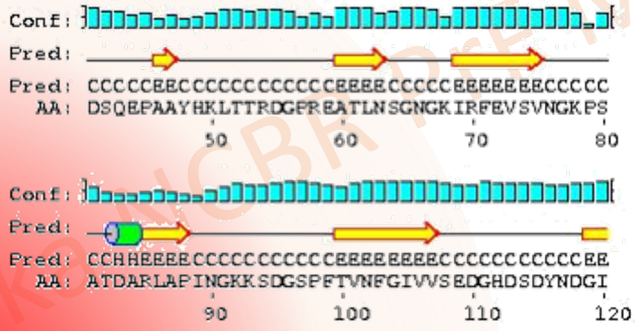
Predikce struktury proteinů

Struktura proteinů

1D

ADSQTSSNRAGEFSIPPNTDFRAIFFANAAE
QQHIKLFIGDSQEPAAYHKLTTRDGPREATL
NSGNGKIRFEVSVNGKPSATDARLAPINGK
KSDGSPFTVNFGIVVSEDGHDSDYNDGIVV
LQWPIG

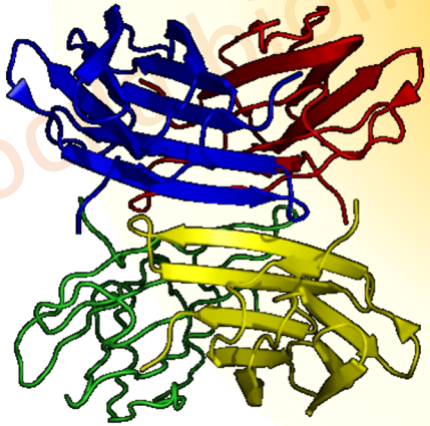
primární (sekvence)



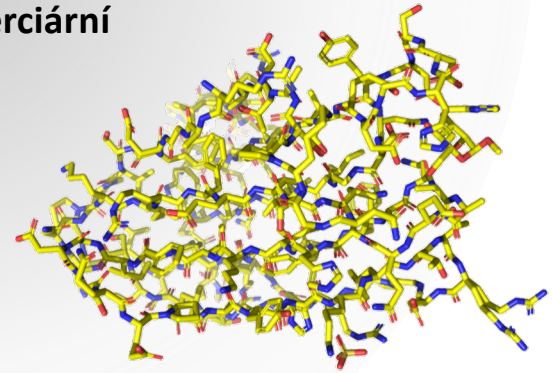
sekundární

2D

kvartérní



terciární

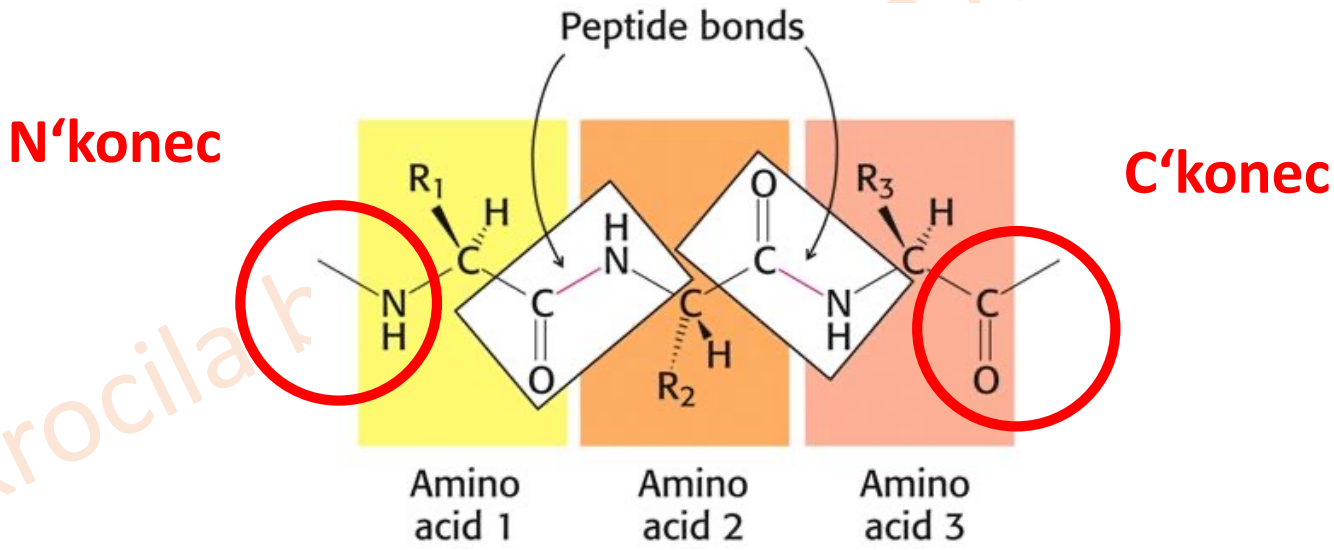


4D

3D

Primární struktura

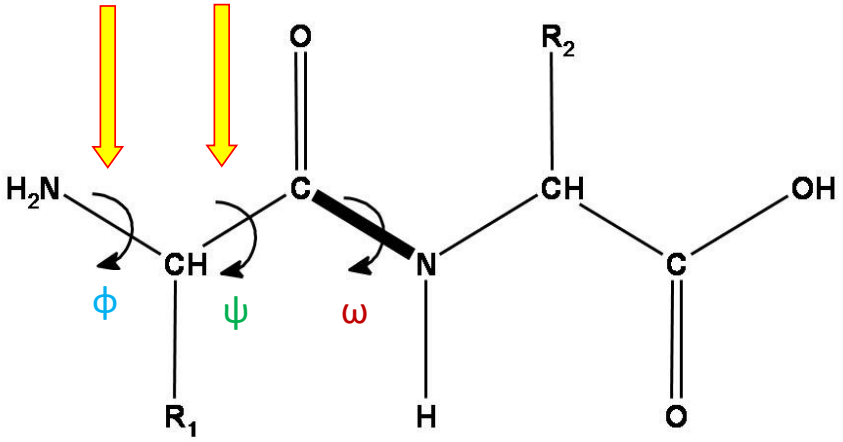
- Sekvence aminokyselin zapsaná od N‘ konce k C‘ konci



Sekundární struktura

2D

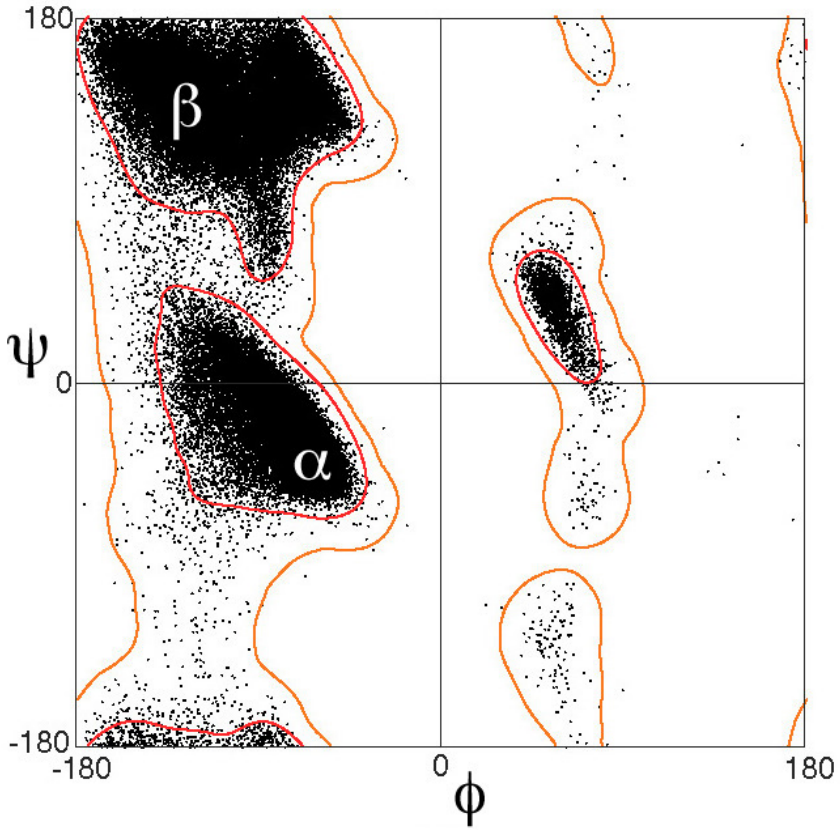
- Definována pomocí **torzních úhlů** peptidové páteře
- Pro každou aminokyselinu lze definovat tři úhly:
  - ϕ – úhel kolem vazby N-Cα
  - ψ – úhel kolem vazby Cα-C(karb.)
  - ω – úhel kolem peptidové vazby (180°, výjimečně 0°)
- Stabilizována pomocí vodíkových můstků mezi atomy peptidové kostry

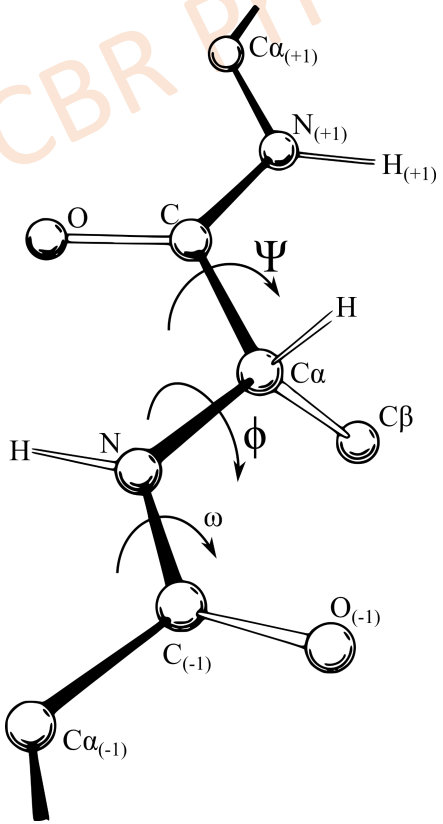


2D

Ramachandranův diagram

- Každé aminokyselině odpovídá jeden bod v diagramu

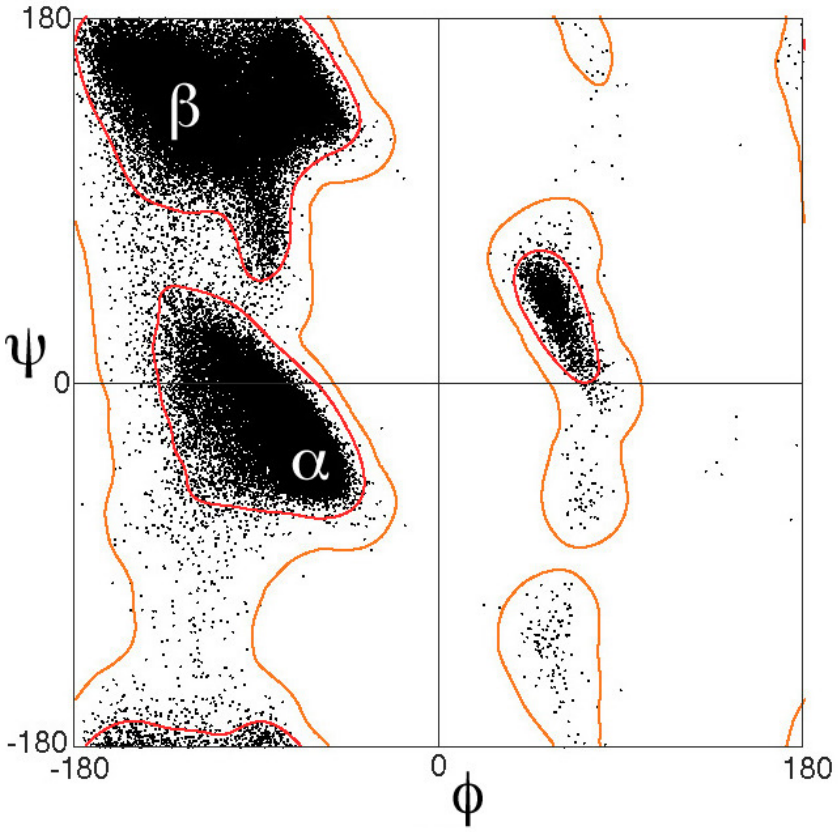


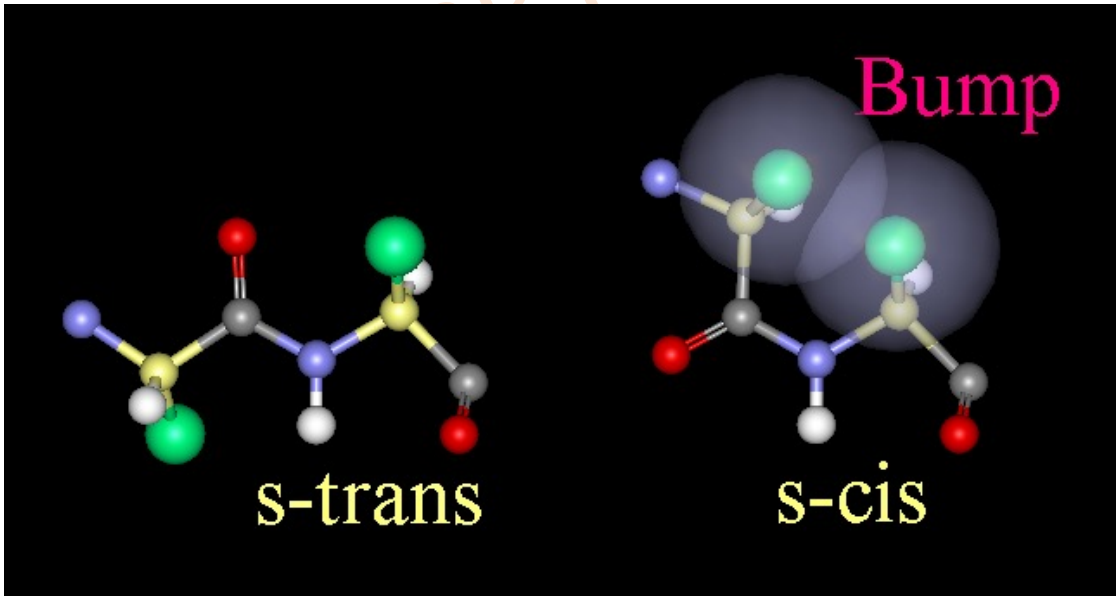


2D

Ramachandranův diagram

- Každé aminokyselině odpovídá jeden bod v diagramu

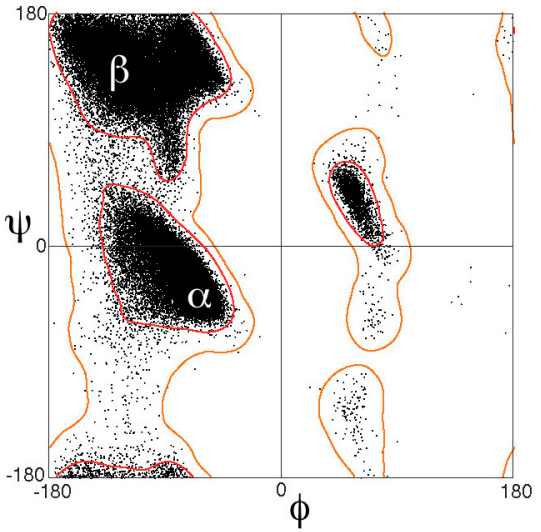




Sekundární struktura

2D

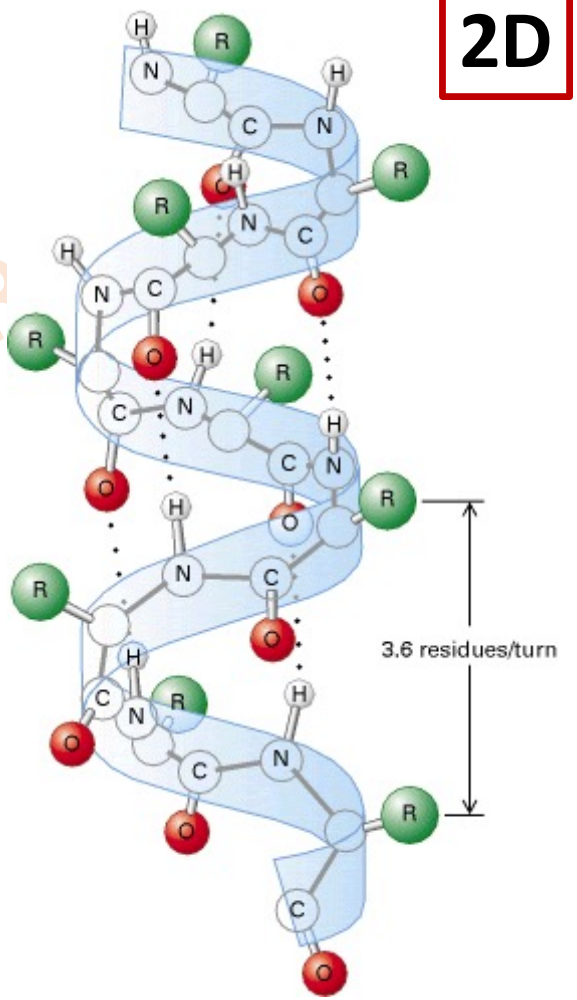
- Stabilní konformace polypeptidového řetězce
- Důležité pro udržení 3D struktury
- α-šroubovice (helix), β-skládaný list (sheet), otáčky, smyčky
- Cca 50 % aminokyselin je součástí α a β struktur



Šroubovice (helix)

2D

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

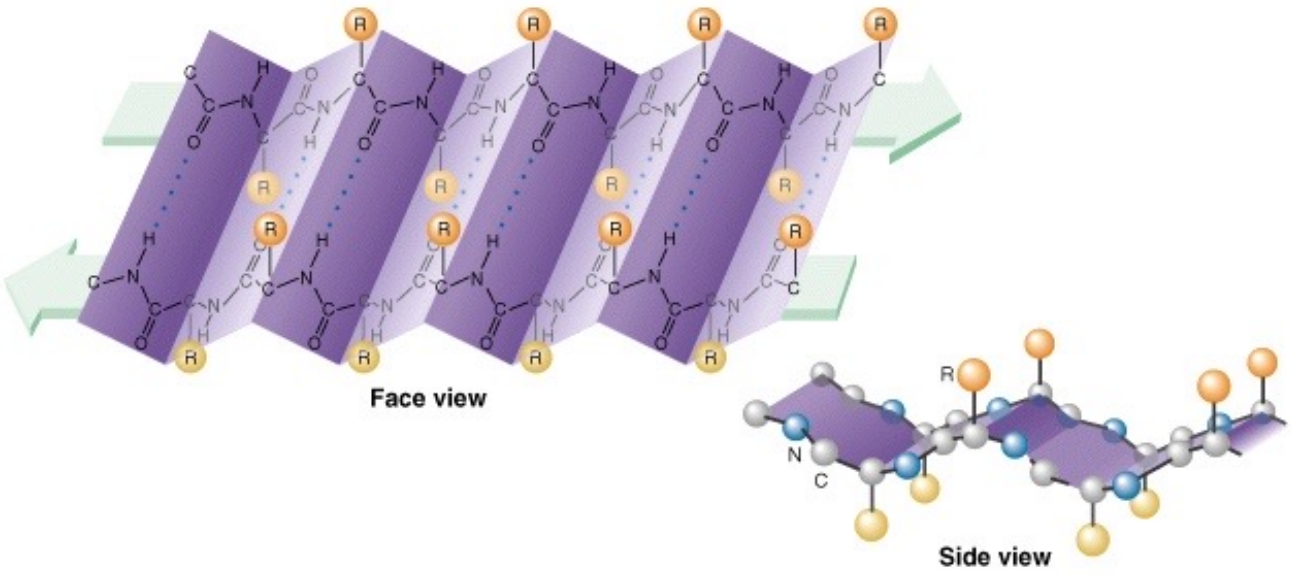


	α-helix	3_{10}-helix	π-helix
Vodíkové můstky	n ... n+4	n ... n+3	n ... n+5
Residua na otáčku	3,6	3	4,4
Vinutí (Å na 1 AK)	1,5	2	1,15

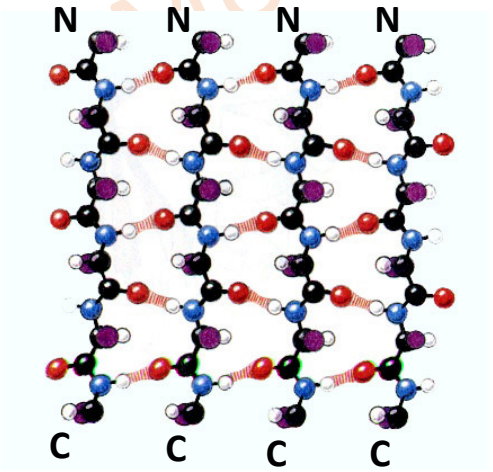
Skládaný list (extended β-sheet)

2D

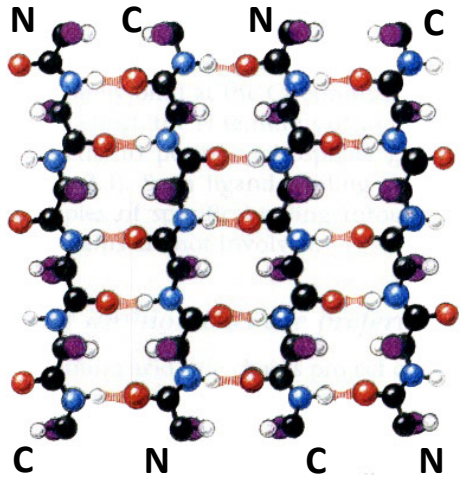
- Paralelní, antiparalelní, mix



Paralelní



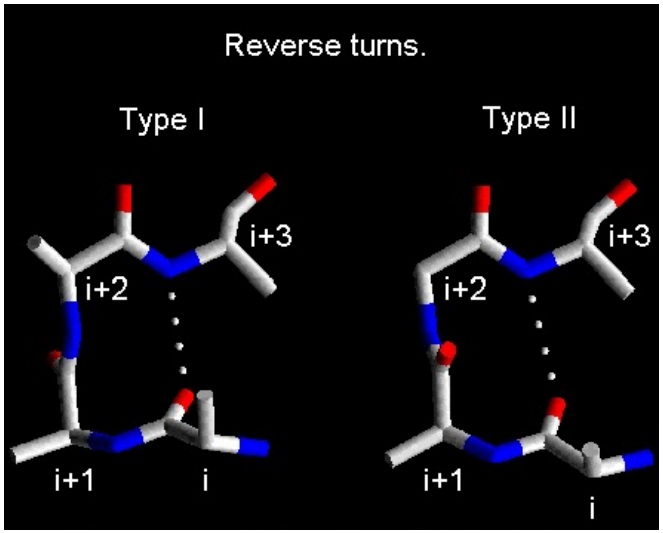
Antiparalelní



Ostatní

2D

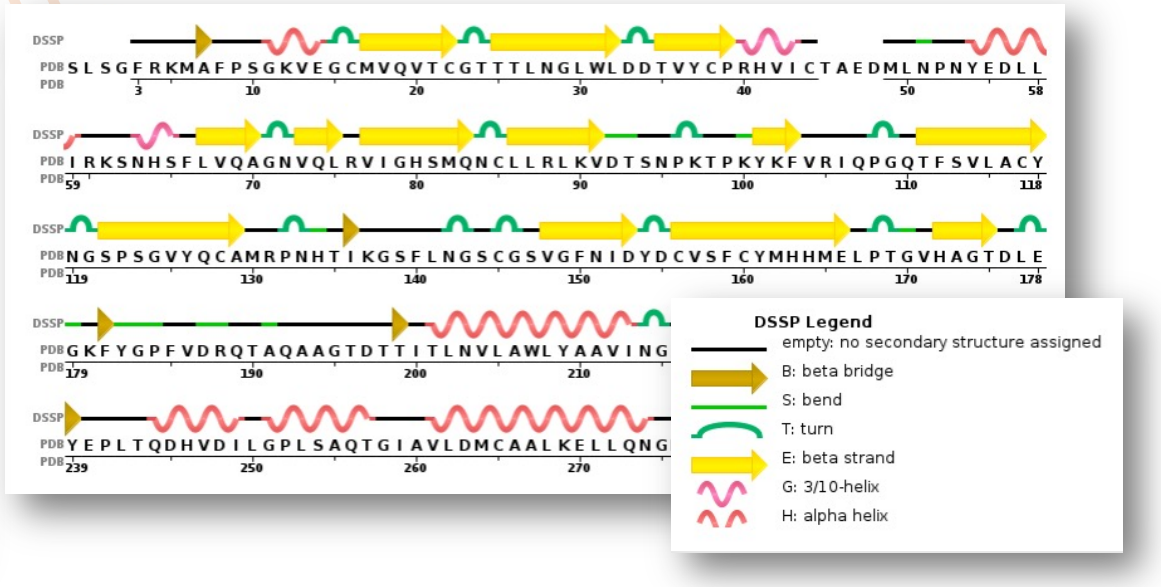
- Úseky které nespadají do kategorií helix nebo list
- **Kombinace** povolených torzních úhlů
- **Nestabilní** konformace
- **Nestandardní** konformace (glycin, prolin)
- **Otáčky** (turns), „**náhodné klubko**“ (**random coil**)



