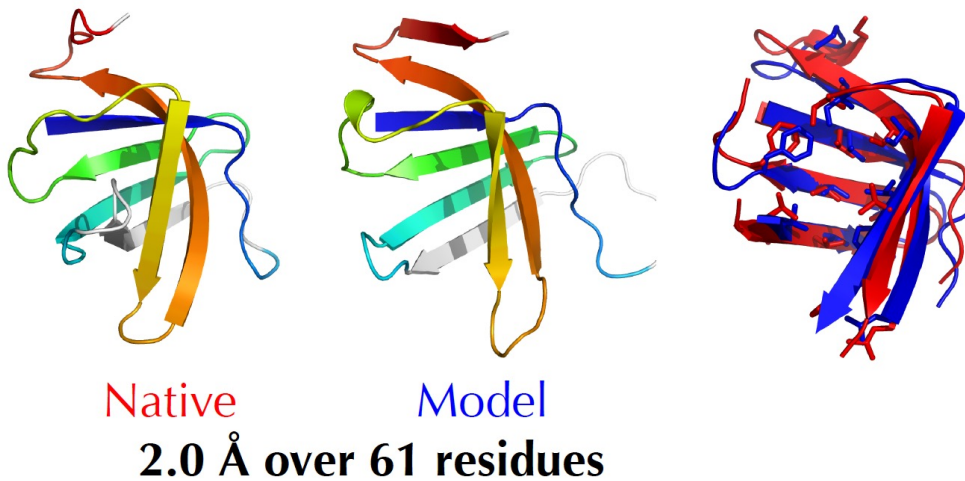
Number of groups registered	210
including: <i>expert groups</i>	123
<i>prediction servers</i>	87
Number of tertiary structure prediction targets released	90
(including <i>all-group targets</i>)	(82)
Number of hetero-multimer targets released	13
Number of refinement targets released	31
Number of assisted prediction targets released	60
Targets canceled (all / human)	(10 / 12)
Targets available/expired for manual non-QA prediction	0 / 72
Targets available/expired for server non-QA prediction	0 / 80
Targets available/expired for QA prediction	0 / 80
Targets available/expired for assisted prediction	0 / 59
Targets available/expired for multimer prediction	0 / 12

Prediction category	Number of groups/servers contributing	Number of models designated as 1	Total number of models
Tertiary structure predictions	107 / 39	7542	35982
Oligomeric predictions	40 / 9	662	2861
Data assisted predictions	24 / 5	456	2017
Residue-residue contacts	46 / 25	3914	3914
Accuracy estimation	52 / 41	4332	8687
Refinement	33 / 6	847	3788
All (unique):	185 / 87	17753	57249

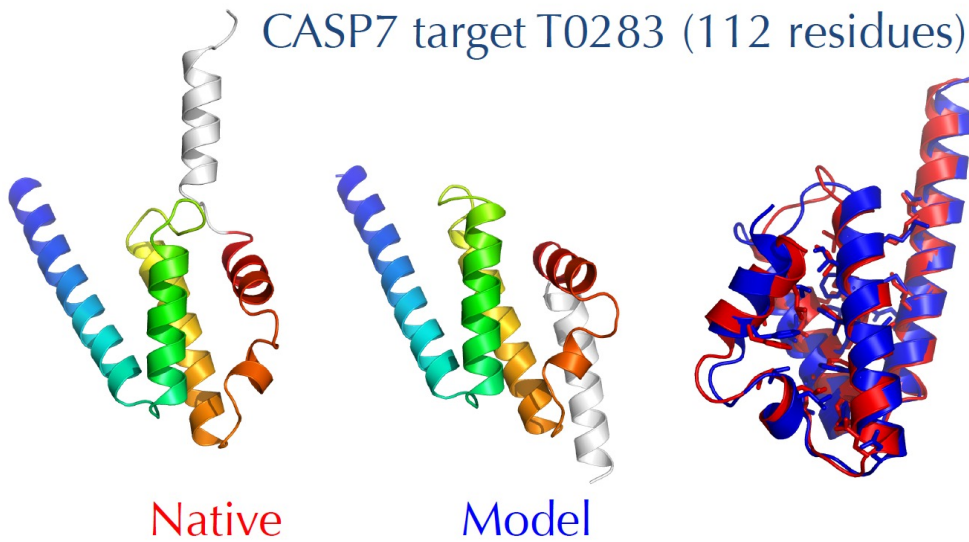
<http://predictioncenter.org/casp13/results.cgi>

De novo successes: all- β

CASP7 target T0316 (domain 3)

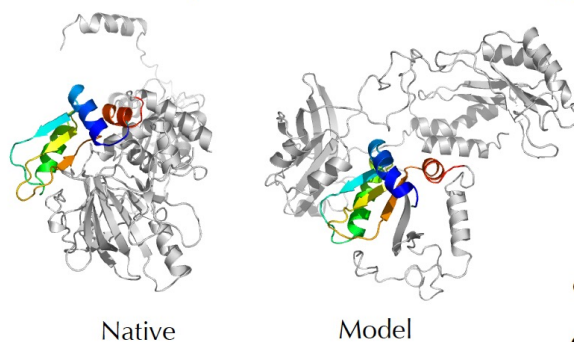


De novo successes: all- α



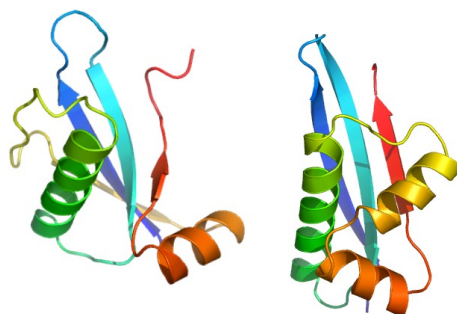
1.4 Å over 90 residues

Is protein folding *solved*?



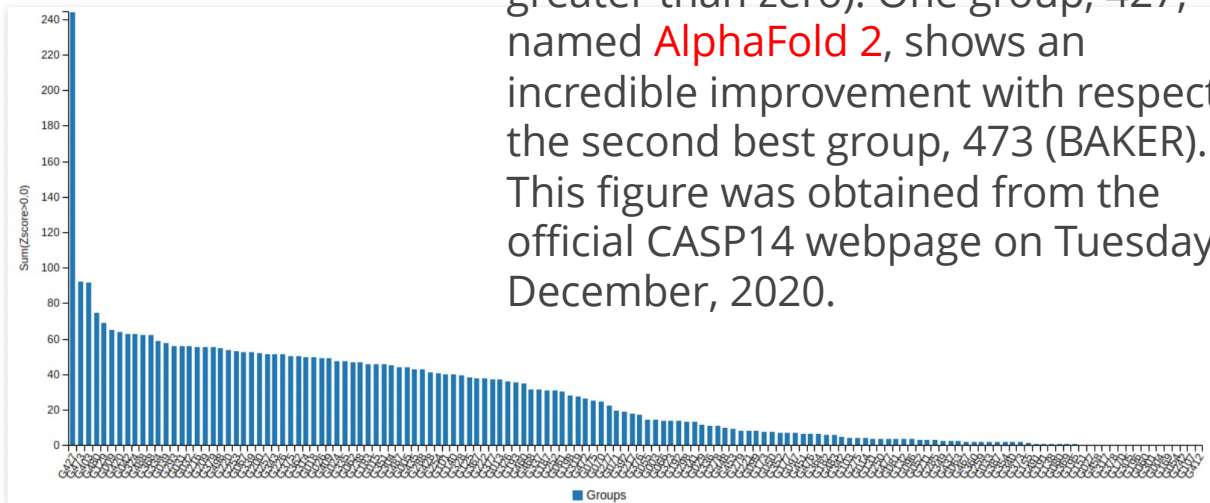
NO! til now?

- Success in <1/3 of cases.
- Conformational sampling still a huge issue



CASP 14 (2020)

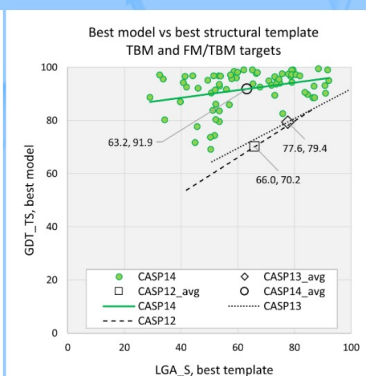
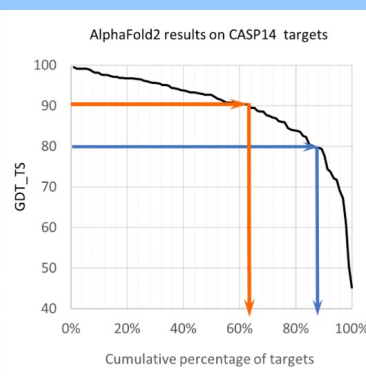
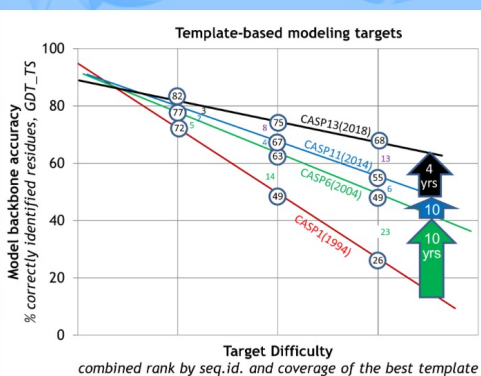
Ranking of participants in CASP14, as per the sum of the Z-scores of their predictions (provided that these are greater than zero). One group, 427, named **AlphaFold 2**, shows an incredible improvement with respect to the second best group, 473 (BAKER). This figure was obtained from the official CASP14 webpage on Tuesday 1st December, 2020.