2D

Znázornění 2D struktury

- **Písmeny** – H (helix), E (extended sheet), C (coil)
- **Barevně** – např. červená (helix), žlutá (skládaný list)
- **Grafickými elementy** – spirála/válec (helix), plochá šipka (skládaný list), linka (ostatní)



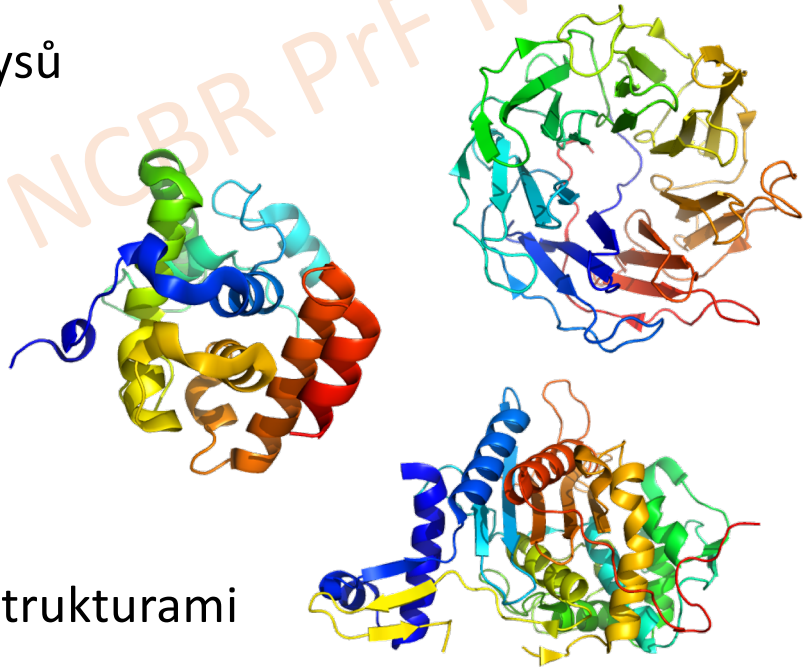
2D

Dělení proteinů dle 2D struktury

Zejména pro účely klasifikace, hledání společných rysů

Každý protein obsahuje mj. smyčky a ohyby

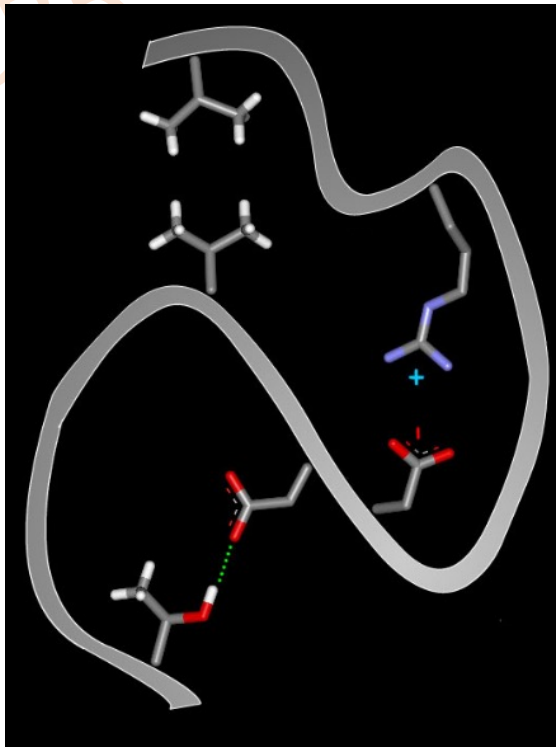
- Jen **α** struktury
- Jen **β** struktury
- **α/β**
  - Motivy kombinující α i β struktury
- **α + β**
  - Oddělené domény tvořené jen α nebo jen β strukturami
- **Malé proteiny**
  - Speciální případy, např. obsahující ionty kovů, stabilizované disulfidickými můstky



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**

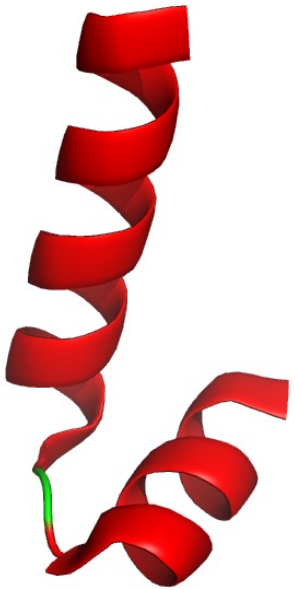


Od 2D ke 3D

- **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)

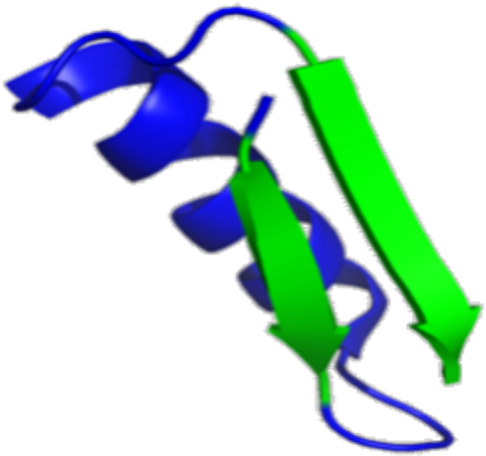
Jednoduché motivy

Helix-turn-helix



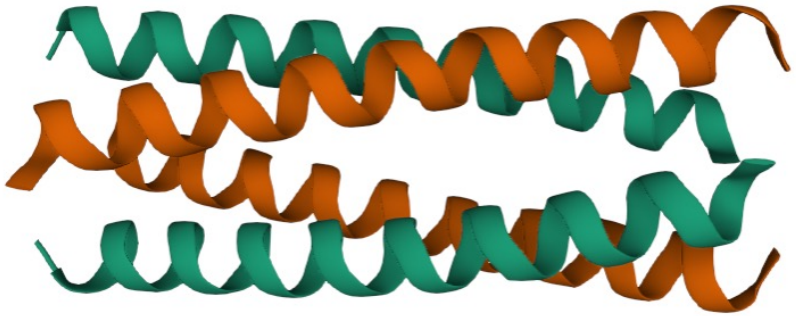
β-vlásenka

β-α-β

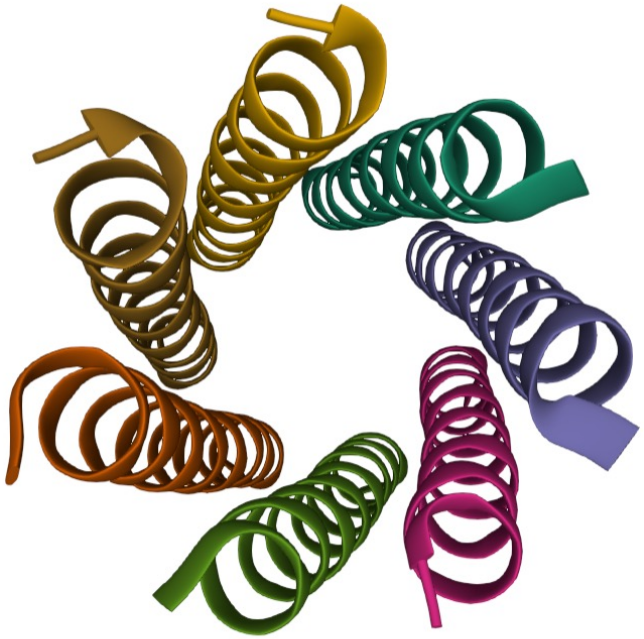


Složené α-motivy/foldy

4-helix bundle

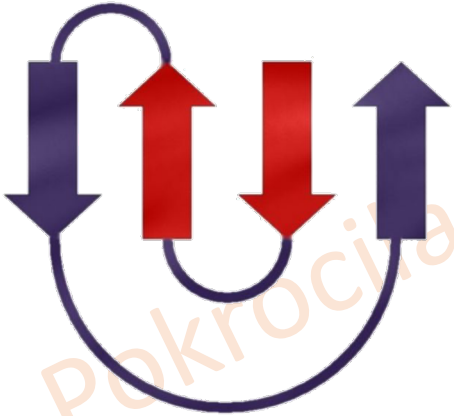


7-helix barel



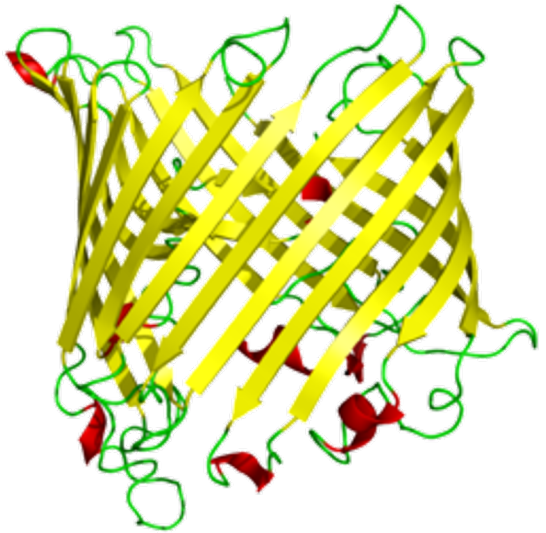
Složené β-motivy/foldy

Řecký klíč



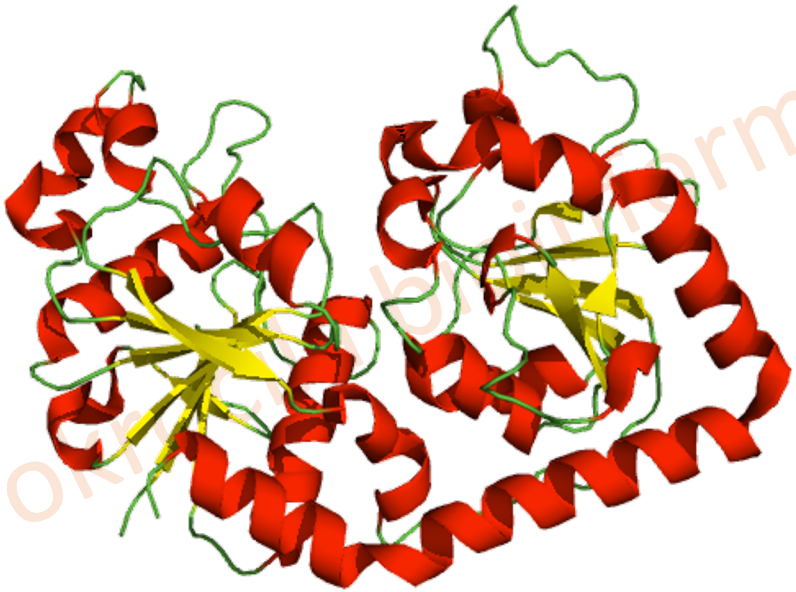
β-meandr

β-barel

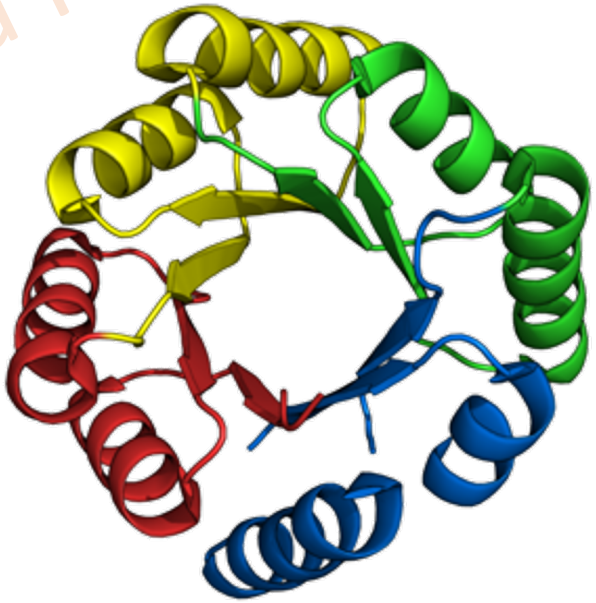


Složené α/β-motivy/foldy

Rossmanův fold



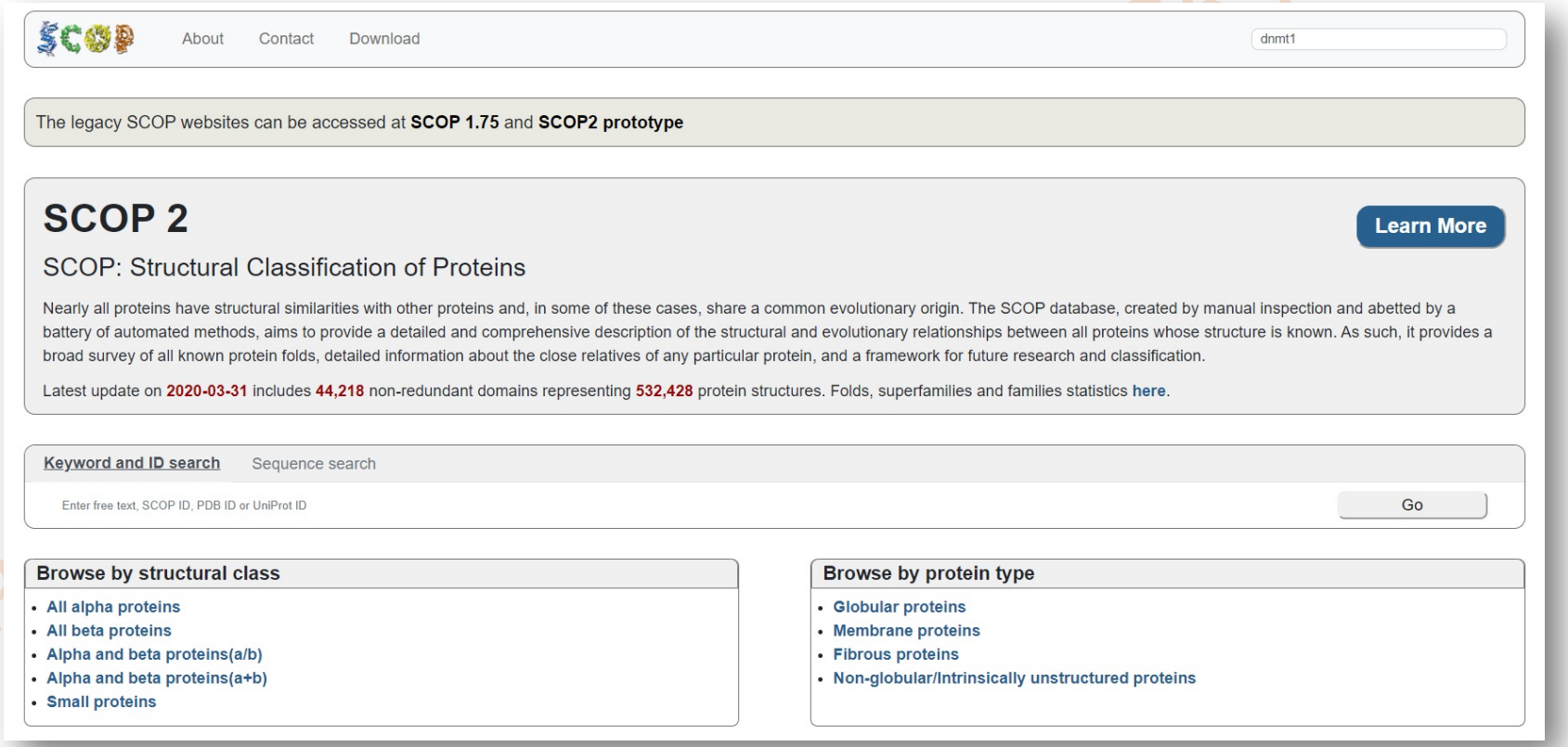
TIM-barel



Structural classification of proteins (SCOP)

2D

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

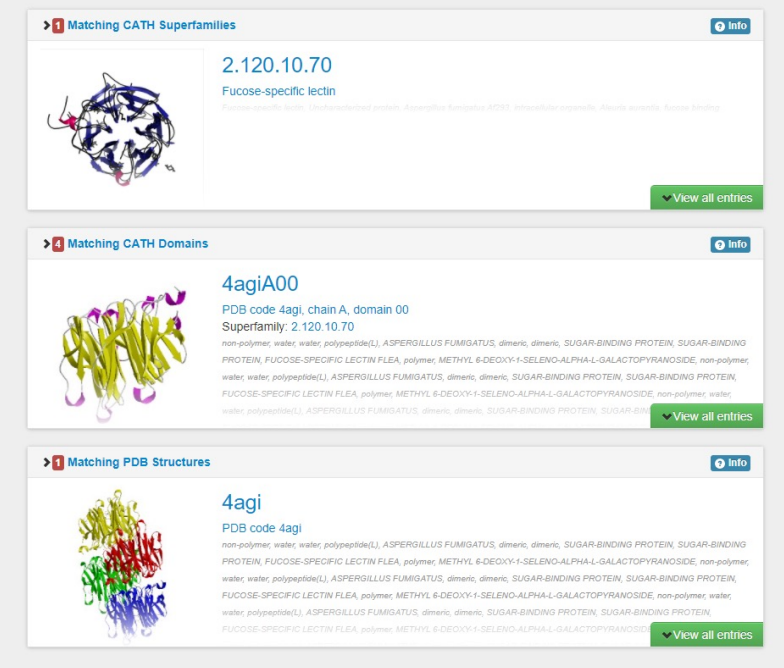


2D

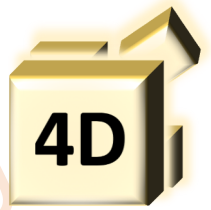
CATH – Protein structure classification database

- Domény jsou klasifikovány podle CATH hierarchie
  - Třída (**C**lass)
    - Podle sekundární struktury
    - Jen α, jen β, α i β, minimum sekundární struktury
  - **A**rchitektura
    - 3D uspořádání sekundární struktury
  - **T**opologie/fold
    - Jak jsou prvky sekundární struktury uspořádané za sebou
  - **H**omologní nadrodina
    - V případě, že jsou domény evolučně příbuzné (homologní proteiny)

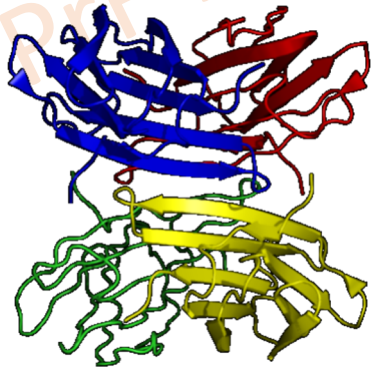
https://www.cathdb.info/

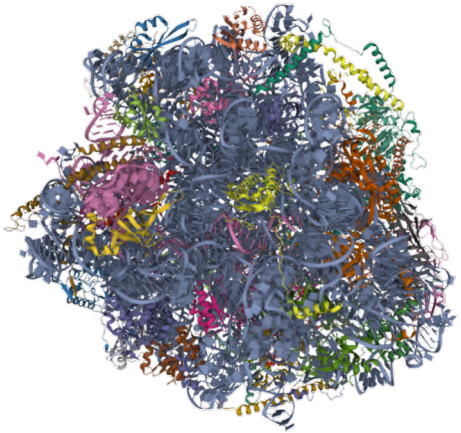


Kvartérní struktura



- 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,...





Způsob uložení 3D (4D) strukturních dat

- Veřejně dostupné **databáze**
  - **Protein Data Bank** (PDB), Biological Magnetic Resonance Data Bank, EMDataBank
- **Koordináty** atomů, doplňkové informace (**meta data**)
- Definovaný **formát**
  - PDB
  - mmCIF

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)

Proč predikovat strukturu?

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

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

1. Generace – *ab inicio*

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

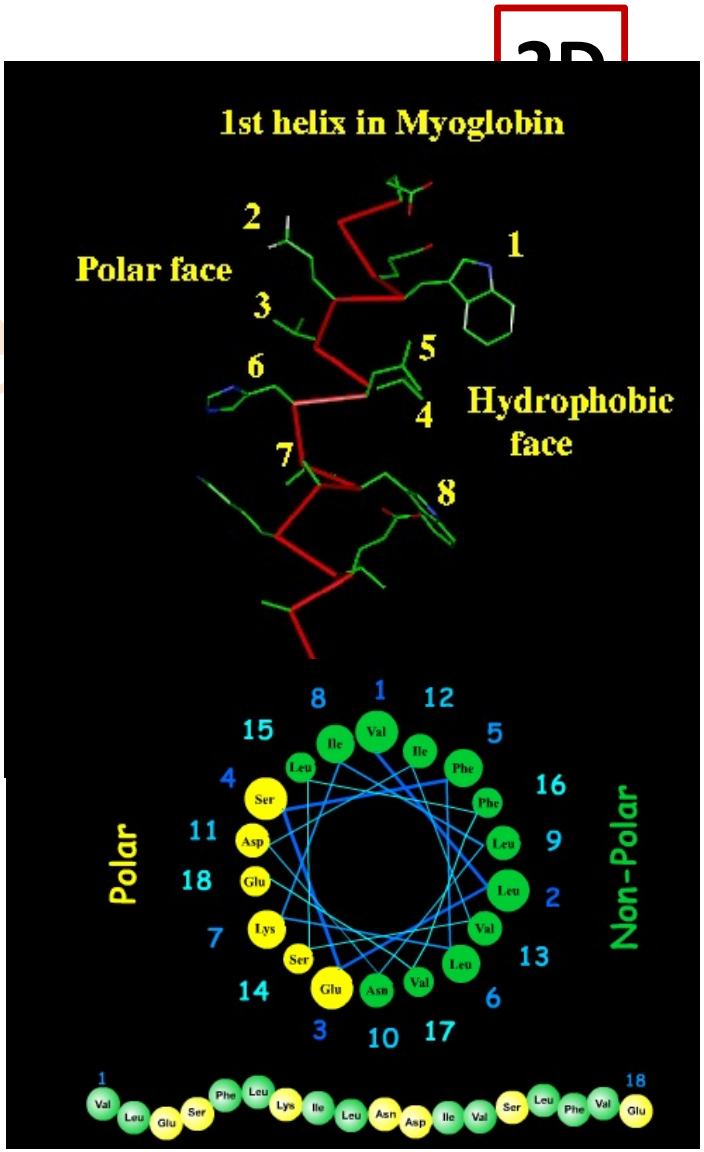
Amino Acid	(α-Helix)	P (β-Strand)	P (Turn)
Alanine	1.42	0.83	0.66
Arginine	0.98	0.93	0.95
Asparagine	0.67	0.89	1.56
Aspartic acid	1.01	0.54	1.46
Cysteine	0.70	1.19	1.19
Glutamic acid	1.51	0.37	0.74
Glutamine	1.11	1.11	0.98
Glycine	0.57	0.75	1.56
Histidine	1.00	0.87	0.95
Isoleucine	1.08	1.60	0.47
Leucine	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

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

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

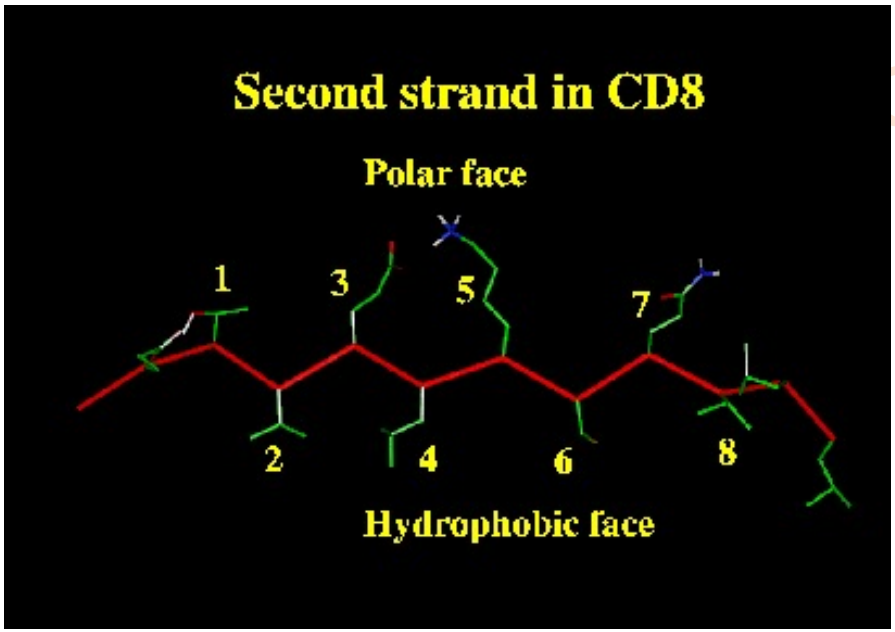
- **Transmembránový** helix
  - Všechny aminokyseliny hydrofobní

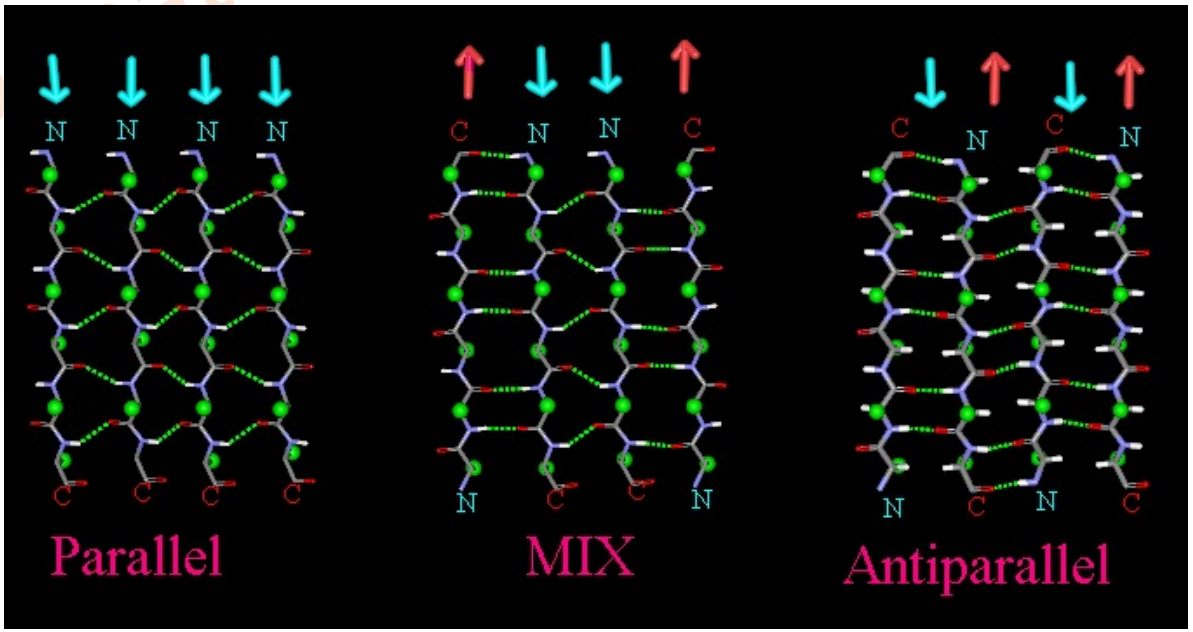


Typické znaky β –list (musí být stabilizován jinou částí polypeptidového řetězce!)

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

ELKAHIRVDLTLQ α

ELKAHIRVDLTLQ

ELKAHIRVDLTLQ β

Polární Nepolární

α-šroubovice nebo β-list?

2D

ELKAHIRVDLTLQ

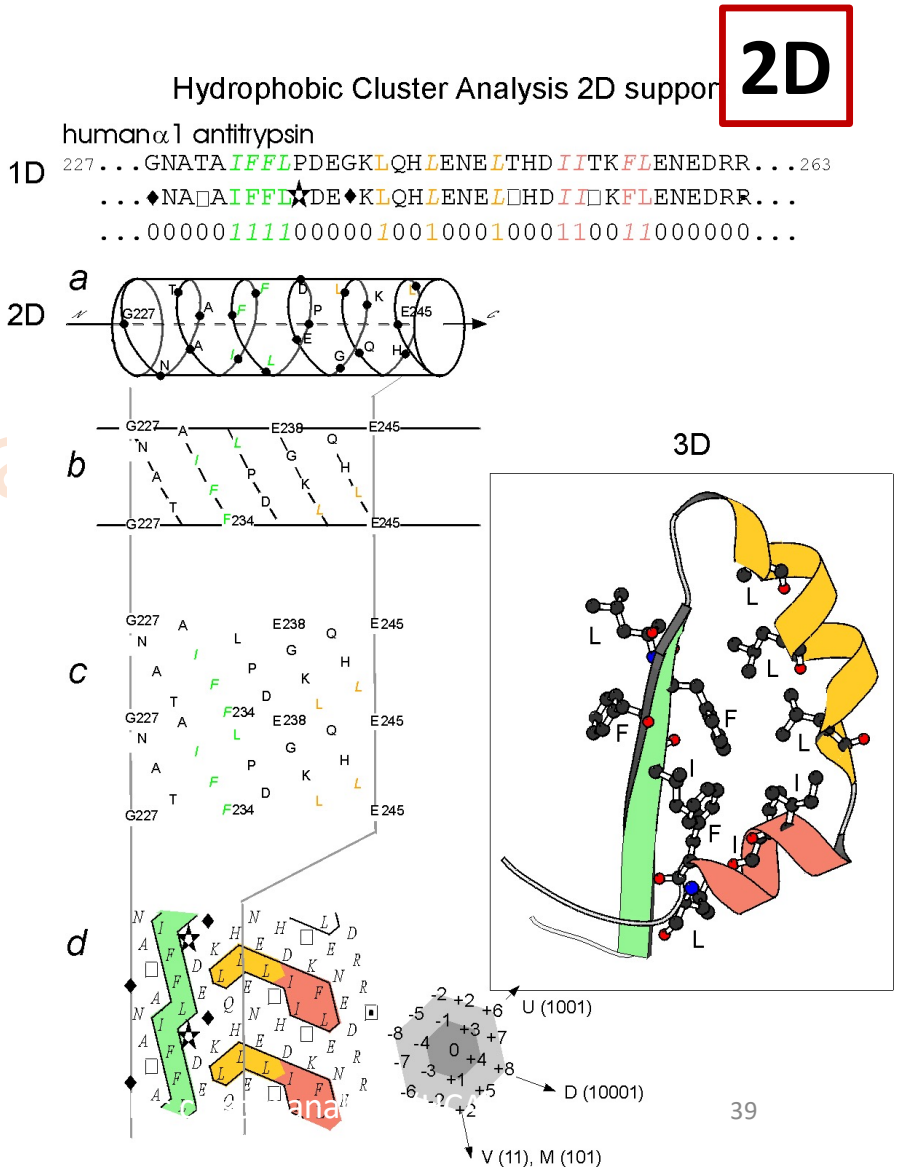
ELKAHIRVDLTLQ α ✗

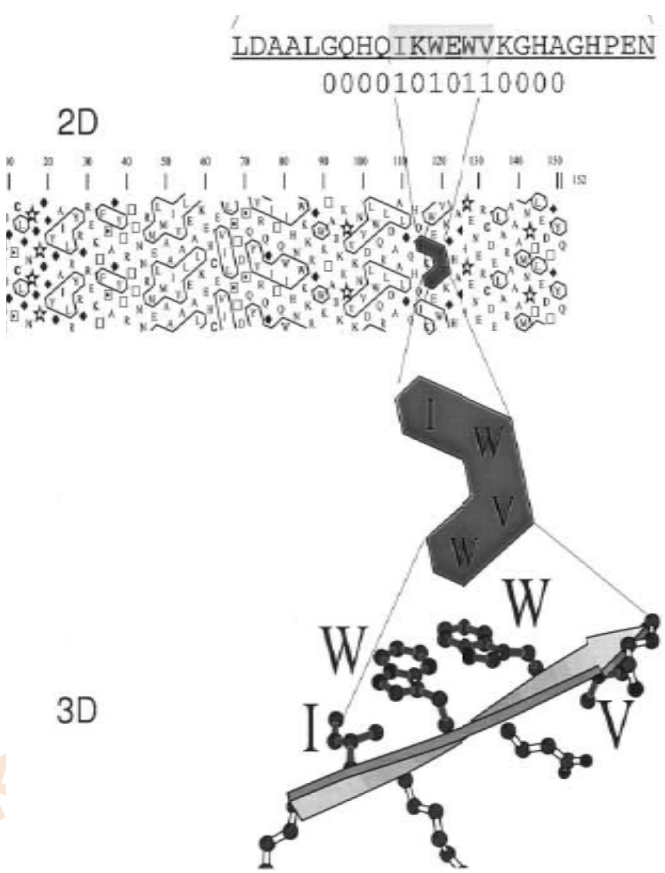
ELKAHIRVDLTLQ β ✓

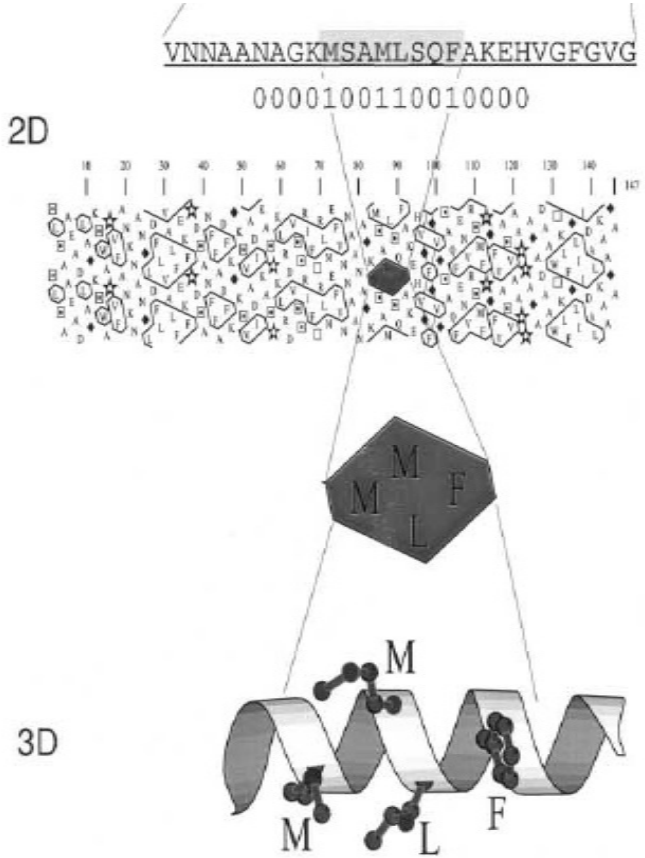
Polární Nepolární

Analýza hydrofobních klastrů (HCA)

- 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



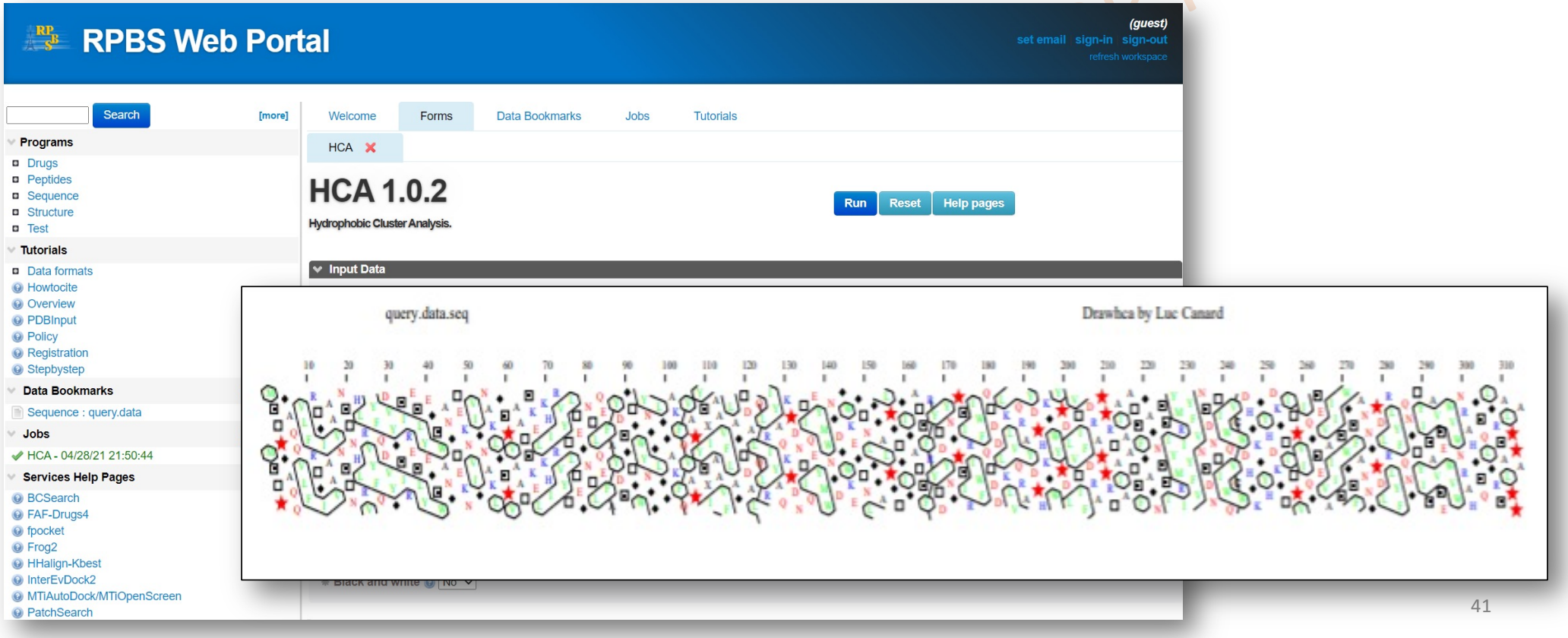

LDAALGQHQIKWEWVKGHAGHPEN
00001010110000
2D
I
W
V
W
W
W
I
V
3D


VNNAANAGKMSAMLSQFAKEHVGFGVG
0000100110010000
2D
M
M
F
L
M
M
L
F
3D

2D

RPBS Web Portal – HCA

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



Predikce sekundární struktury

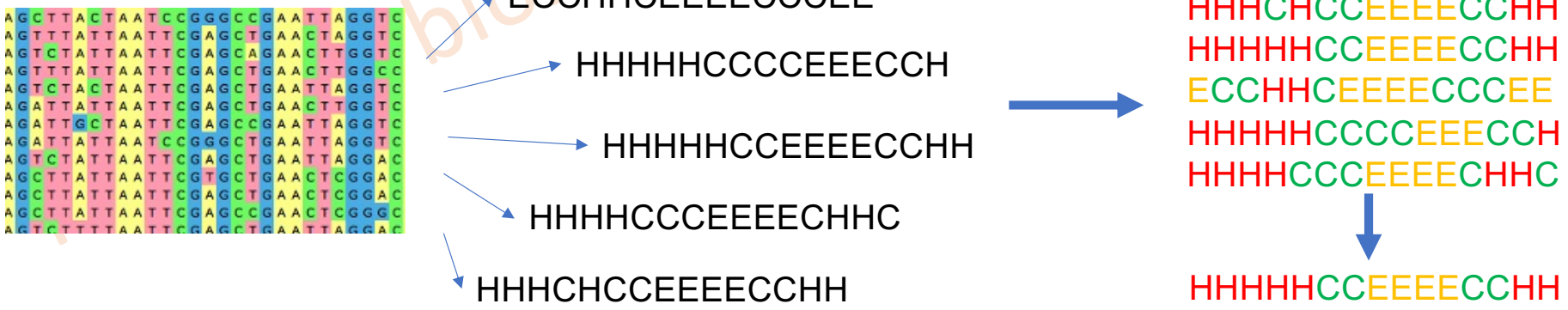
2D

- 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

2D

Metody založené na homologii (*Homology-based*)

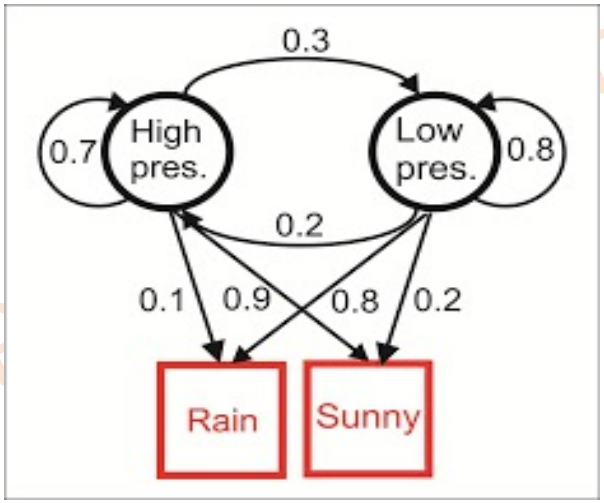
- Vychází z předpokladu, že 2D struktura je více konzervovaná než sekvence
  1. Multiple sequence alignment
  2. Predikce sekundárních struktur pro každou sekvenci zvlášť
  3. Porovnání predikovaných sekundárních struktur s alignmentem
  4. Konsenzus sekundární struktury

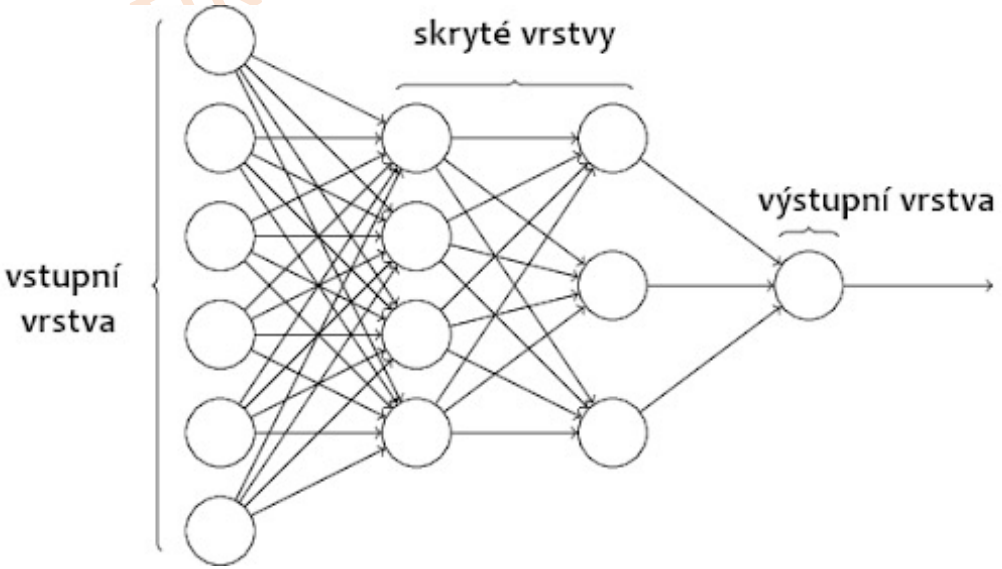


Metody strojového učení (*Machine learning*)