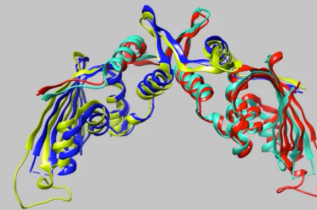
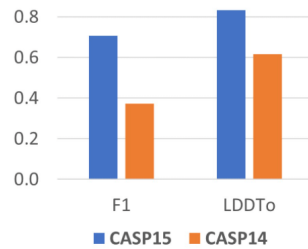
Models based on templates identified by sequence similarity remain the most accurate. Over the course of the CASP experiments there have been enormous improvements in this area. However, the overall accuracy improvements that we have seen in the first 10 years of CASP remained unmatched until CASP12 (2016), when a new burst of progress happened [Kryshtafovych et al, 2018]. In two years from 2014 to 2016, the backbone accuracy of the submitted models improved more than in the preceding 10 years. The next CASP continued the trend [Croll et al, 2019], and the 2014-2018 model accuracy improvement doubled that of 2004-2014 (see left plot). Several factors contributed to this, including more accurate alignment of the target sequence to that of available templates, combining multiple templates, improved accuracy of regions not covered by templates, successful refinement of models, and better selection of models from decoy sets due to improved methods for estimation of model accuracy.

CASP14 marked an extraordinary increase in the accuracy of the computed three-dimensional protein structures with the emergence of the advanced deep learning method AlphaFold2. Models built with this method proved to be competitive with the experimental accuracy (GDT_TS>90) for ~2/3 of the targets and of high accuracy (GDT_TS>80) for almost 90% of the targets (middle plot). The accuracy of CASP14 models for TBM targets significantly superseded accuracy of models that can be built by simple transcription of information from templates, and reached the level of GDT_TS=92 on average, which is significantly higher than the corresponding averages in previous two CASPs (right plot).

template-based modeling



CASP15 (2022) showed enormous progress in modeling multimolecular protein complexes. The assembly modeling (a.k.a. quaternary structure modeling, oligomeric modeling, multimeric modeling) has been assessed in CASP since 2016 (CASP12). Typically, models were of good accuracy when templates were available for the structure of the whole target complex. After the success of AlphaFold2 in CASP14 (2020), it was expected that deep learning methodology that brought monomeric modeling to qualitatively new level will be extended to multimeric modeling. Indeed, CASP15 showed that newly developed methods are capable of accurate reproducing structures of oligomeric complexes and outperform CASP14 methods by a large margin. In particular, the accuracy of models almost doubled in terms of the Interface Contact Score (ICS a.k.a. F1) and increased by 1/3 in terms of the overall fold similarity score LDDTo (left panel). An impressive example of multimeric modeling is shown in the right panel below.



CASP15: T1113o
model 239_2: F1=92.2; LDDTo=0.913

Jakou metodu zvolit?



1. Mám homologní protein se známou strukturou → homologní modelování
2. Využiji experimentální data
 - Threading
 - Kombinace více templátů pro jednotlivé části struktury
 - Různé predikční nástroje
3. *Ab initio* modelování smyček a částí sekvence bez vhodného templátu
4. Mám unikátní sekvenci – *ab initio*

Predikce kvartérní struktury



Zahrnuje různé úrovně, např.:

- Predikce vazebných míst
- Predikce aminokyselin podílejících se na interakci
- Odhad oligomerního stavu
- Protein-protein docking (protein-nukleová kyselina docking)

- SW dosud často nedokonalý, **nízká spolehlivost** predikce
- Složitější postupy většinou nejsou automatizované

109

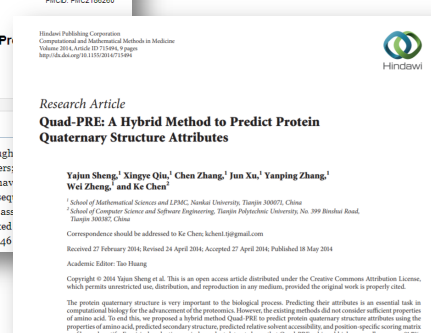
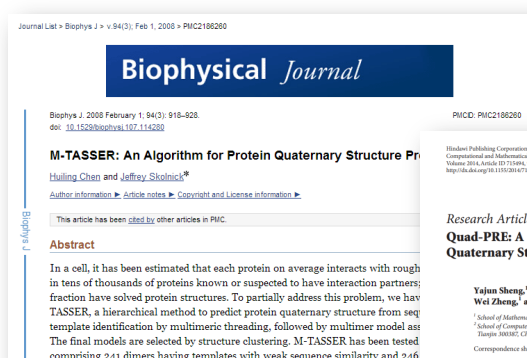
Predikce kvartérní struktury



Programy většinou vycházejí z podobnosti sekvence a/nebo 3D struktury se známými proteiny