2D

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

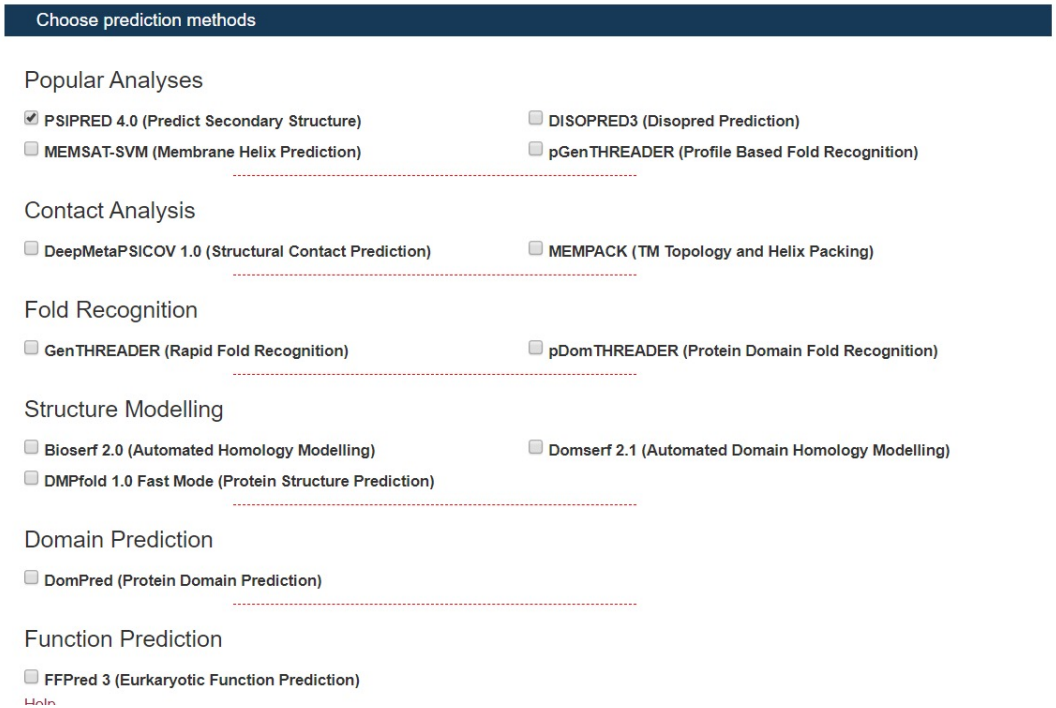




PSIPRED

2D

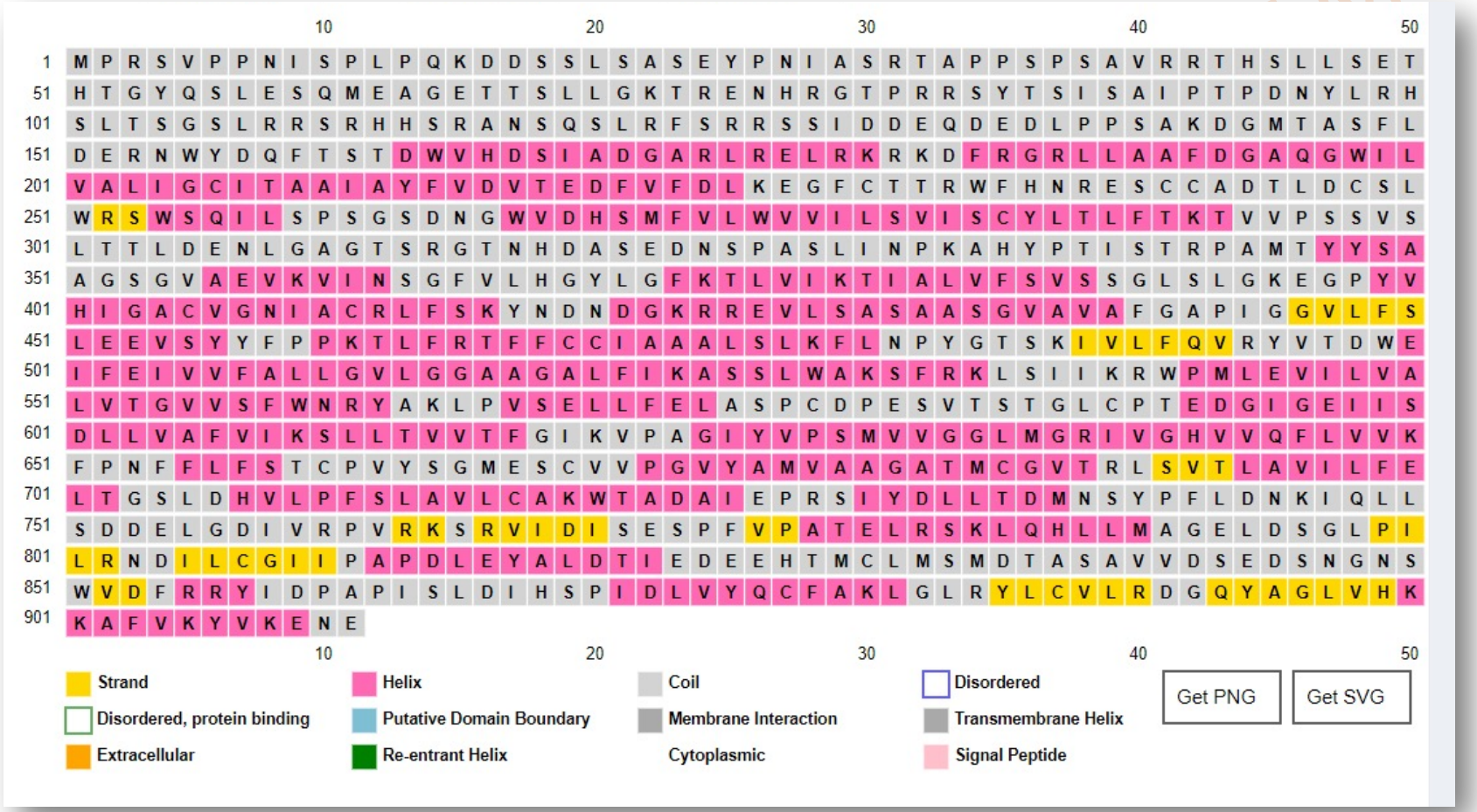
- Predikce sekundární struktury pomocí 2 neuronových sítí
- Časově náročnější
- Ve srovnání s většinou programů na predikci sekundární struktury má lepší výsledky



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

PSIPRED

2D


10 20 30 40 50
1 M P R S V P P N I S P L P Q K D D S S L S A S E Y P N I A S R T A P P S P S A V R R T H S L L S E T
51 H T G Y Q S L E S Q M E A G E T T S L L G K T R E N H R G T P R R S Y T S I S A I P T P D N Y L R H
101 S L T S G S L R R S R H H S R A N S Q S L R F S R R S S I D D E Q D E D L P P S A K D G M T A S F L
151 D E R N W Y D Q F T S T D W V H D S I A D G A R L R E L R K R K D F R G R L L A A F D G A Q G W I L
201 V A L I G C I T A A I A Y F V D V T E D F V F D L K E G F C T T R W F H N R E S C C A D T L D C S L
251 W R S W S Q I L S P S G S D N G W V D H S M F V L W V V I L S V I S C Y L T L F T K T V V P S S V S
301 L T T L D E N L G A G T S R G T N H D A S E D N S P A S L I N P K A H Y P T I S T R P A M T Y Y S A
351 A G S G V A E V K V I N S G F V L H G Y L G F K T L V I K T I A L V F S V S S G L S L G K E G P Y V
401 H I G A C V G N I A C R L F S K Y N D N D G K R R E V L S A S A A S G V A V A F G A P I G G V L F S
451 L E E V S Y Y F P P K T L F R T F F C C I A A A L S L K F L N P Y G T S K I V L F Q V R Y V T D W E
501 I F E I V V F A L L G V L G G A A G A L F I K A S S L W A K S F R K L S I I K R W P M L E V I L V A
551 L V T G V V S F W N R Y A K L P V S E L L F E L A S P C D P E S V T S T G L C P T E D G I G E I I S
601 D L L V A F V I K S L L T V V T F G I K V P A G I Y V P S M V V G G L M G R I V G H V V Q F L V V K
651 F P N F F L F S T C P V Y S G M E S C V V P G V Y A M V A A G A T M C G V T R L S V T L A V I L F E
701 L T G S L D H V L P F S L A V L C A K W T A D A I E P R S I Y D L L T D M N S Y P F L D N K I Q L L
751 S D D E L G D I V R P V R K S R V I D I S E S P F V P A T E L R S K L Q H L L M A G E L D S G L P I
801 L R N D I L C G I I P A P D L E Y A L D T I E D E E E H T M C L M S M D T A S A V V D S E D S N G N S
851 W V D F R R Y I D P A P I S L D I H S P I D L V Y Q C F A K L G L R Y L C V L R D G Q Y A G L V H K
901 K A F V K Y V K E N E
Strand
Helix
Coil
Disordered
Disordered, protein binding
Putative Domain Boundary
Membrane Interaction
Transmembrane Helix
Extracellular
Re-entrant Helix
Cytoplasmic
Signal Peptide
Get PNG
Get SVG

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**

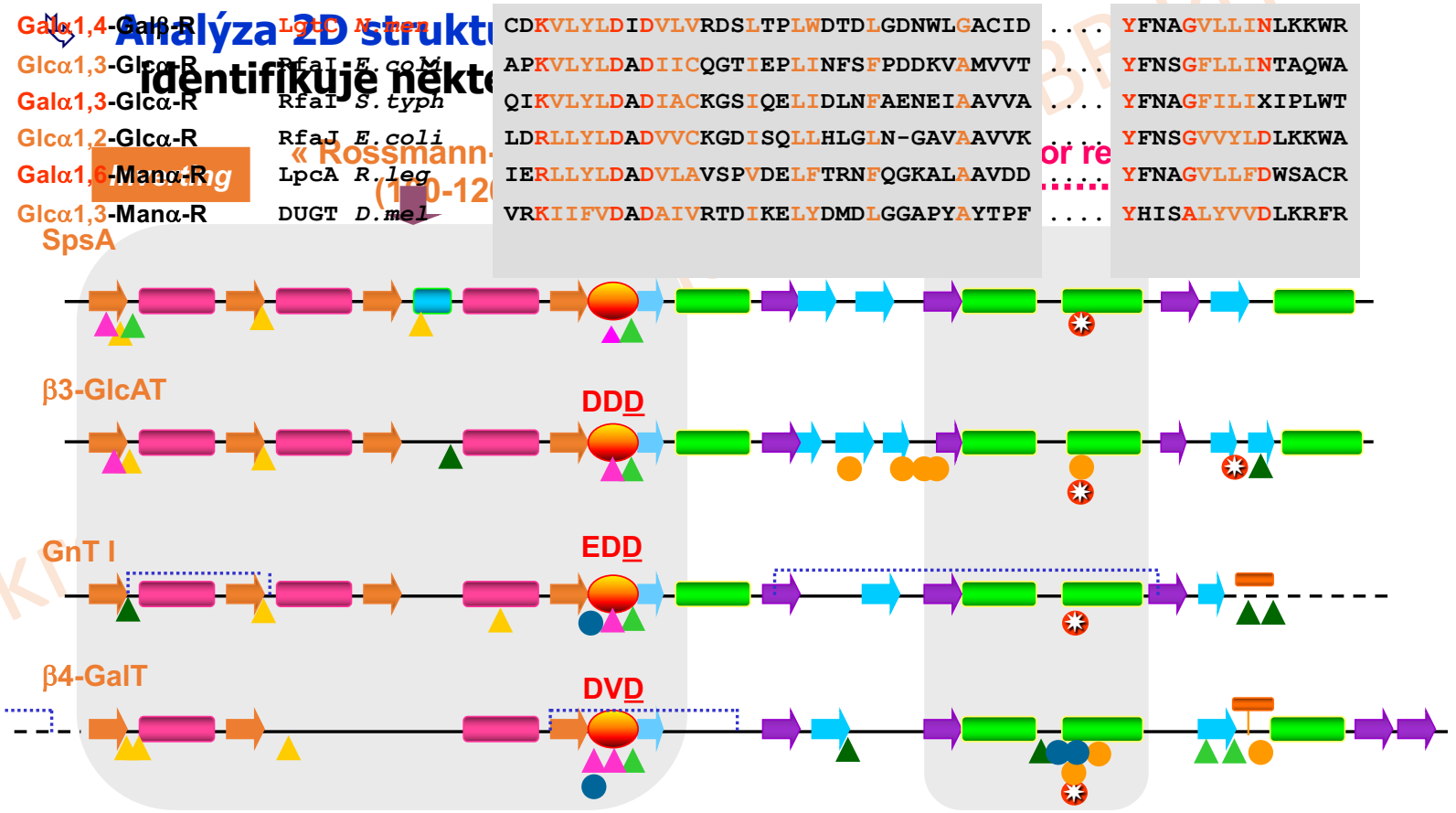
Predikce terciární struktury

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

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

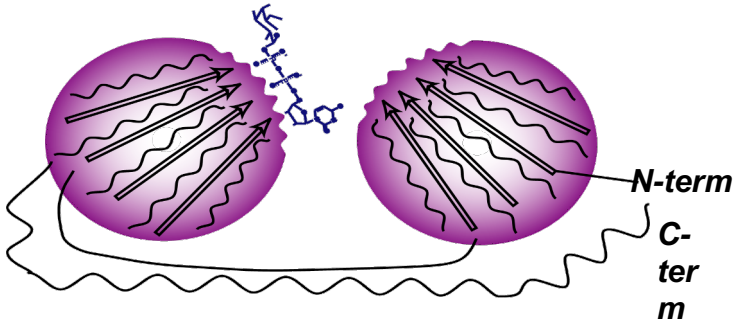
Metody pro predikci funkce

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



Dvě pozorované topologie 3D struktur glykosyltransferas

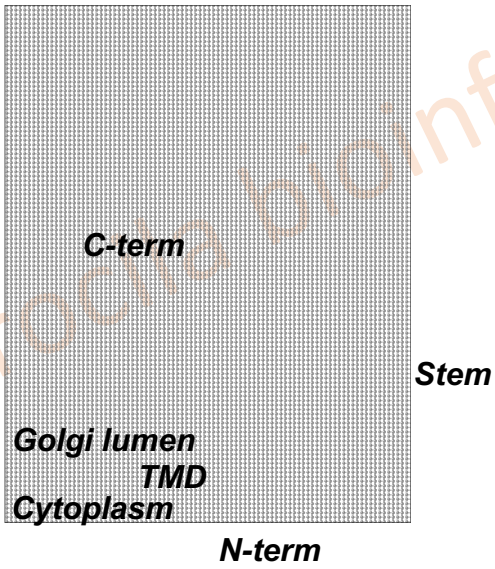
BGT-fold



(Procaryotes/Phage)

β-GlcT (BGT, phage T4)	n.c.	inv
β4-GlcNAcT (MurG, *E.coli*)	GT28	inv
β-GlcT (GtfB, *M. orientalis*)	GT1	inv

SpsA-fold



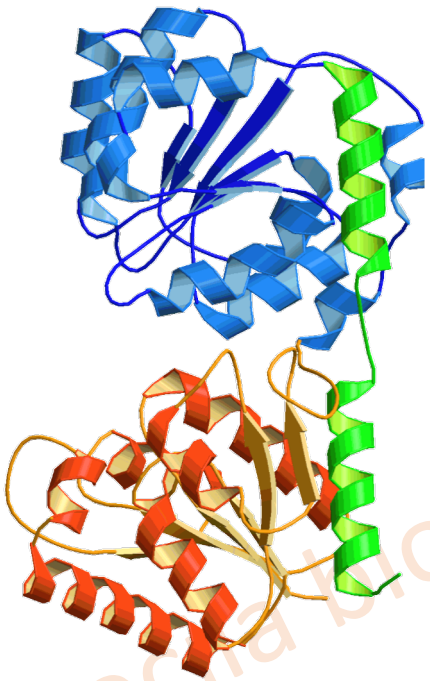
(Procaryotes)

SpsA (B. subtilis)	GT2	inv
α4-GalT (LgtC, N.meningitis)	GT8	ret

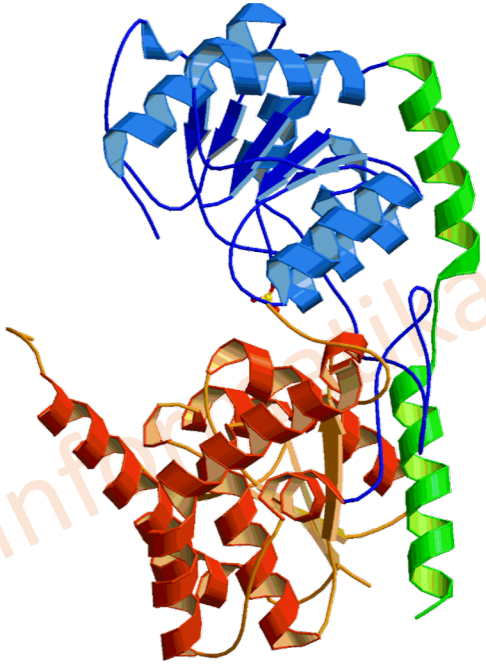
(Eucaryotes)

β4-GalT1 (bovine)	GT7	inv
β2-GlcNAcT (GnT I, rabbit)	GT13	inv
β3-GlcAT I (human)	GT43	inv
α3-GalT (bovine)	GT6	ret
Glycogenin (rabbit)	GT8	ret
α3-GalNacT (GTA, human)	GT6	ret
α3-GalT (GTB, human)	GT6	ret

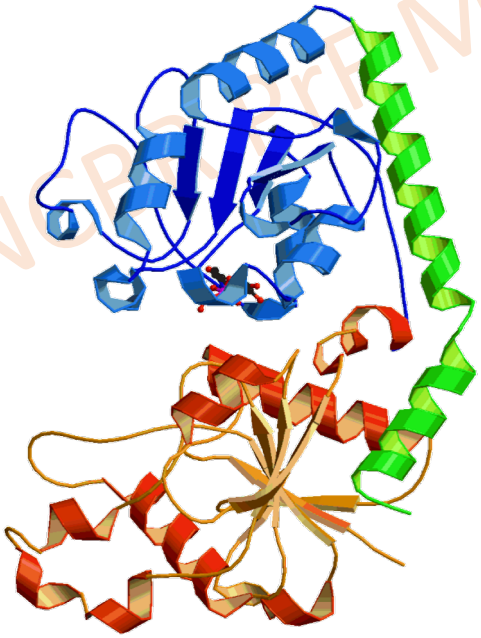
Nadrodina s BGT foldem



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



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

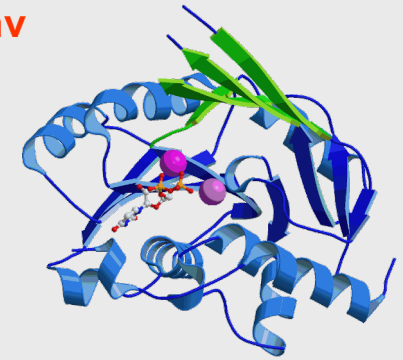


BGT (β-GlcT)
n.c.
Phage T4
Vrielink *et al*, 1994

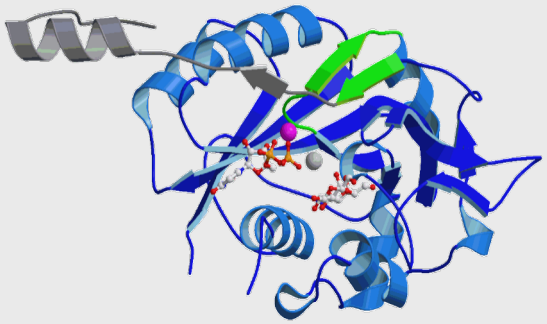
Nadrodina s SpsA foldem

Společná NBD

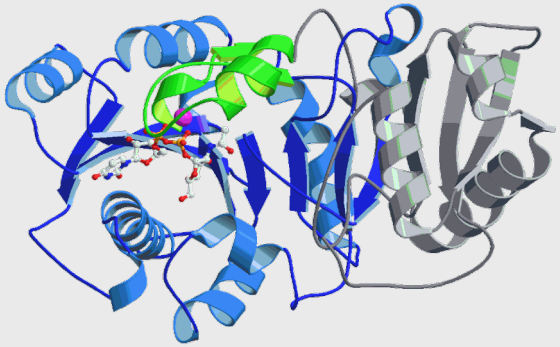
Inv



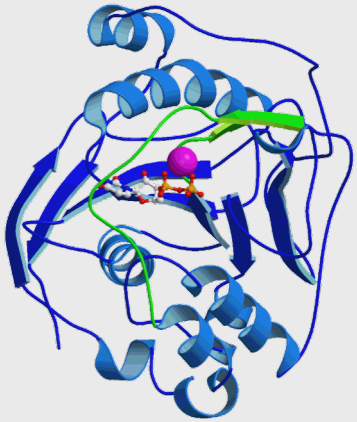
SpsA [GT2]
Charnok *et al*, 1999, 2001



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

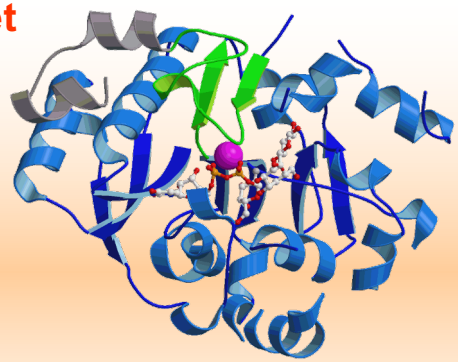


Rabbit GnT I [GT13]
Ünligil *et al*, 2000

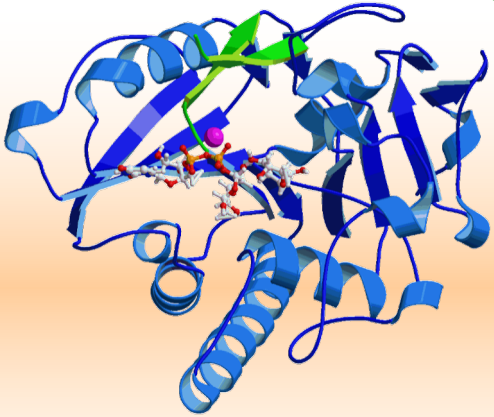


Bovine β4-GalT [GT7]
Gastinel *et al*, 1999
Ramakrishnan *et al*, 2001, 2002

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

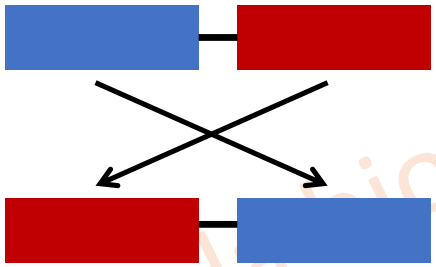
- ***Ab initio***
- **Homologní modelování**
- **Threading** („navlékání“)

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 multidoménových struktur je potřeba aminokyselinovou sekvenci rozdělit na jednotlivé domény a analyzovat je separátně

PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKV
PESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIG
GGGERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGR
PSRLGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG

ERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSR
LGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLGPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRDGKLQ
VILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAM
HIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGGKLAAALEIKRASQPELAPEDPEDVEHHHHHH



```
#
#=======================================

EMBOSS_001         1 --------------------------------------------------      0

EMBOSS_001         1 ERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQ     50

EMBOSS_001         1 --------------------------------------------------      0

EMBOSS_001        51 NLGTKVLDSGNGRVRVIVMANGRPSRLGSRQVDIFKKSYFGIIGSEDGAD    100

EMBOSS_001         1 ---------------PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRD     35
                                    |||||||||||||||:||||||||||||||||||
EMBOSS_001       101 DDYNDGIVFLNWPLGPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRD    150

EMBOSS_001        36 GKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGR     85
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       151 GKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGR    200

EMBOSS_001        86 MPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQ    135
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       201 MPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQ    250

EMBOSS_001       136 GSGNQGAETGGTGAGNIGGGGERDGTFNLPPHIKFGVTALTHAANDQTID    185
                     ||||||||||||||||||||
EMBOSS_001       251 GSGNQGAETGGTGAGNIGGGG----------------------------    271

EMBOSS_001       186 IYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSRLGS    235
                             |.||:          |..|                  |:|
EMBOSS_001       272 -------KLAAA----------LEIK----------------RAS----    283

EMBOSS_001       236 RQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG    271
                      |.:         ::.||..|.:.:.       ...
EMBOSS_001       284 -QPE---------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

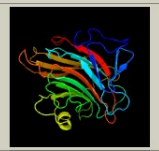
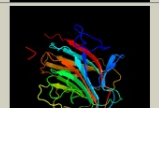
https://zhanglab.ccmb.med.umich.edu/I-TASSER/

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/

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c1r1zB_	Alignment		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-533 reveals a novel metal binding site and conformational4 changes associated with calcium ion binding Run Investigator
2	c2a6yA_	Alignment		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 Run Investigator
3	d2a6za1	Alignment		100.0	20	**Fold:**Concanavalin A-like lectins/glucanases **Superfamily:**Concanavalin A-like lectins/glucanases **Family:**Lectin leg-like Run Investigator
4	c2dupB_	Alignment		100.0	42	**PDB header:**protein transport **Chain:** B: **PDB Molecule:**vesicular integral-membrane protein vip36; **PDBTitle:** crystal structure of vip36 exoplasmic/lumenal domain, metal-free form

Phyre2

ARDLVIPMIYCGHGY

User sequence

Homologous sequences

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

Phyre2

User sequence

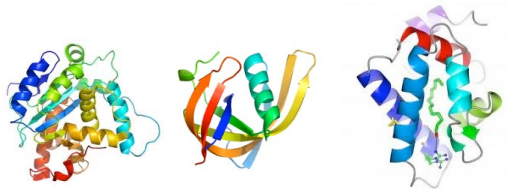
PSI-Blast

Hidden Markov model

Capture the mutational propensities at each position in the protein

An evolutionary fingerprint

Phyre2

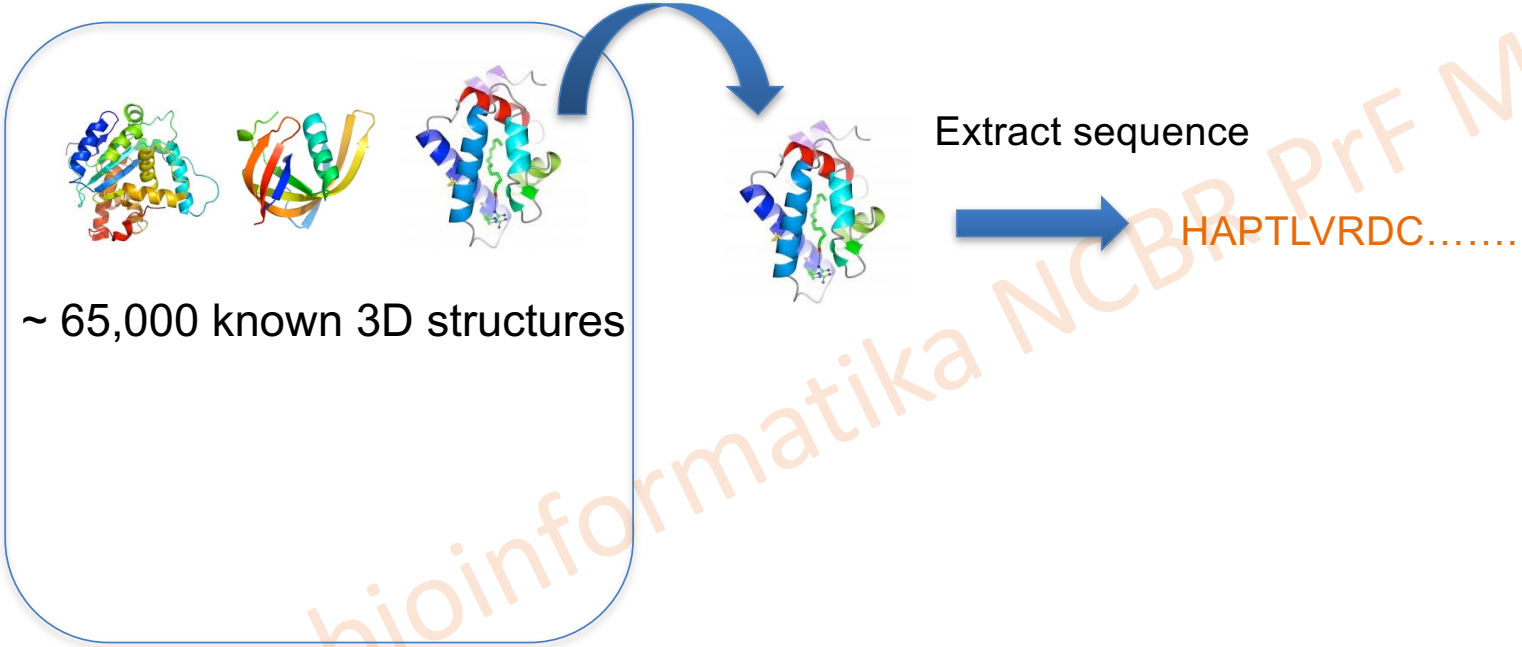


~ 65,000 known 3D structures

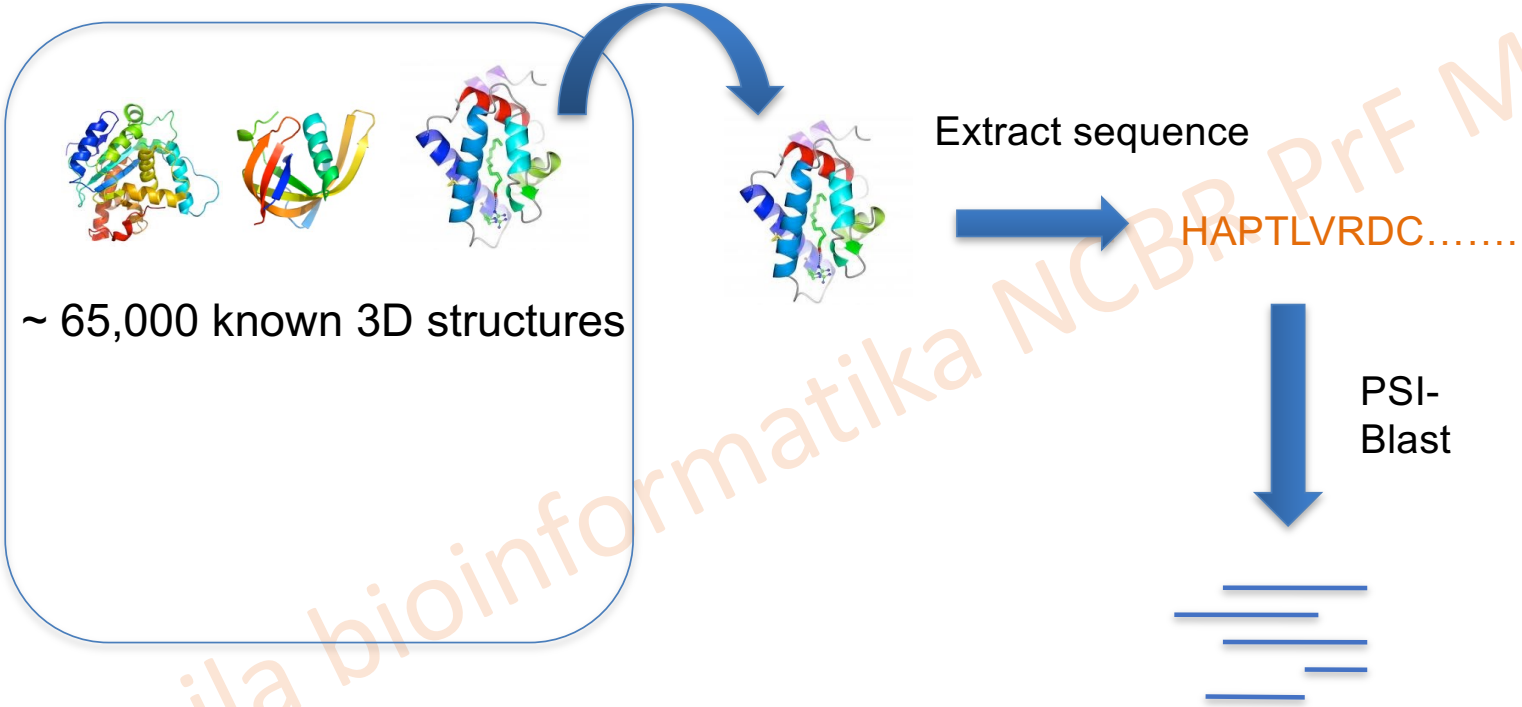
Phyre2

~ 65,000 known 3D structures

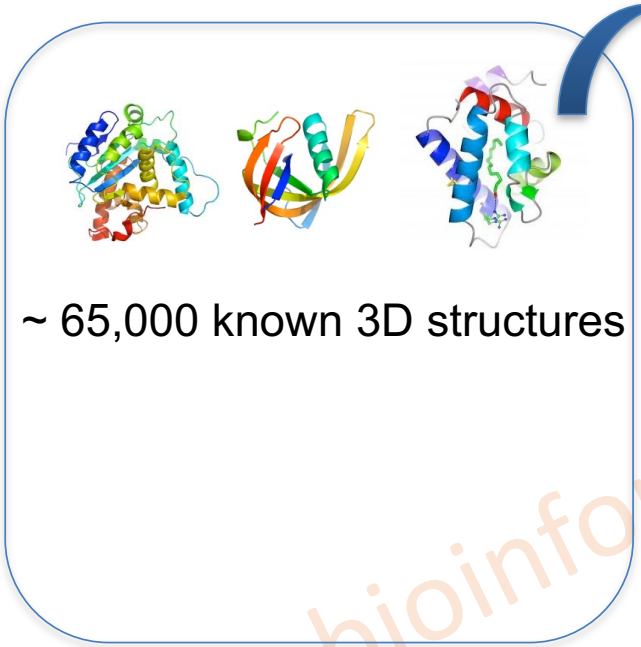
Phyre2



Phyre2



Phyre2



~ 65,000 known 3D structures

Extract sequence

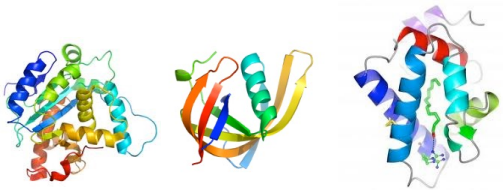
HAPTLVRDC.....

PSI-
Blast

HMM

Hidden Markov model
for sequence of KNOWN structure

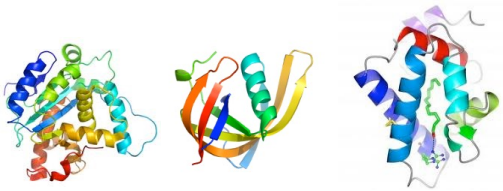
Phyre2



~ 65,000 known 3D structures

~ 65,000 hidden Markov models

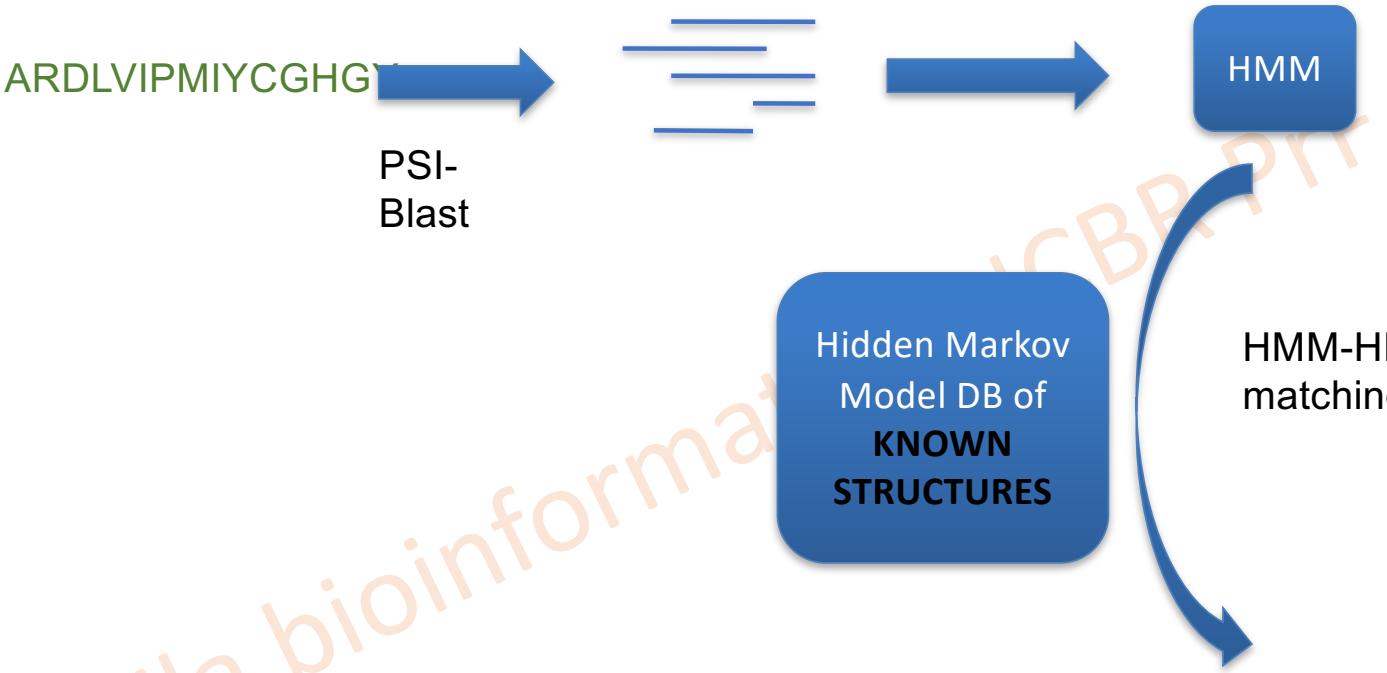
Phyre2



~ 65,000 known 3D structures

Hidden Markov Model
Database of
KNOWN STRUCTURES

Phyre2



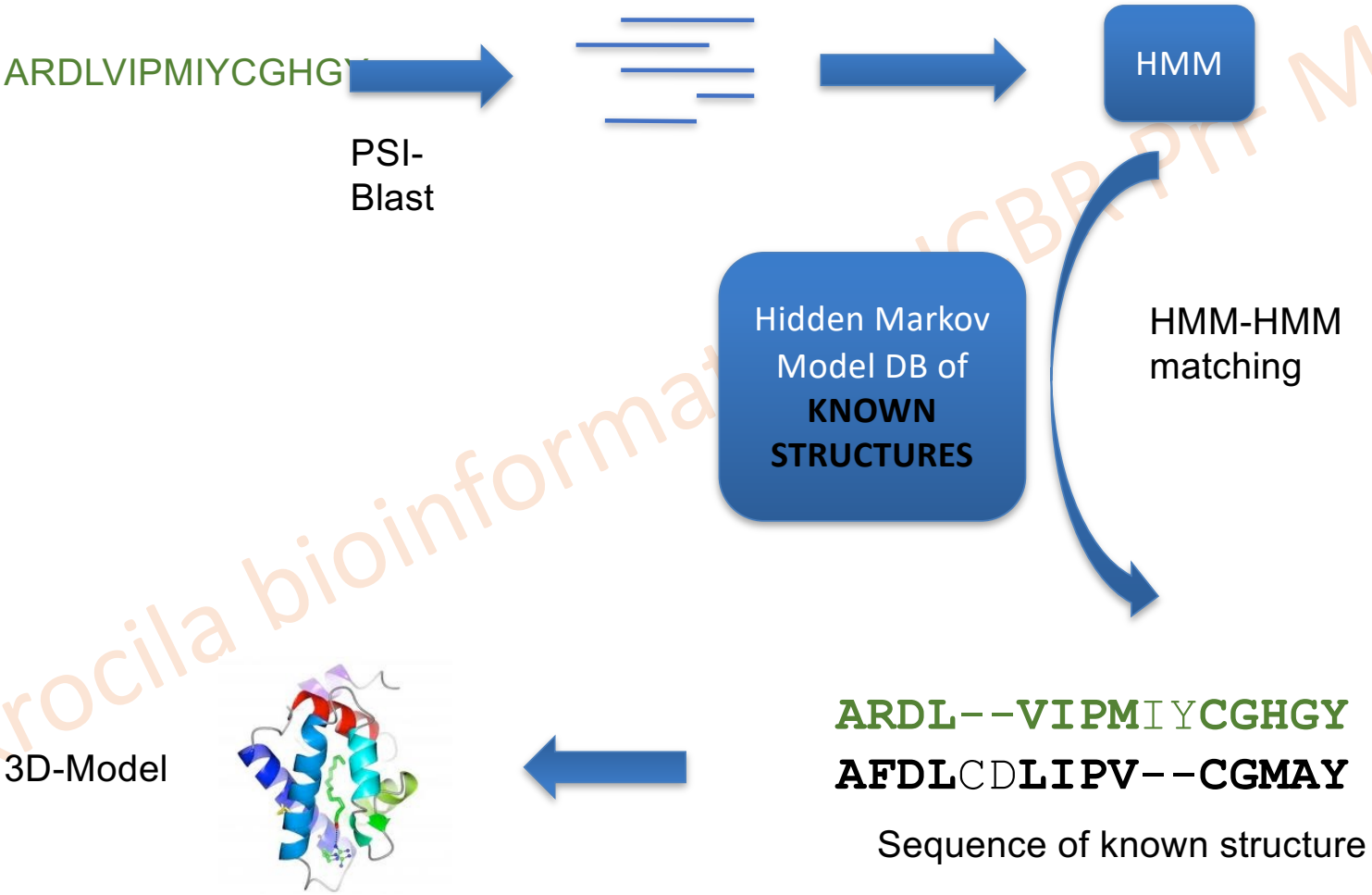
Alignments of user sequence to known structures ranked by confidence.

ARDL--VIPMIY**CGHGY**

AFDLCD**LIPV--CGMAY**

Sequence of known structure

Phyre2



Phyre2

ARDLVIPMIYCGHGY

PSI-Blast

HMM

Hidden Markov Model DB of **KNOWN STRUCTURES**

HMM-HMM matching

Very powerful – able to reliably detect extremely remote homology

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

ARDL--VIPMIYCGHGY
AFDLCDLIPV--CGMAY

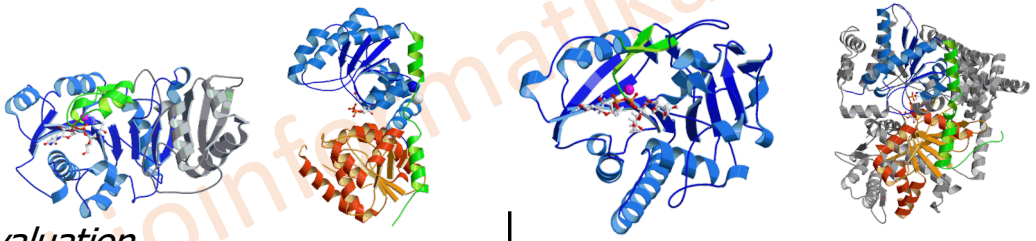
Sequence of known structure

3D-Model

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

SDVDIEAGQTLVQVVNISNGETWVAIQLPAQYRSFDLVFENVSPSTSGSVLVAQMAPQSGGVYGSNYS
GSGWGNDLGGGGFYGYSEAKWMCLWPANRSGPNSKTGIYGTCKLMNLNQSNAVPSVTSNLFAPTAY
KNEPGYANVGGCCQKIRGLASSIQFAFALHGGNVPQNTDTFSGGTIKVYGWN

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

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

Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda

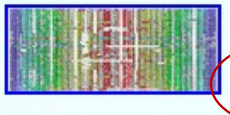
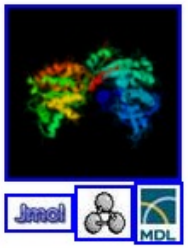
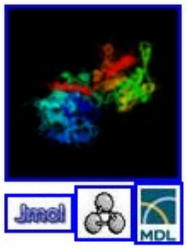
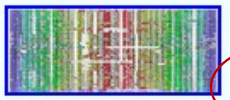
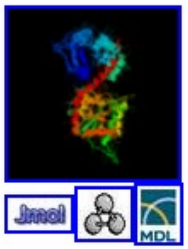
http://www.sbg.bio.ic.ac.uk/phyre/qphyre_output/95cbaa7600a9bfff/summary.html

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

Quickphyre results for job cand____ | Quickphyre results for job synt____ | Quickphyre results for job synt____

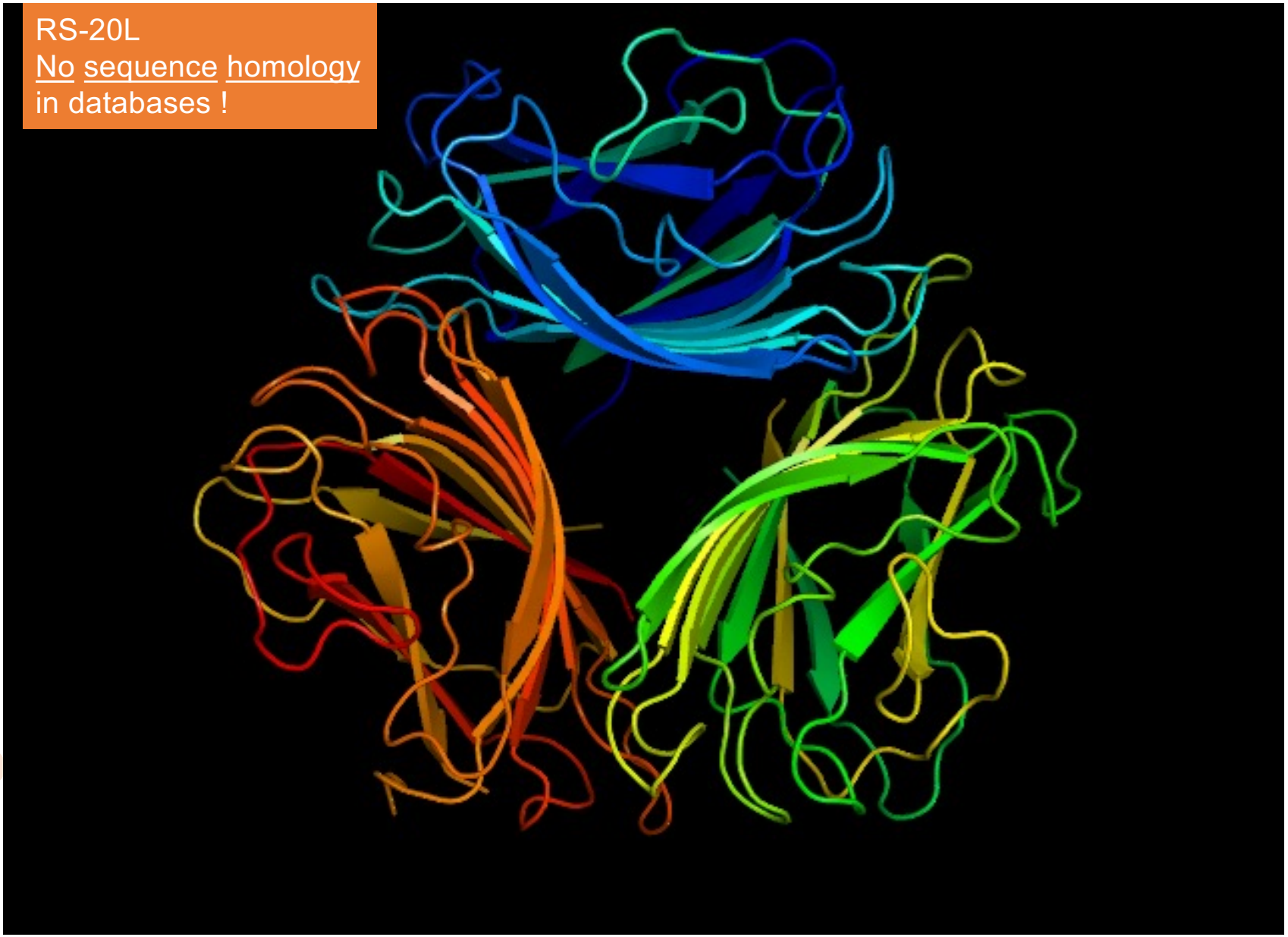
To predict functional residues and GO classification, try ConFunc

Fold Recognition

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

RS-20L
No sequence homology
in databases !



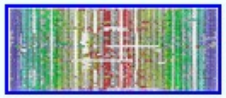
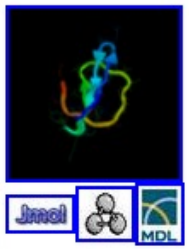
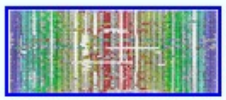
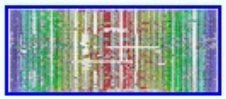
Quickphyre results for job rs20____ - Mozilla Firefox

Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda

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

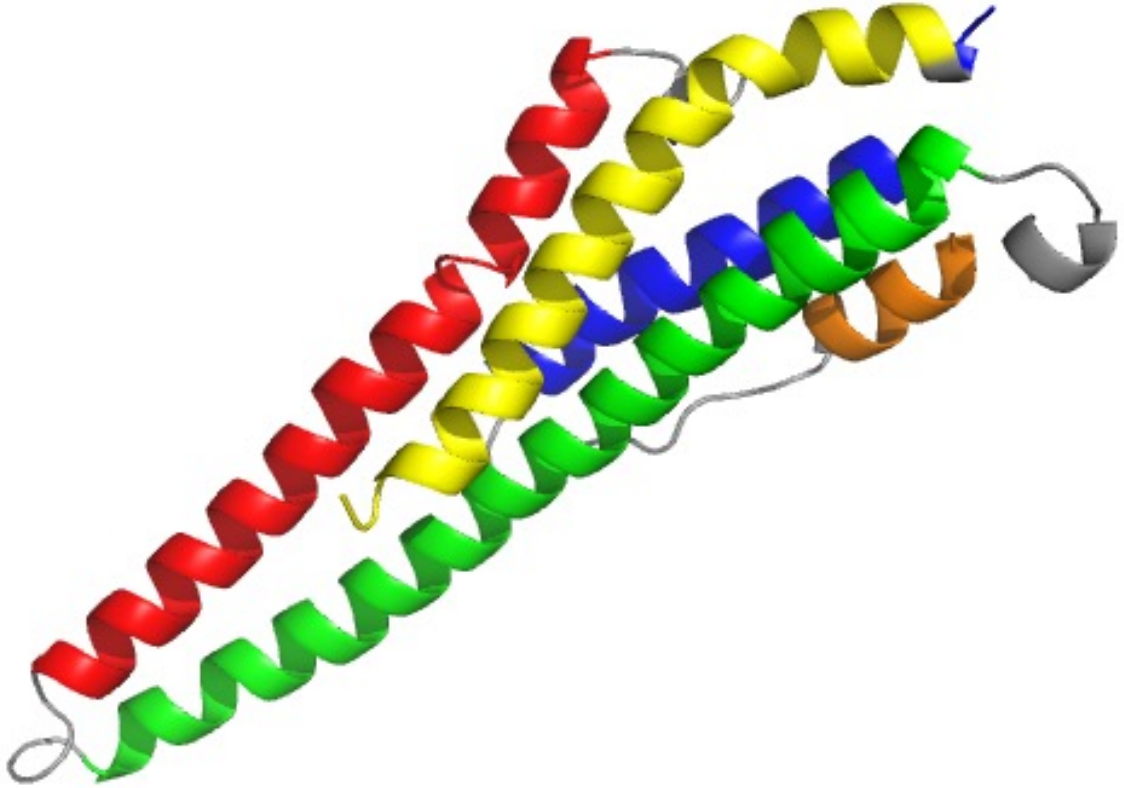
Nejnavštěvovanější Jak začít Přehled zpráv http://www.ncbi.nlm.... http://www.glycoscie... CHMI Radar Departme...

Fold Recognition

View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	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;	**PDB Title:** a common fold for the receptor binding domains of2 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;	**PDB Title:** solution strutcure of the sh3 domain of the cdc42-2 interacting protein 4	n/a

AB2L structure overview

Structure: 4 helical bundle



Top model

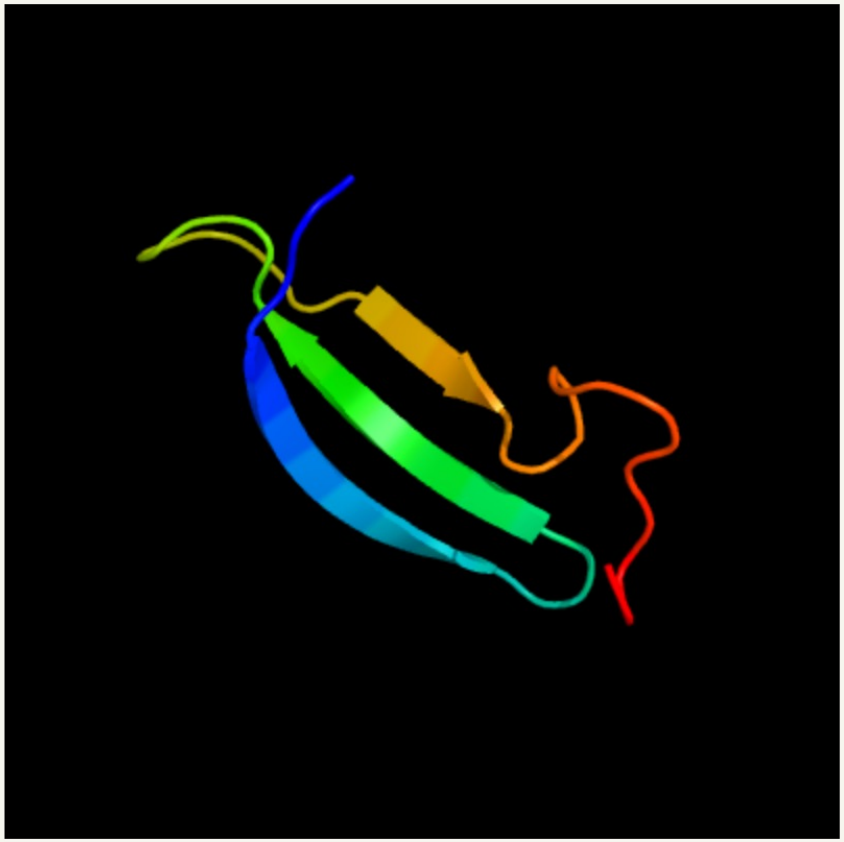


Image coloured by rainbow N → C terminus

Model dimensions (Å): **X**:24.236 **Y**:23.853 **Z**:38.403

Model (left) based on template d2ja9a1

Top template information

Fold:OB-fold
Superfamily: Nucleic acid-binding proteins
Family: Cold shock DNA-binding domain-like

Confidence and coverage

Confidence: **24.1%** Coverage: **20%**

38 residues (20% of your sequence) have been modelled with 24.1% confidence by the single highest scoring template.

Phyre alarm

You may wish to submit your sequence to Phyrealarm. This will automatically scan your sequence every week for new potential templates as they appear in the Phyre2 library.

Please note: You must be registered and logged in to use Phyrealarm.

3D viewing

I-TASSER

- Několikrát vyhodnocen jako nejlepší predikční server

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

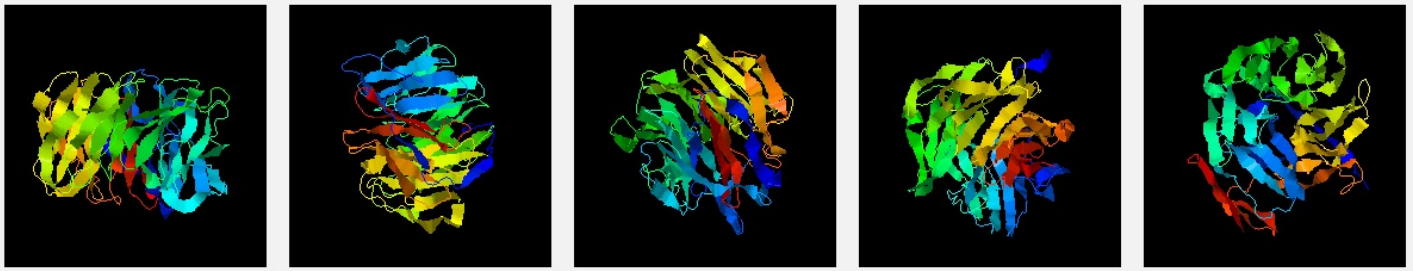
Predicted Secondary Structure