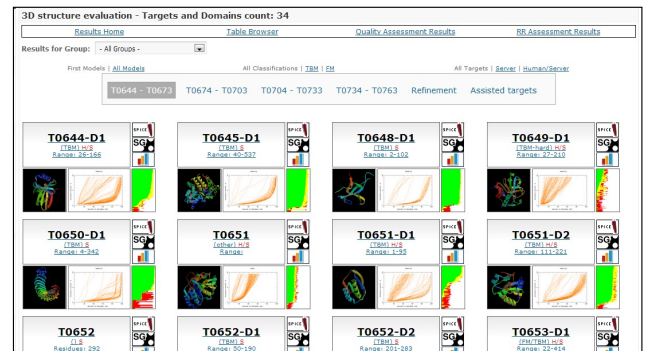
Příklady SW:

- QuatIdent
- QuaBingo
- M-TASSER
- Quad-PRE



Hodnocení kvality predikčních nástrojů - CASP

- *Critical Assessment of Techniques for Protein Structure Prediction*
- 2020 – CASP14
- Predikce vyřešených, ale zatím nepublikovaných struktur
- **Rozsáhlá analýza predikčních programů**
 - Predikce terciárních struktur
 - Identifikace neuspořádaných oblastí
 - Funkční predikce (predikce vazebných míst)
 - Interakce mezi doménami, podjednotkami a proteiny
 - Hodnocení spolehlivosti



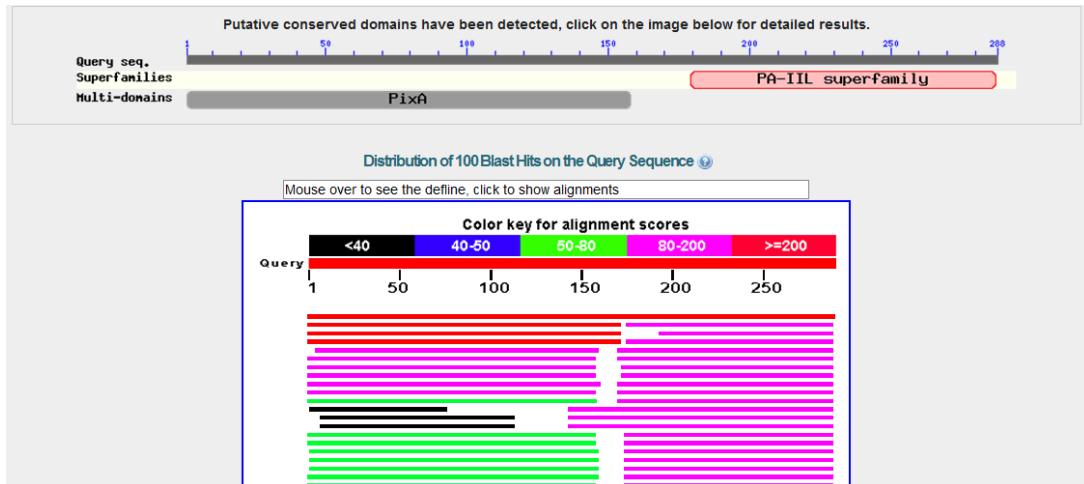
Ale!

! pozor na domény !

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 [cl|15110] [View concise result](#)

Local query sequence

Graphical summary [show options](#)

Query seq. Non-specific hits Superfamilies Multi-domains

PixA PA-IIL PA-IIL superfamily

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

List of domain hits

Description	PssmId	Multi-dom	E-value
PA-IIL[pfam07472]. 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[pfam12306]. 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

pfam07472: PA-IIL

Fucose-binding lectin II (PA-IIL)
 In *Paucimonas scrugosa* 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 ions that mediate the binding of fucose in a recognition mode unique among carbohydrate-protein interactions.

PubMed References

Structural basis for oligosaccharide binding to fucose in *Paucimonas scrugosa* in the lungs of cystic fibrosis patients. *Nat Struct Biol* 2002 Dec 9(12):919-921

pfam07472 is a member of the superfamily cl06486.

Sequence Alignment

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

100L_A	6	FILPANIYGVYAFANAAQIQIVLVDS	VFR	ATFSSQITSA	[1].LGS	[2].LMSGS	GASK	83
qt_81680028	7	FILPANIYGVYLVNNAATQVLEIVDS	EFR	AAFSQVYDGN	[1].LGT	[2].LMSGS	GRVR	84
qt_75485512	254	FQLPNSIKLGLSAGVHTIQIKVIVDD	QIV	DTLSSQVMSV	LGF	[2].VSSST	GRVC	290
qt_123468540	14	FSIPVPIIPRALYFANAAQIQIKLVFDD	SQK	[2].AVKLTTRDGP	[1].EAT	LMSGS	GKIR	71
qt_123570095	187	FSIPVPIIPRALYFANAAQIQIKLVFDD	APK	[2].ATFVMSIDGV	[1].LFT	LMSGS	GKIR	244
qt_123585156	174	FSIPVPIIPRALYFANAAQIQIKLVFDD	EPK	[2].ATFVMSIDGV	[1].LGT	[2].LMSGS	GRVR	233
208A_A	7	FILPANIYGVYLVNNAATQVLEIVDS	EPK	[2].ATFVMSIDGV	[1].LGT	[2].LMSGS	GRVR	88
280I_A	6	FILPANIYGVYLVNNAATQVLEIVDS	EFR	AAFSQVYDGN	[1].LGT	[2].LMSGS	GRVR	83
qt_107102893	2	FILPANIYGVYAFANAAQIQIVLVDS	EIA	ATFSSQITSA	[1].LGT	[2].LMSGS	[1].GRVQ	80
2V9V_A	14	FSIPVPIIPRALYFANAAQIQIKLVFDD	[2].EFA	AVKLTTRDGP	[1].EAT	LMSGS	GKIR	71

NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí
 ➤ celý protein

pfam12306: PixA

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. PixA is thought to be specifically produced in *Xenorhabdus nematophila*. It is an inclusion body protein.

PubMed References

Analysis of the PixA inclusion body protein of *Xenorhabdus nematophila*. *J. Bacteriol* 2008 Apr; 190(7):2708-2710

pfam12306 is classified as a model that may span more than one domain.

pfam12306 is not assigned to any domain superfamily.

Sequence Alignment

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

qt_123655921	2	[2].NIVDILVVIDEY	ILK	[17].S	[2].PIQL	[4].SNG	[7].VDRVARD	[7].GKELAVMLRQGD	84
qt_123464695	13	[2].QSIQHLAVIEDY	IKK	[10].M	PIGI	[1].STA	LNLNGST	[8].TDRGLKLNFGD	77
qt_123180777	10	[2].QDMLAVIEDY	VKK	[10].A	PIGI	[1].SNG	QFLCTGA	[7].TADLEIIAYFGD	73
qt_53717990	9	[2].QKIRLVVIDEY	IRS	[10].Q	PIGI	[1].SDE	QFLCTGS	[8].TDRLEFRANFGD	73
qt_254248506	27	[2].QDILAVIEDY	IKL	[10].L	PIAV	[1].SRA	VRLYFGA	[8].PADPVLILYFGD	81
qt_170734850	2	[2].VRCDALVYDAV	LLS	[10].A	PVVI	[1].GRH	IKVLSGD	[7].DNGLYAGLSPGD	85
qt_53748592	18	[2].LIINVCITDQDA	ILA	[10].M	PIAL	[1].SAY	IKVLSGD	[8].PDRILNANVFD	82
qt_134279425	20	[2].SRVDLVVIDEY	VKK	[10].I	PVPP	[1].SRA	LVVICAGS	[8].SREALCTANVGD	82
qt_170702239	10	[2].QKTLAVIAAEK	[1].IKK	[10].R	PVQL	[1].SSE	QFLCTDP	[8].ANRDKYANVGD	79
qt_258424079	20	[2].QIVVYGLVYDAV	IYA	[11].K	SNPI	[1].SRS	EYNGASTV	[7].TADLSTVYRQGS	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 interface. At the top, there are navigation tabs: Research, Training, Industry, About Us, Help, and Site Index. Below this, the breadcrumb trail reads: EBI > Tools > Protein Functional Analysis > InterProScan Sequence Search. The main heading is "InterProScan Results", with sub-tabs for Summary Table, Tool Output, Visual Output (selected), Submission Details, and Submit Another Job. A "Download in SVG format" button is visible. The analysis details include: InterProScan (version: 4.8), Sequence: Sequence_1, Length: 288, and CRC64: 3FAE4C40C2498B64. The launch and finish times are both listed as Wed, May 16, 2012 at 17:31:03 and 17:35:39 respectively. The main content area shows a table of InterPro Matches. The first match is IPR010907, Calcium-mediated lectin, with a description of "no description". The second match is IPR021087, Uncharacterised protein family PixA/AidA, with a description of "PixA". A legend at the bottom identifies various domain databases: 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.html

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c2mvvC <input type="radio"/>			100.0	60	PDB header: sugar-binding protein Chain: C; PDB Molecule: bda; PDB title: crystal structure of bda lectin from burkholderia2 cenocepacia in complex with alpha-methyl-mannoside at 1.73 angstrom resolution
2	c2xr4A <input type="radio"/>			100.0	43	PDB header: sugar binding protein Chain: A; PDB Molecule: lectrn; PDB title: c-terminal domain of bc2l-c lectin from burkholderia cenocepacia
3	d2chhal <input type="radio"/>			100.0	37	Fold: Calcium-mediated lectin Superfamily: Calcium-mediated lectin Family: Calcium-mediated lectin

NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

➤ celý protein

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

Local query sequence

Graphical summary [show options](#)

List of domain hits

Description	PssmId	Multi-dom	E-value
PA-IIL[pfam07472] ; 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[pfam12306] ; 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

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c2xr4A 		100.0	44	PDB header: sugar binding protein Chain: A; PDB Molecule: lectin; PDBTitle: c-terminal domain of bc2l-c lectin from burkholderia cenocepacia	
2	c2vrvC 		100.0	62	PDB header: sugar-binding protein Chain: C; PDB Molecule: bcda; PDBTitle: crystal structure of bcda lectin from burkholderia2 cenocepacia in complex with alpha-methyl-mannoside at 1.73 angstrom resolution	
3	dluzva 		100.0	30	Fold: Calcium-mediated lectin Superfamily: Calcium-mediated lectin Family: Calcium-mediated lectin	