```
                    20                  40                  60                  80                 100                 120                 140
                    |                   |                   |                   |                   |                   |                   |
Sequence    YPFFDNPNYTNTYATNEDFVCPYFLDYYNNSQDDYKNFRGENYDFEDTEENIENRNIEETEYEGLFRAWNPWNNLGGNITSGLGASSWAANRIDLFARGRGGELIHNWFDNGKWNYWENLGGILTSSPKAVSWGFNRIDVVCRGI
Prediction  CCCCCCCCCCCCCCCCCCSSCCCHHHHHHCCCCCCCCCCCCSSCCSSSSSSSSSCCCSSSSSCCCCCCCCCCSSCCCCCCCCSSSSCCCCCSSSSSSCCCCCSSSSSSCCCCCCCCSCCCCCCCCCCSSSSSCCCCSSSSSSCC
Conf.Score  995677866788886653677523565512600045898303601114799982572588751698988877118997789874798089968999995999589998259987666406897658867899389948999998 9
```

Predicted Solvent Accessibility

```
                    20                  40                  60                  80                 100                 120                 140
                    |                   |                   |                   |                   |                   |                   |
Sequence    YPFFDNPNYTNTYATNEDFVCPYFLDYYNNSQDDYKNFRGENYDFEDTEENIENRNIEETEYEGLFRAWNPWNNLGGNITSGLGASSWAANRIDLFARGRGGELIHNWFDNGKWNYWENLGGILTSSPKAVSWGFNRIDVVCRGI
Prediction  401030012233233433010111054122347417314333131433231032331223213233231231230313221200000237210000000341101012233230321331223222200000137220000010 3
            Values range from 0 (buried residue) to 9 (highly exposed residue)
```

Top 5 Models predicted by I-TASSER



Download Model 1	Download Model 2	Download Model 3	Download Model 4	Download Model 5
C-score=-2.09	C-score=-2.42	C-score=-3.44	C-score=-3.46	C-score=-3.53

Estimated accuracy of Model1: 0.47±0.15 (TM-score) 11.3±4.5Å (RMSD) (Read more about C-score of generated models)

Prozkoumání možností a principů fungování I-TASSERu bude domácím úkolem

Homologní modelování

- Je založeno na existenci blízkého **strukturního homologu** (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

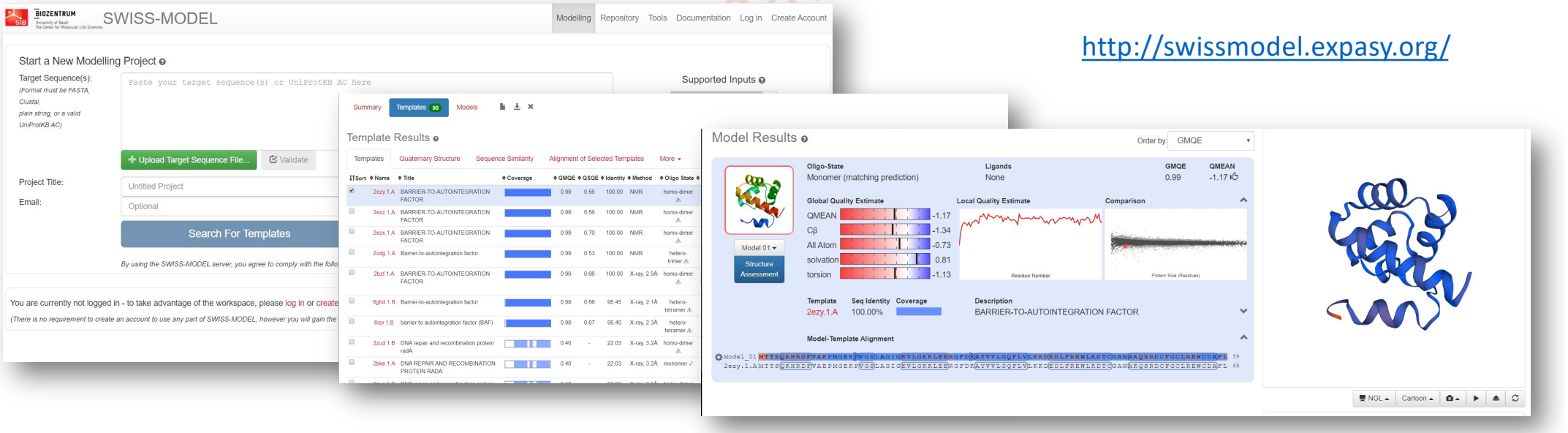
An automated knowledge-based protein modelling server

Homologní modelování

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

Swiss-Model

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



http://swissmodel.expasy.org/

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)
TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVLKKDEDLFREW
LKDTCGANAKQSRDCFGCLREWCDAFL
```

Predicted Secondary Structure

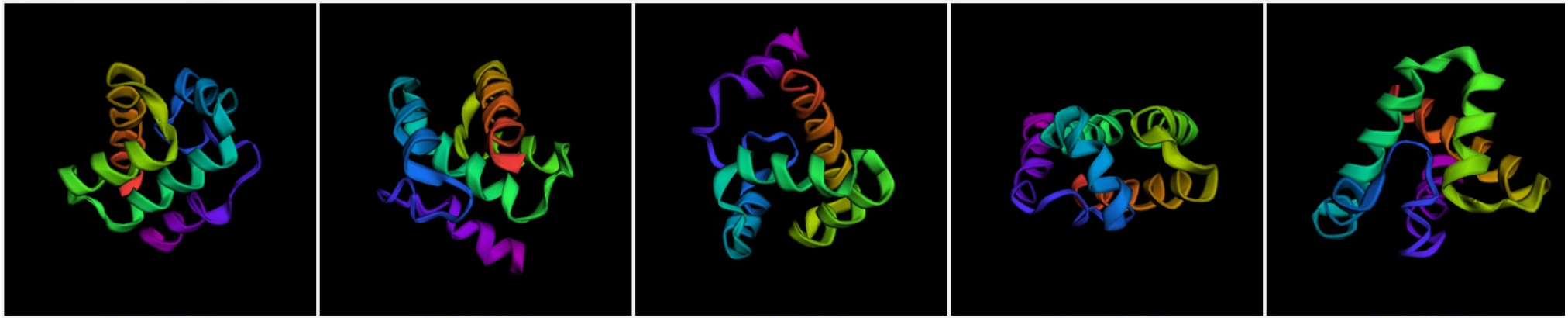
```
                          20                  40                  60                  80
                          |                   |                   |                   |
Sequence    TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVLKKDEDLFREWLKDTCGANAKQSRDCFGCLREWCDAFL
Prediction CCCHHHHHHHCCCCCCCCCCCCCCCHHHHHHHHHCCCHHHHHHHHHHHHCCCHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHC
Conf.Score 988899999879999987447898899999999979659999999999958889999999999968899999999999999985 9
           H:Helix; S:Strand; C:Coil
```