Phyre – n-term http://www.sbg.bio.ic.ac.uk/phyre2/phyre2_output/e332b1ecabb8d0a6/summary.html


#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c1sddb 		83.7	9	PDB header: blood clotting Chain: B; PDB Molecule: coagulation factor v; PDBTitle: crystal structure of bovine factor vai	
2	c3cdzB 		76.1	6	PDB header: blood clotting Chain: B; PDB Molecule: coagulation factor viii light chain; PDBTitle: crystal structure of human factor viii	
3	d1kbva2 		68.0	13	Fold: Cupredoxin-like Superfamily: Cupredoxins Family: Multidomain cupredoxins	


Swissprot – whole protein


   **SWISS-MODEL Workspace**
Modelling Tools Repository Documentation


[myWorkspace] [login]

Workunit: P000007 - Overview




Print/Save this page as 

Model Summary 




Model information:
Modelled residue range: 169 to 288
Based on template: [2vnnD] (1.7 Å)
Sequence Identity [%]: 56.35
Evaluate: 0.00e-1

Quality information: [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  [+/-]

http://swissmodel.expasy.org/workspace/index.phpuserid=michaw@chemi.muni.cz&key=0f449e99bc0176edfa75fba19b2d96e48&func=workspace_modelling&prid=P000007



Rosetta@home  Project Computing Community Site Sign Up Login





You don't have to be a scientist to do science.


By simply running a free program, you can help advance research in medicine, clean energy, and materials science.

[Join Rosetta@home](#)


  **HHMI**
HOWARD HUGHES MEDICAL INSTITUTE

 **INSTITUTE FOR Protein Design**
UNIVERSITY OF WASHINGTON

 **UNIVERSITY OF WASHINGTON**



Rosetta@home needs your help to determine the 3-dimensional shapes of proteins in research that may ultimately lead to finding cures for some major human diseases. By running the Rosetta program on your computer while you don't need it you will help us speed up and extend our research in ways we couldn't possibly attempt without your help. You will also be helping our efforts at designing new proteins to fight diseases such as HIV, Malaria, Cancer, and Alzheimer's. Please [join us](#) in our efforts!


08:05:04 GMT

[PUZZLES](#) [CATEGORIES](#) [GROUPS](#) [PLAYERS](#) [RECIPES](#) [CONTESTS](#)
[BLOG](#) [FEEDBACK](#) [FORUM](#) [WIKI](#) [FAQ](#) [ABOUT](#) [CREDITS](#)

The Science Behind Foldit

Foldit is a revolutionary crowdsourcing computer game enabling you to contribute to important scientific research. This page describes the science behind Foldit and how your playing can help.

Page Contents:

- [What is protein folding?](#)
- [Why is this game important?](#)
- [Foldit Scientific Publications](#)
- [News Articles about Foldit](#)
- [News Articles about Rosetta](#)
- [Rosetta@Home Screensaver](#)
- [Community Rules](#)
- [Let's Foldit Podcast](#)
- [Instructions for Educators](#)
- [Terms of Service and Consent](#)
- [Credits](#)


<http://fold.it/portal/>


What is protein folding?

What is a protein? Proteins are the workhorses in every cell of every living thing. Your body is made up of trillions of cells, of all different kinds: muscle cells, brain cells, blood cells, and more. Inside those cells, proteins are allowing your body to do what it does: break down food to power your muscles, send signals through your brain that control the body, and transport nutrients through your blood. Proteins come in thousands of different varieties, but they all have a lot in common. For instance, they're made of the same <https://fold.it/portal/> consists of a long chain of

GET STARTED: DOWNLOAD


Win Beta
Windows (XP/Vista/7/8)


Mac Beta
OSX (10.7 or later)


Linux Beta
Linux (64-bit)

[Are you new to Foldit? Click here.](#)
[Are you a student? Click here.](#)
[Are you an educator? Click here.](#)

SEARCH

 Only search fold.it

RECOMMEND FOLDIT

USER LOGIN

Username: *

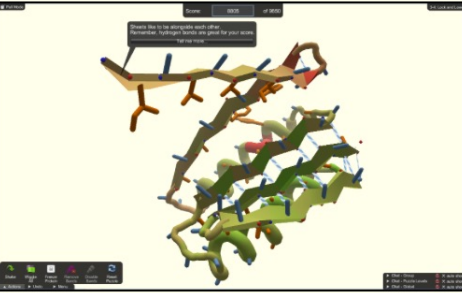
Password: *

[Create new account](#)
[Request new password](#)



Folded up Streptococcal Protein Puzzle
(+) [Enlarge This Image](#)

Just a game?



This is an example of a puzzle that a human can see the obvious answer to - fix the sheet that is sticking out!