Predicted Solvent Accessibility

```
                          20                  40                  60                  80
                          |                   |                   |                   |
Sequence    TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVLKKDEDLFREWLKDTCGANAKQSRDCFGCLREWCDAFL
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.



Download Model 1 | Download Model 2 | Download Model 3 | Download Model 4 | Download Model 5

De novo modelling with Rossetta

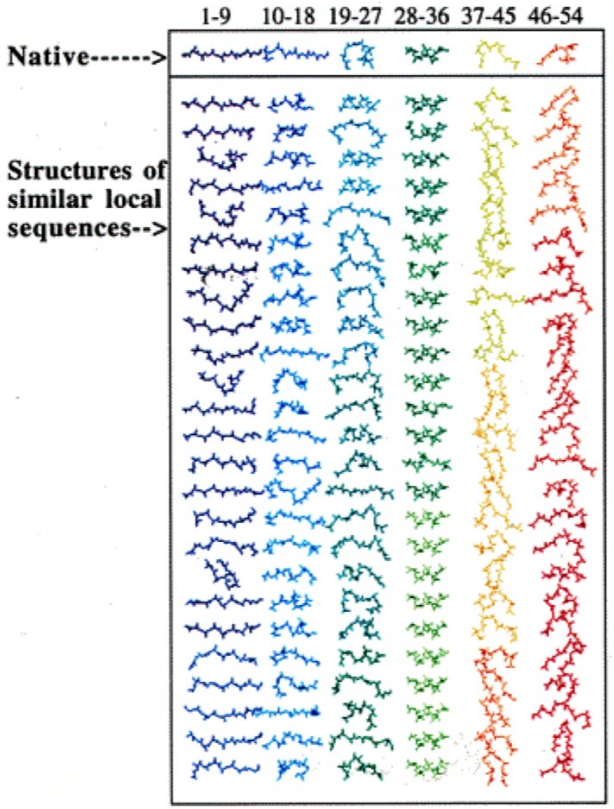
(David Baker lab, Univ. of Washington)

•• In contrast to threading, Rosetta does *de novo* prediction – doesn't use templates/homologous structures

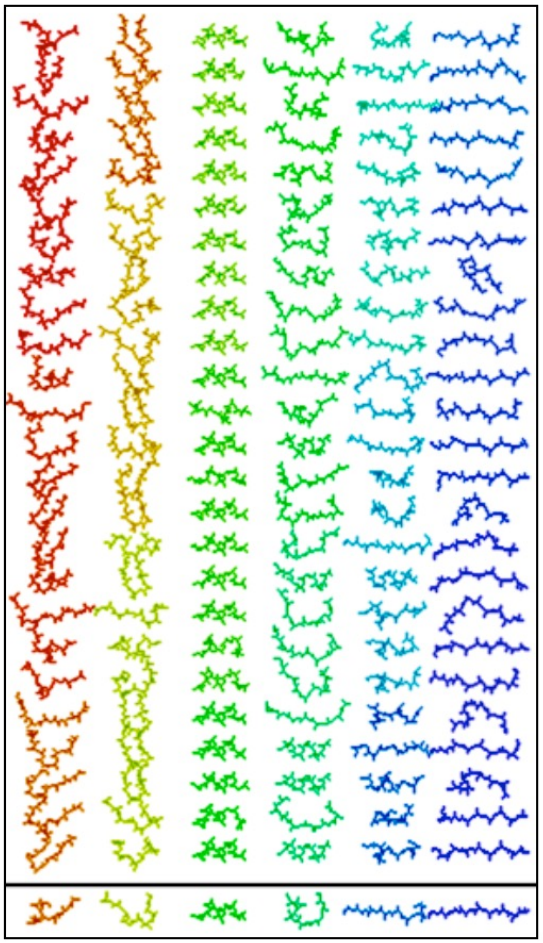
•• instead performs Monte Carlo search through space of conformations to find minimal energy conformation

De novo modelling with Rossetta

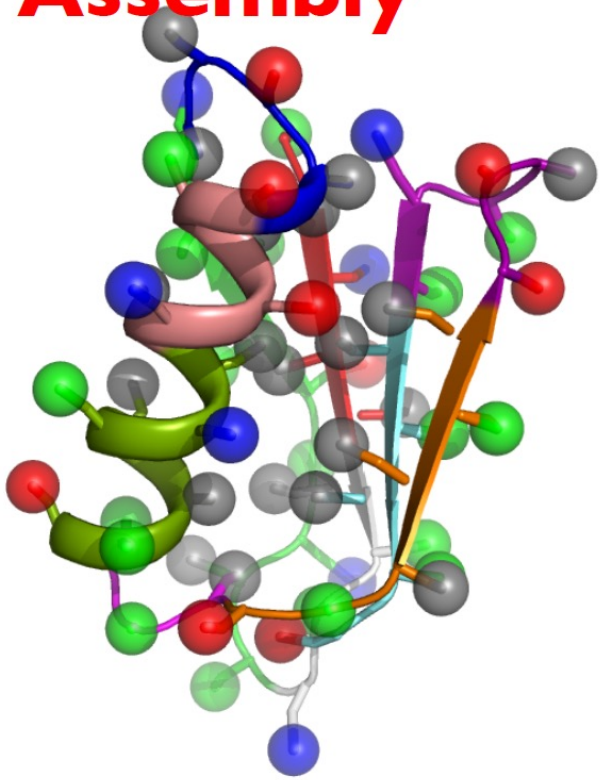
- fragments are selected from known structures
- the window-fragment matches are calculated using
  - PSI-BLAST to build a profile model of the sequence
  - the predicted secondary structure of the sequence



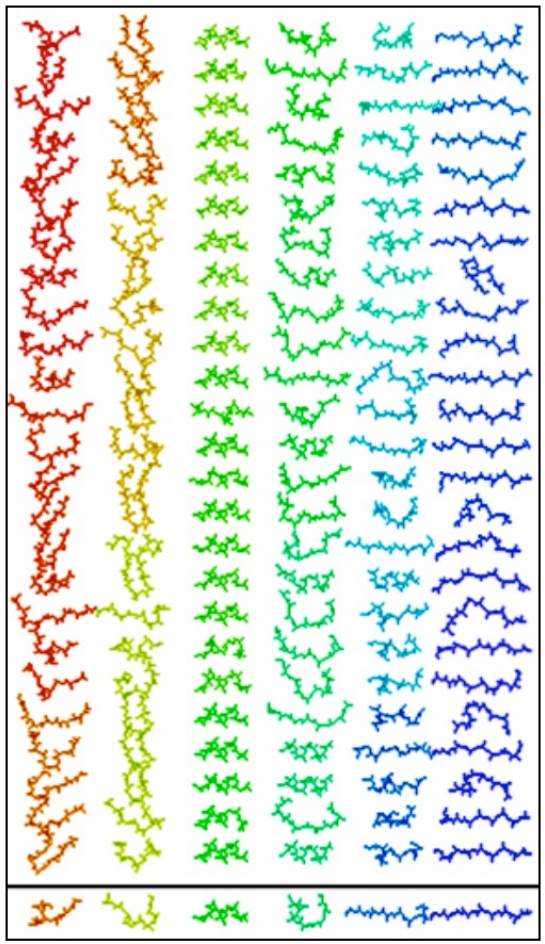
De novo Modeling with Rosetta



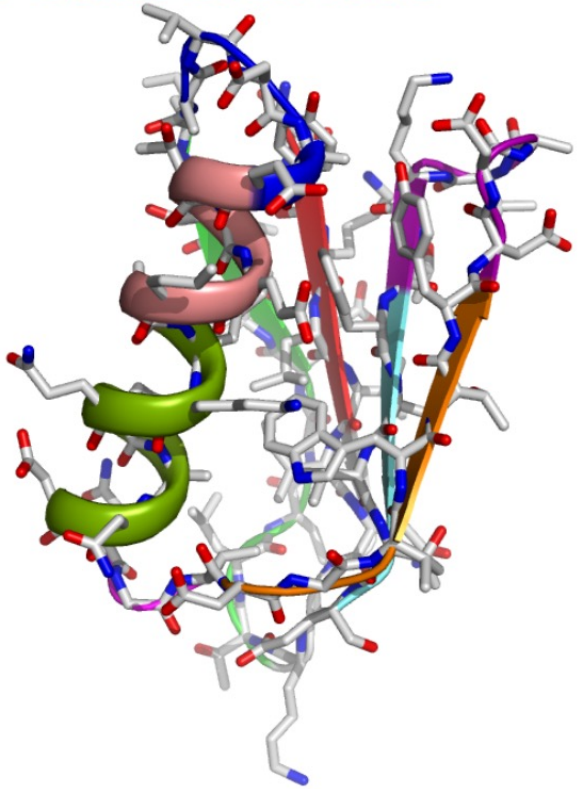
Stage I. Fragment Assembly



De novo Modeling with Rosetta

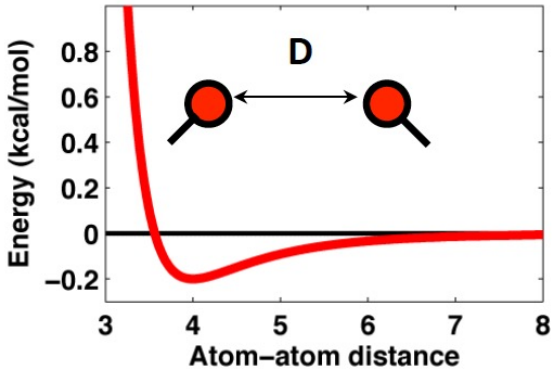


Stage II. All-atom refinement

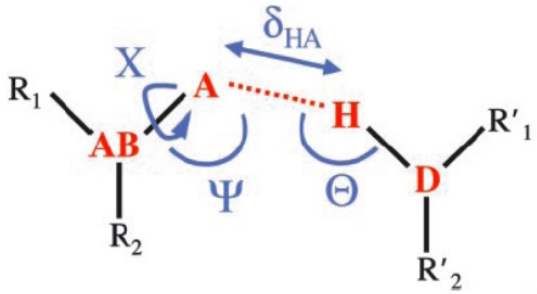


Ingredients of a high resolution potential

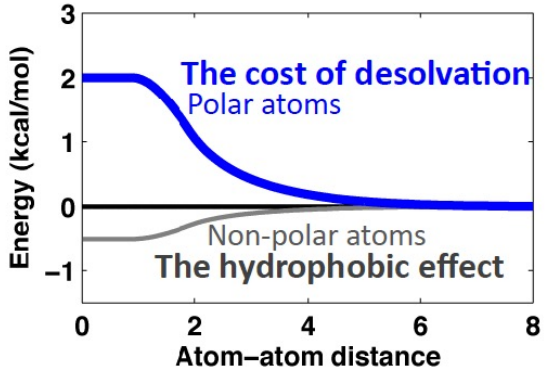
1. Van der waals packing



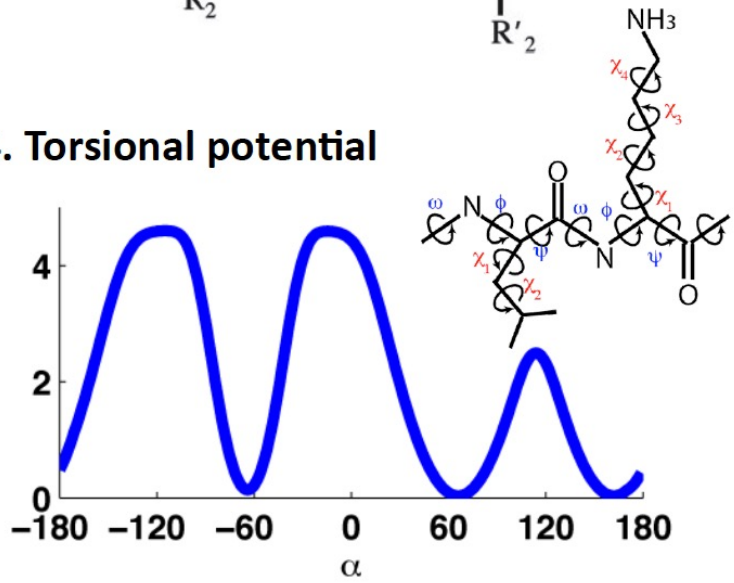
2. Hydrogen bonds



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 on 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 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.

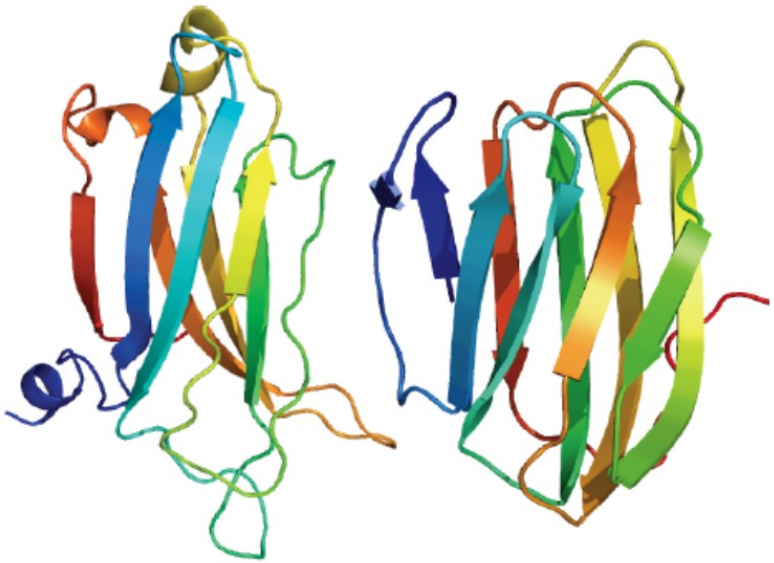
Cut and paste your sequence (in FASTA format, less than 200 AA. Example input

Driving innovation in protein structure prediction: "CASP"

Critical Assessment of Structure Prediction

Five *blind* predictions per target

CASP1 (1994)



CASP1 TARGET (1rsy)

"successful" fold recognition 2tbv

RMSD: 16.0 Å

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:Biology (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: *expert groups*	*112*
prediction servers	*80*
Number of regular targets released	**82**
including *all-group (human) targets*	*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

13th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction

CASP13 in numbers

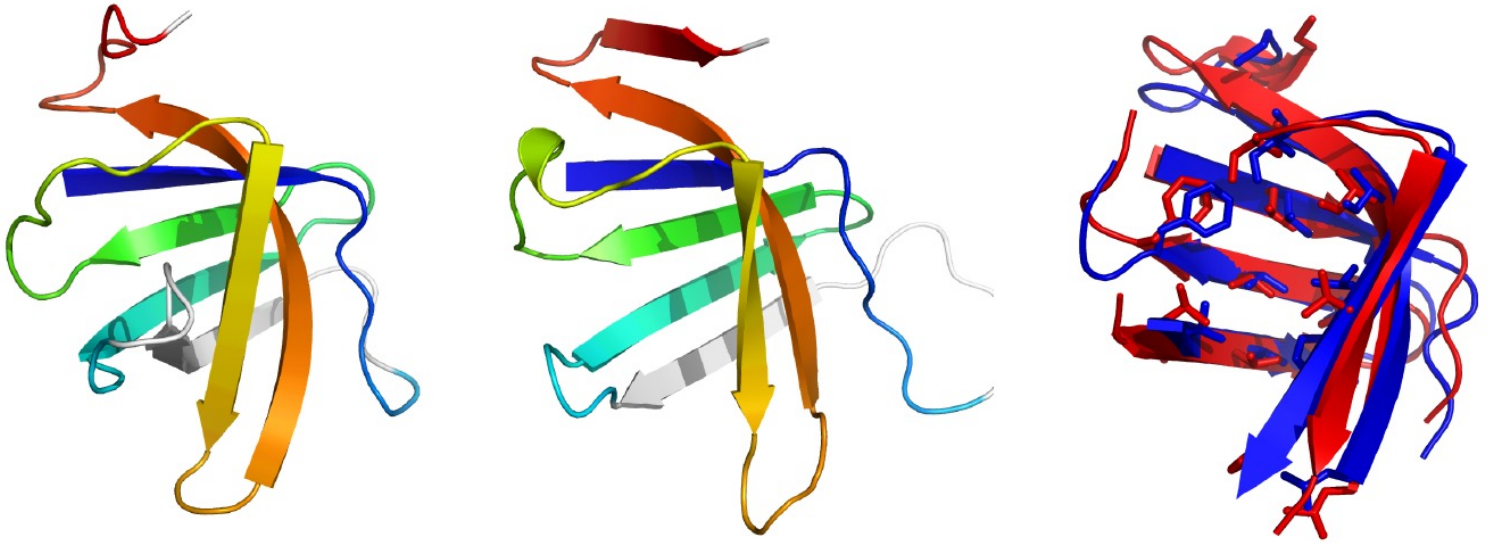
Number of groups registered	**210**
including: *expert groups*	*123*
prediction servers	*87*
Number of tertiary structure prediction targets released	**90**
(including *all-group targets*)	*(82)*
Number of hetero-multimer targets released	**13**
Number of refinement targets released	**31**
Number of 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

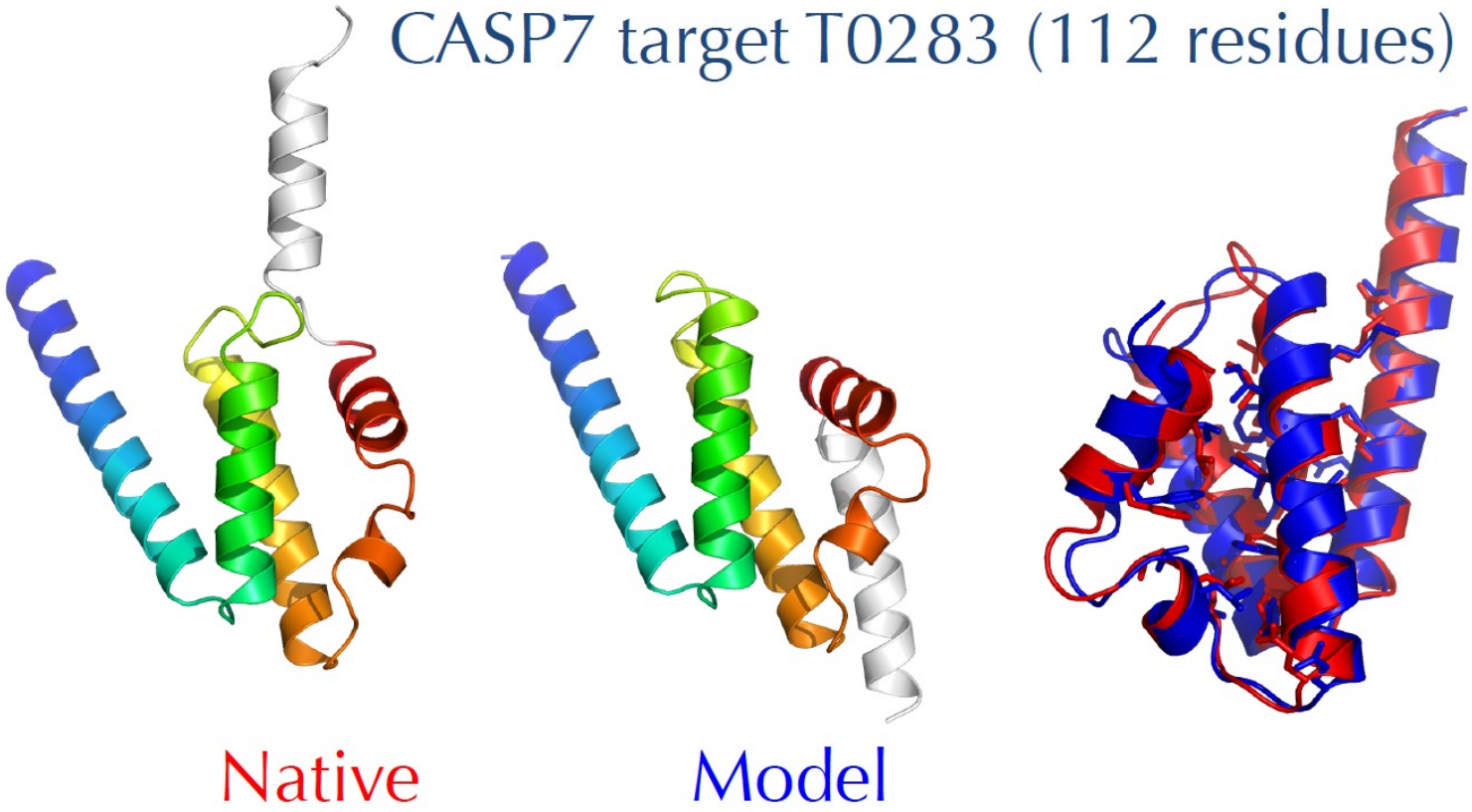
CASP7 target T0316 (domain 3)



Native Model

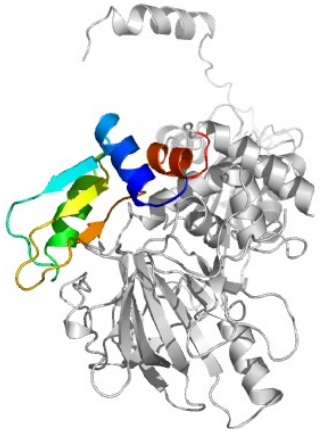
2.0 Å over 61 residues

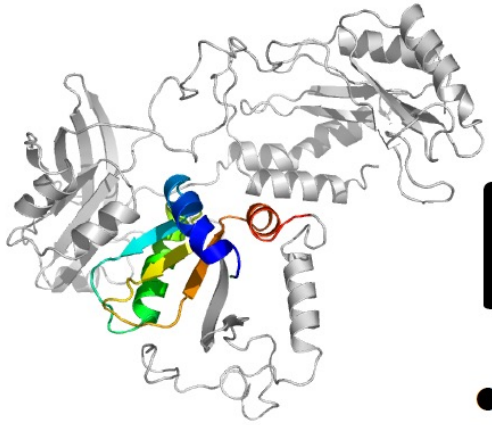
De novo successes: all-α



1.4 Å over 90 residues

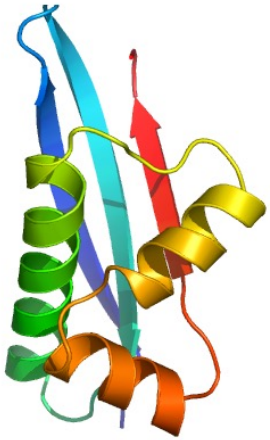
Is protein folding *solved?*





Native

Model

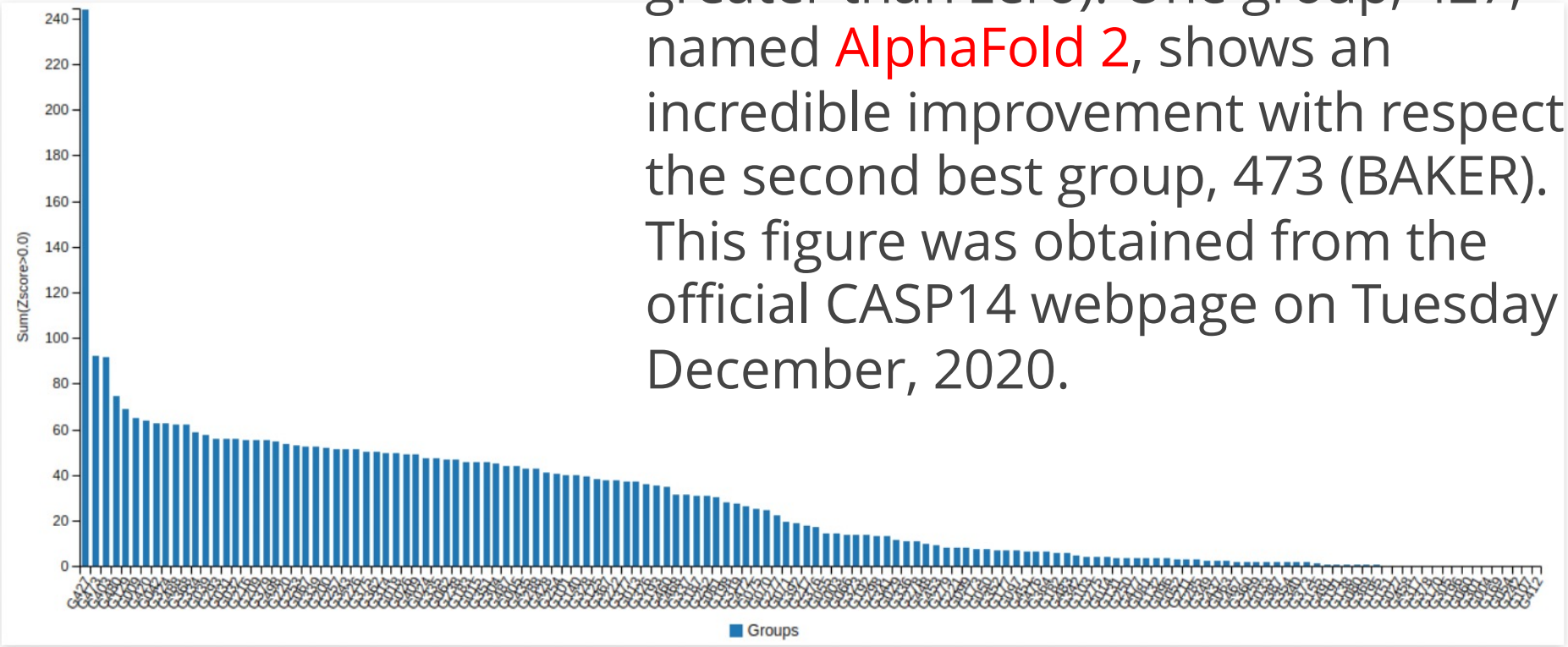


NO! til now?

- 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.



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

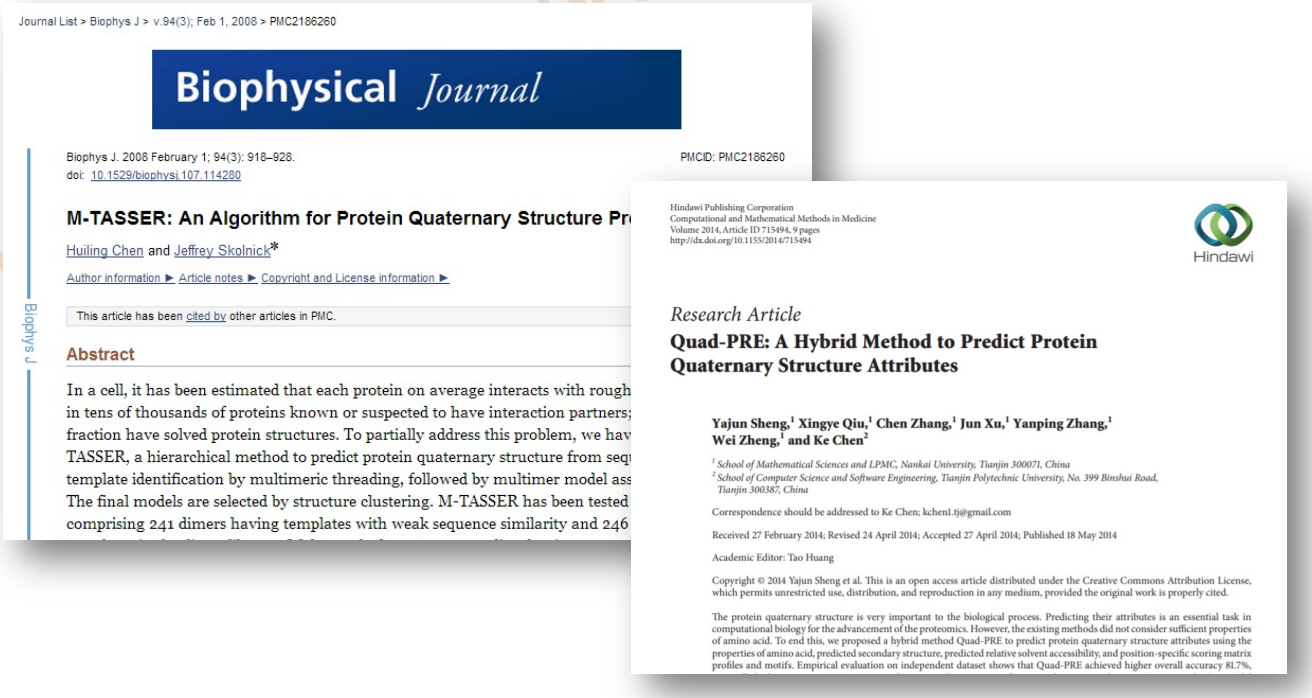
➢ Složitější postupy většinou nejsou automatizované

Predikce kvartérní struktury

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

Příklady SW:

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



Journal List > Biophys J > v.94(3); Feb 1, 2008 > PMC2186260

Biophysical Journal

Biophys J. 2008 February 1; 94(3): 918–928. PMCID: PMC2186260
doi: 10.1529/biophysj.107.114280

M-TASSER: An Algorithm for Protein Quaternary Structure Pr

Huiling Chen and Jeffrey Skolnick*

Author information ▶ Article notes ▶ Copyright and License information ▶

This article has been cited by other articles in PMC.