(+) [Enlarge This Image](#)

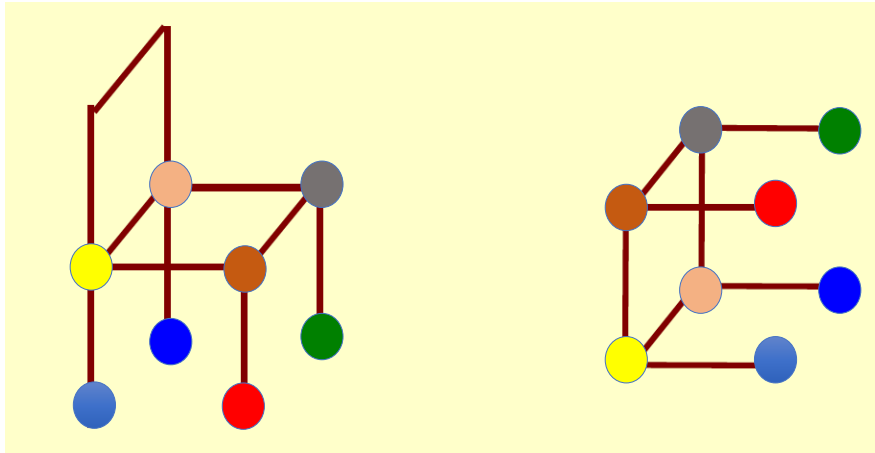
What other good stuff am I contributing to by playing?

Proteins are found in all living things, including plants. Certain types of plants are grown and converted to biofuel, but the conversion process is not as fast and efficient as it could be. A critical step in turning plants into fuel is breaking down the plant material, which is currently done by microbial enzymes (proteins) called "cellulases". Perhaps we can find new proteins to do it better.

Can humans really help computers fold proteins?

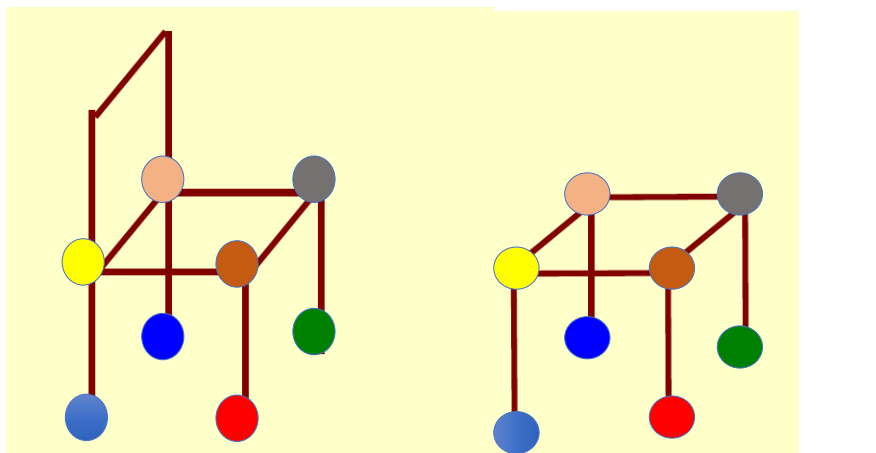
We're collecting data to find out if humans' pattern-recognition and puzzle-solving abilities make them more efficient than existing computer programs at pattern-folding tasks. If this turns out to be true, we can then teach human strategies to computers and fold proteins faster than ever!

Structure Superposition



The key is finding corresponding points between the two structures

Structure Superposition



The key is finding corresponding points between the two structures

Algorithms for Structure Superposition

Distance based methods:

DALI (Holm & Sander): Aligning scalar distance plots

SSAP (Orengo & Taylor): Dynamic programming using intra-molecular vector distances

MINAREA (Falicov and Cohen): Minimizing soap-bubble surface area

CE (Shindyalov & Bourne)

Vector based methods:

VAST (Bryant): Graph theory based secondary structure alignment

3D Search (Singh and Brutlag) & 3D Lookup (Holm and Sander): Fast secondary structure index lookup

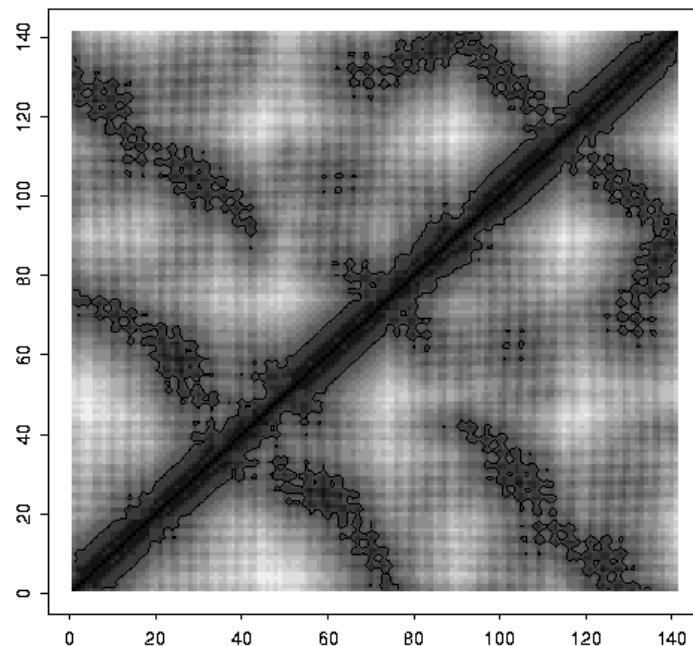
Both

LOCK (Singh & Brutlag) LOCK2 (Ebert & Brutlag): Hierarchically uses “Adaptive”

FATCAT(Flexible structure **A**lignment**T** by **C**haining **A**ligned fragment pairs allowing **T**wists, Ye & Godzik) – not further maintained?

<http://fatcat.godziklab.org/fatcat/>

DALI



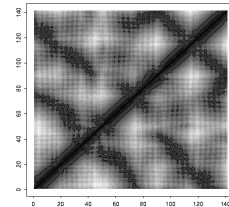
An intra-molecular distance plot for myoglobin

DALI

Based on aligning 2-D intra-molecular distance matrices

Computes the best subset of corresponding residues from the two proteins such that the similarity between the 2-D distance matrices is maximized

Searches through all possible alignments of residues using Monte-Carlo and Branch-and-Bound algorithms

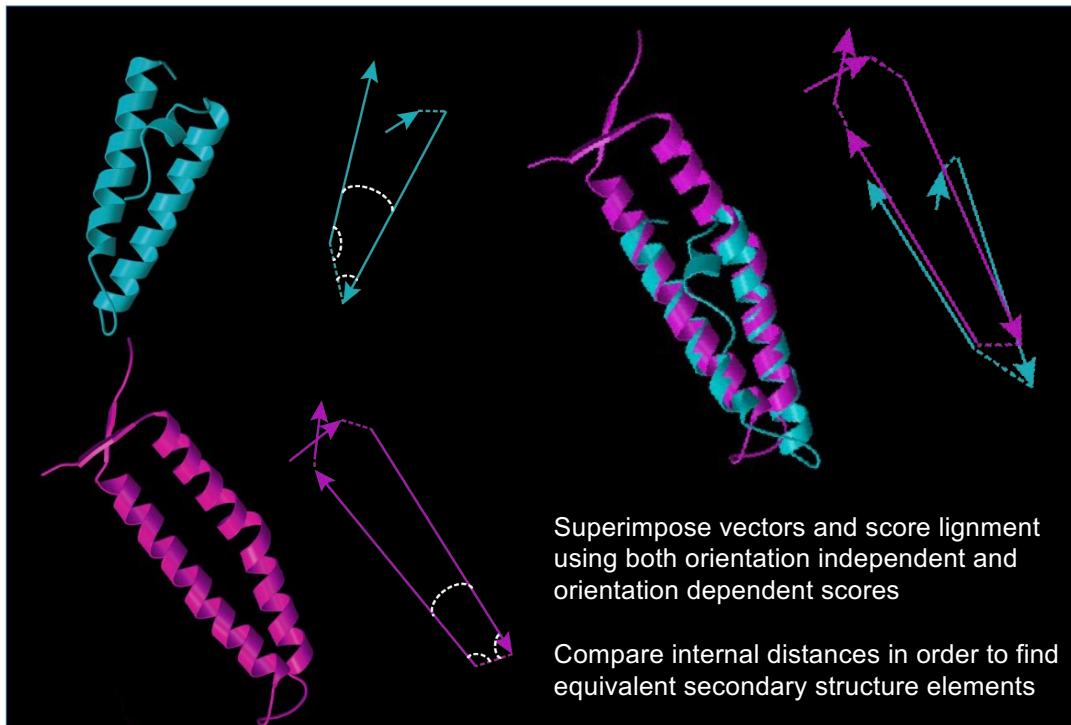


VAST – Vector Alignment Search Tool

Identifying similar structures by **purely geometric criteria** (and to identify distant homologs that cannot be recognized by sequence comparison). Find similarly shaped individual protein molecules or 3D domains (VAST+: similarly shaped macromolecular complexes)

- Aligns only secondary structure elements (SSE)
- Represents each SSE as a vector
- Finds all possible pairs of vectors from the two structures that are similar
- Uses a graph theory algorithm to find maximal subset of similar vector pairs
- Overall alignment score is based on the number of similar pairs of vectors between the two structures

LOCK2



FoldMiner: Structure Similarity Search Based on LOCK2 Alignment

FoldMiner aligns query structure with all database structures using LOCK2

FoldMiner up weights secondary structure elements in query that are aligned more often

FoldMiner outperforms CE and VAST is searches for structure similarity

The best to test as first:

Distance based methods

DALI

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

Vector and distance based method

FoldMiner (LOCK2) – local installation needed

“Adaptive”

FATCAT

<http://fatcat.godziklab.org/fatcat/>

Závěrem

- Struktura je klíčová pro správnou funkci proteinu
- Predikovat na základě sekvence (1D) lze 2D, 3D i 4D strukturu
- Vždy je nutné **kriticky kontrolovat** výstupy programů
- Ideální je využít více predikčních programů s různou metodologií a porovnat výsledky