Biophys J

Abstract

In a cell, it has been estimated that each protein on average interacts with rough
in tens of thousands of proteins known or suspected to have interaction partners;
fraction have solved protein structures. To partially address this problem, we hav
TASSER, a hierarchical method to predict protein quaternary structure from seq
template identification by multimeric threading, followed by multimer model ass
The final models are selected by structure clustering. M-TASSER has been tested
comprising 241 dimers having templates with weak sequence similarity and 246

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 715494, 9 pages
http://dx.doi.org/10.1155/2014/715494

Hindawi

Research Article

Quad-PRE: A Hybrid Method to Predict Protein Quaternary Structure Attributes

Yajun Sheng,[1] Xingye Qiu,[1] Chen Zhang,[1] Jun Xu,[1] Yanping Zhang,[1] Wei Zheng,[1] and Ke Chen[2]

[1] School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China
[2] School of Computer Science and Software Engineering, Tianjin Polytechnic University, No. 399 Binshui Road, Tianjin 300387, China

Correspondence should be addressed to Ke Chen; kchen1.tj@gmail.com

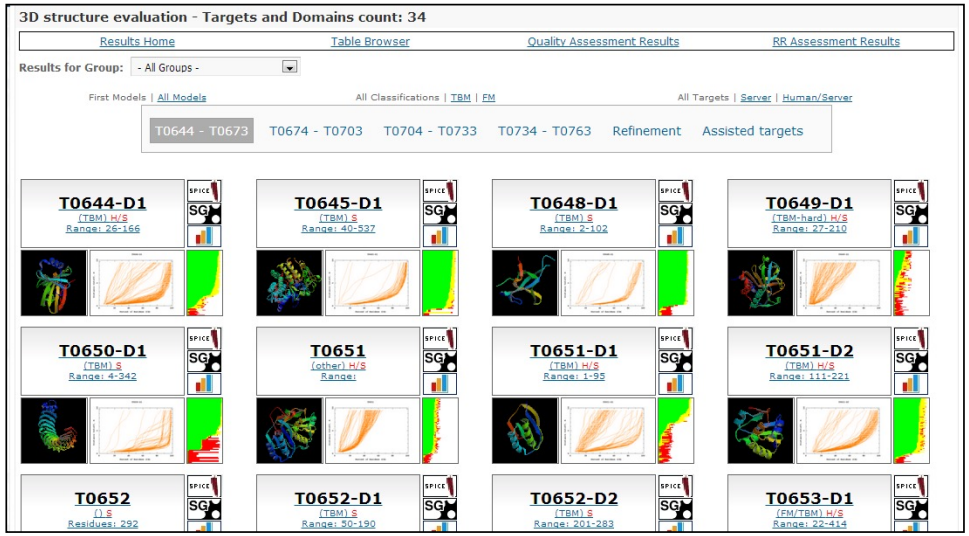
Received 27 February 2014; Revised 24 April 2014; Accepted 27 April 2014; Published 18 May 2014

Academic Editor: Tao Huang

The protein quaternary structure is very important to the biological process. Predicting their attributes is an essential task in computational biology for the advancement of the proteomics. However, the existing methods did not consider sufficient properties of amino acid. To end this, we proposed a hybrid method Quad-PRE to predict protein quaternary structure attributes using the properties of amino acid, predicted secondary structure, predicted relative solvent accessibility, and position-specific scoring matrix profiles and motifs. Empirical evaluation on independent dataset shows that Quad-PRE achieved higher overall accuracy 81.7%,

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



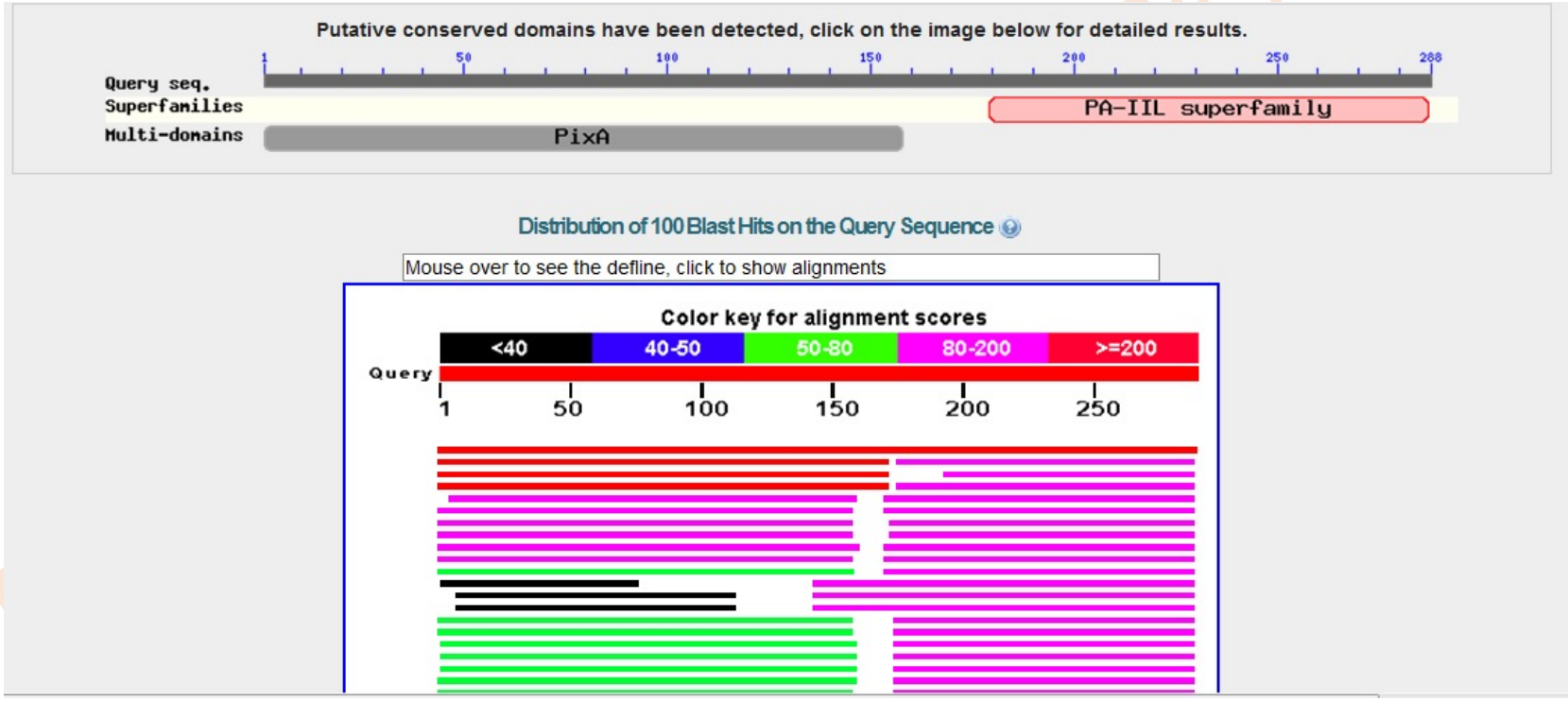
Ale!

! pozor na domény !

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

Prohledávání databází známých aminokyselinových sekvencí

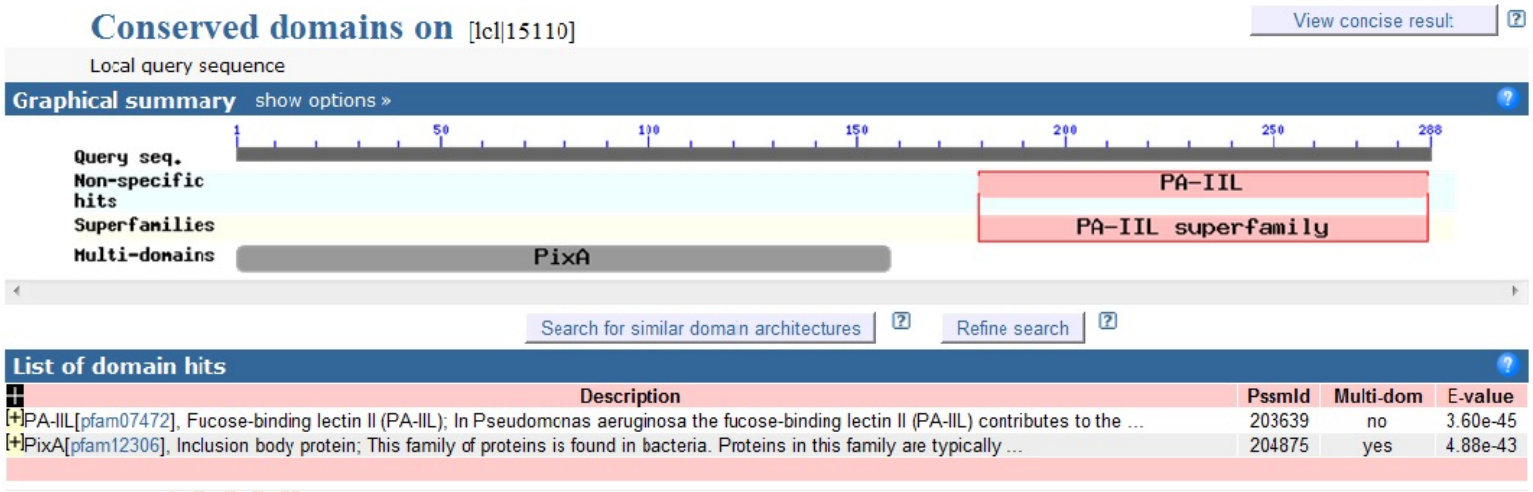
- **celý protein**



NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

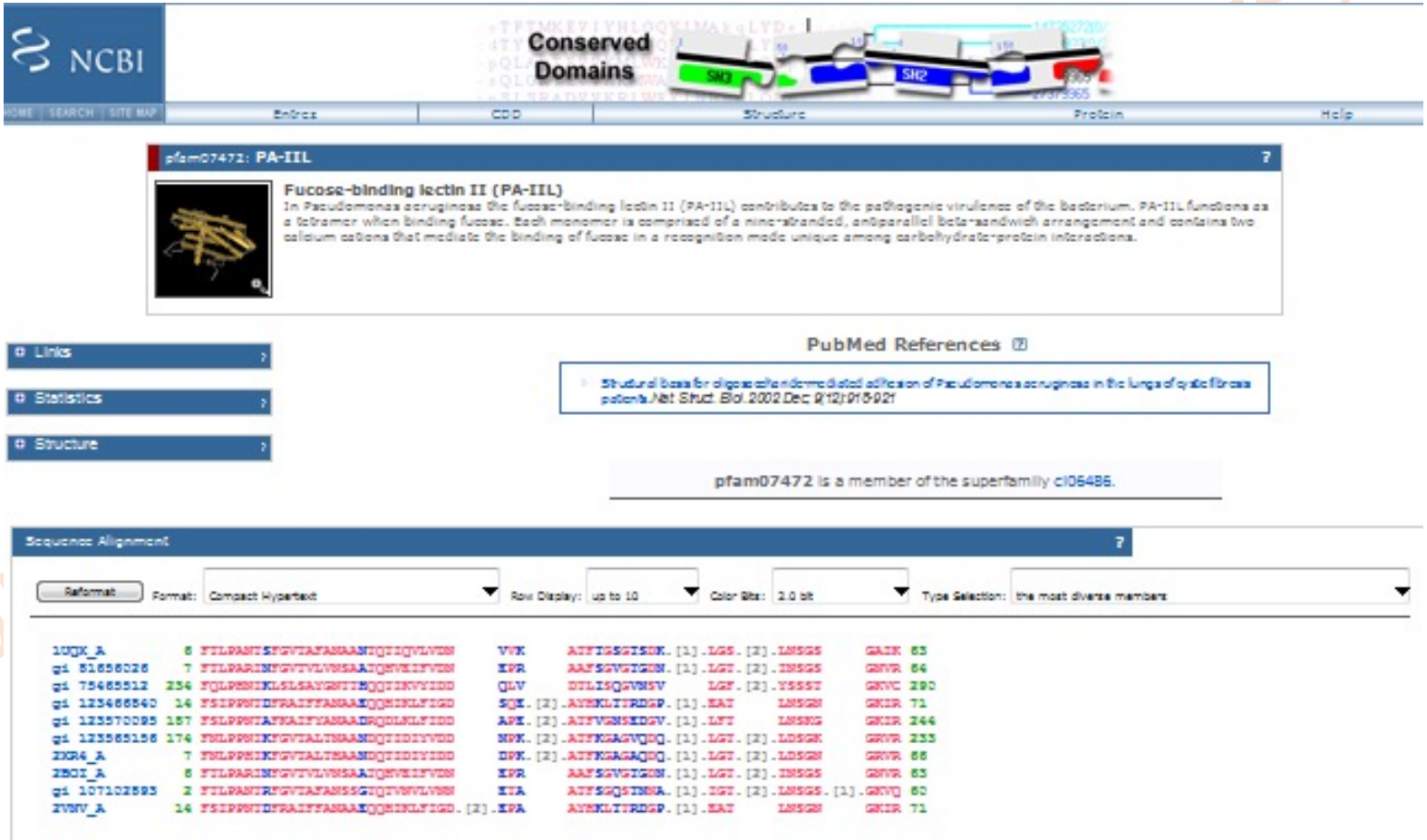
- **celý protein**



NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

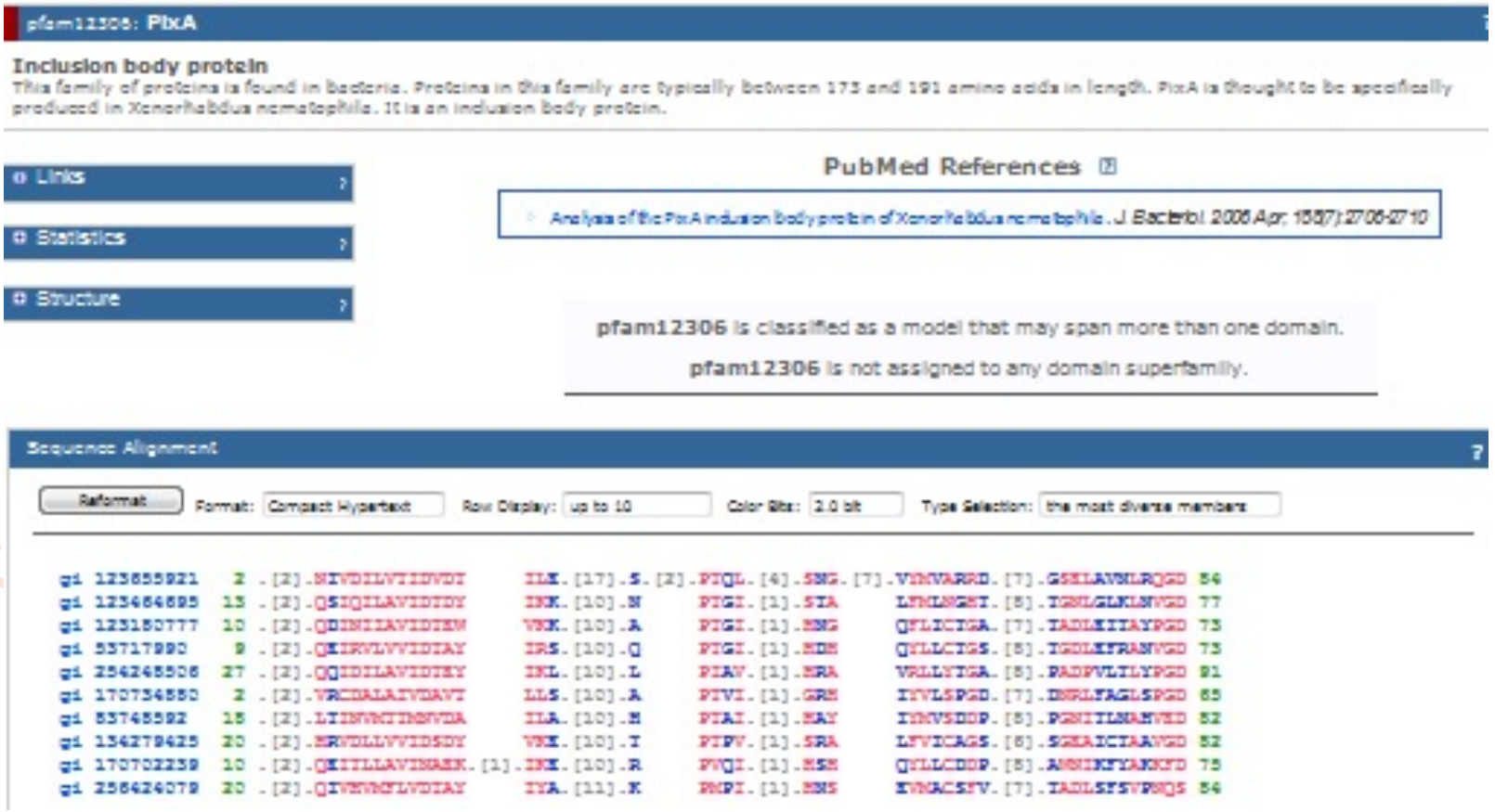
- **celý protein**



NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

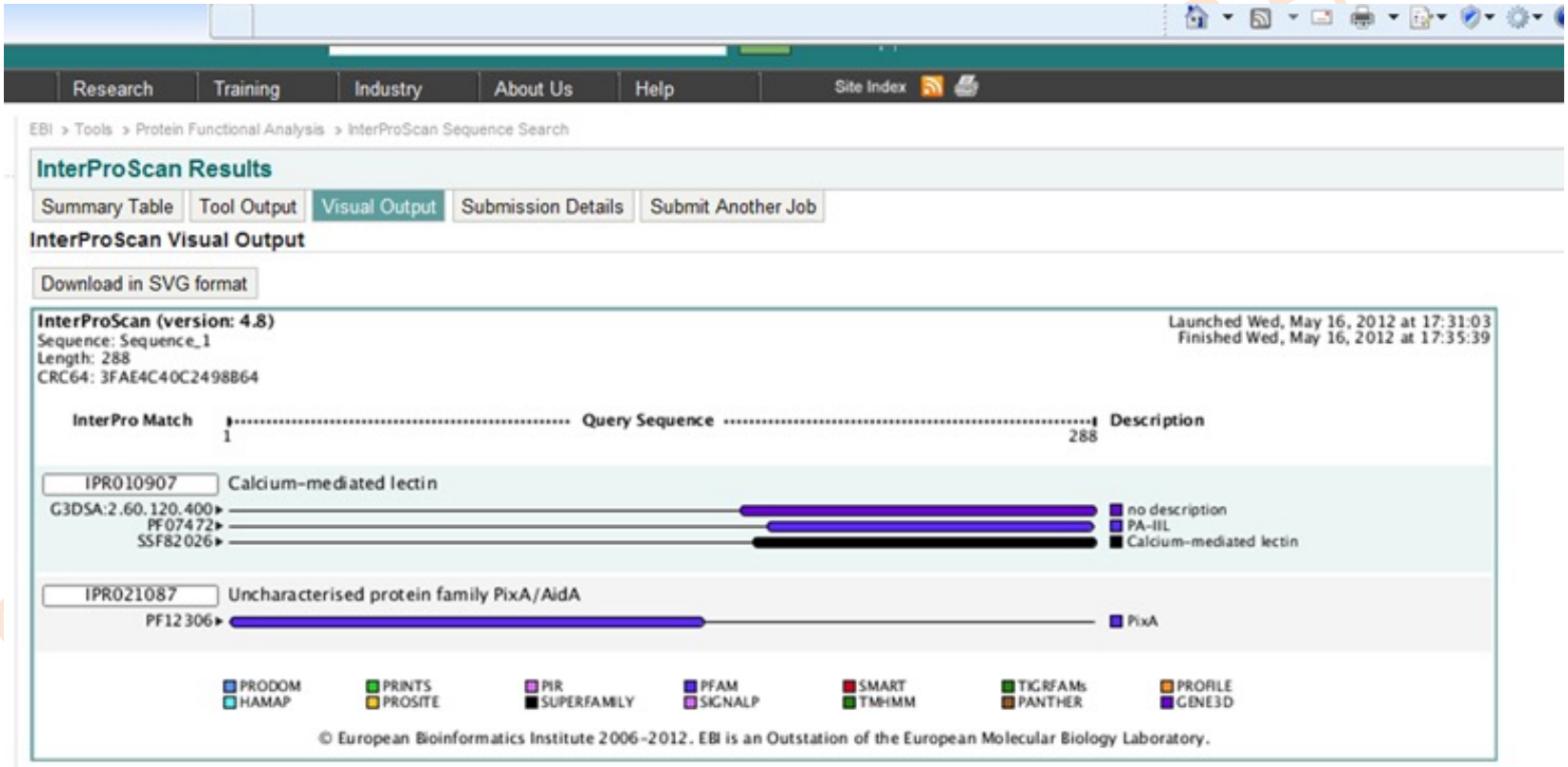
- **celý protein**



InterPro protein sequence analysis & classification

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

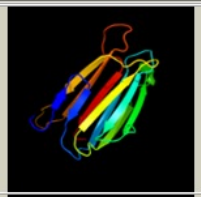
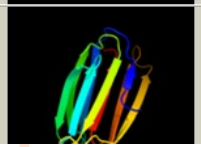
European Bioinformatics Institute - http://www.ebi.ac.uk/



Proč potřebujeme predikci domén

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

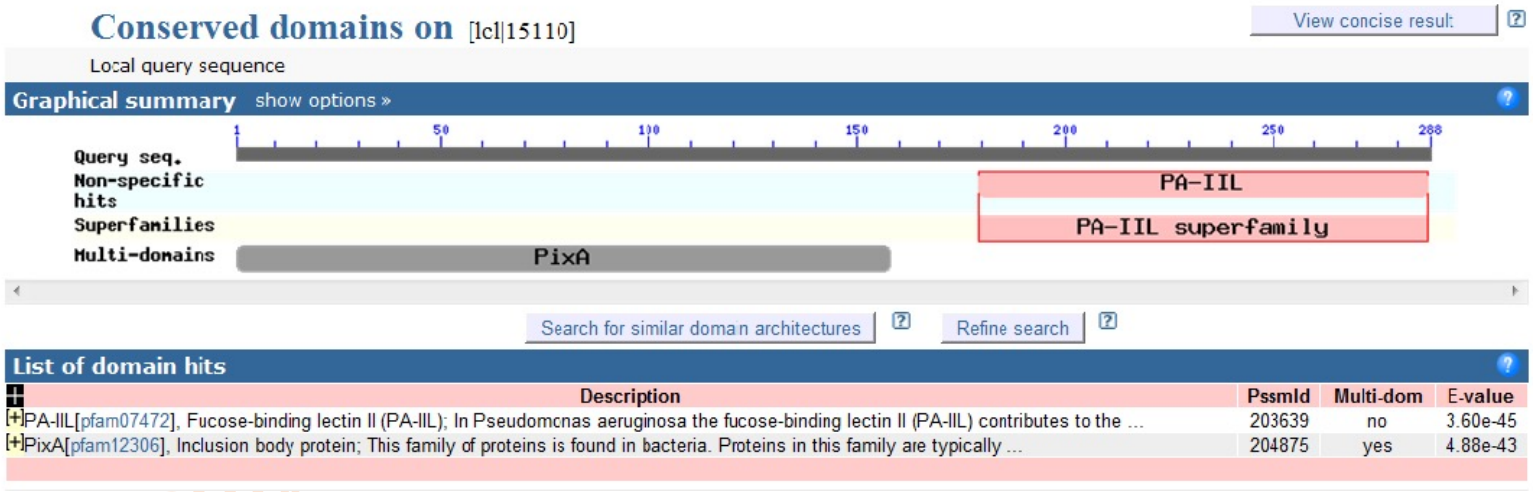
Phyre – whole protein http://www.sbg.bio.ic.ac.uk/phyre2/phyre2_output/a132b051273537c4/summary.html

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c2vnvC_	Alignment		100.0	60	**PDB header:**sugar-binding protein **Chain:** C: **PDB Molecule:**bcla; **PDBTitle:** crystal structure of bcla lectin from burkholderia2 cenocepacia in complex with alpha-methyl-mannoside at 1.73 angstrom resolution
2	c2xr4A_	Alignment		100.0	43	**PDB header:**sugar binding protein **Chain:** A: **PDB Molecule:**lectin; **PDBTitle:** c-terminal domain of bc2l-c lectin from burkholderia cenocepacia
3	d2chha1	Alignment		100.0	37	**Fold:**Calcium-mediated lectin **Superfamily:**Calcium-mediated lectin **Family:**Calcium-mediated lectin

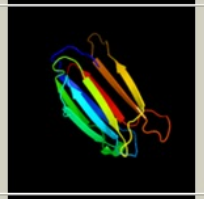
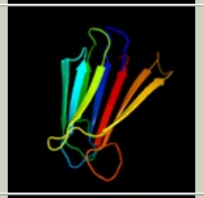
NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

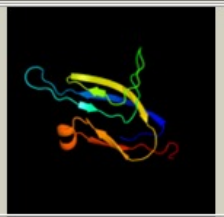
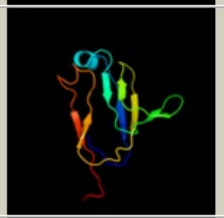
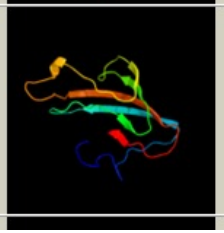
- **celý protein**



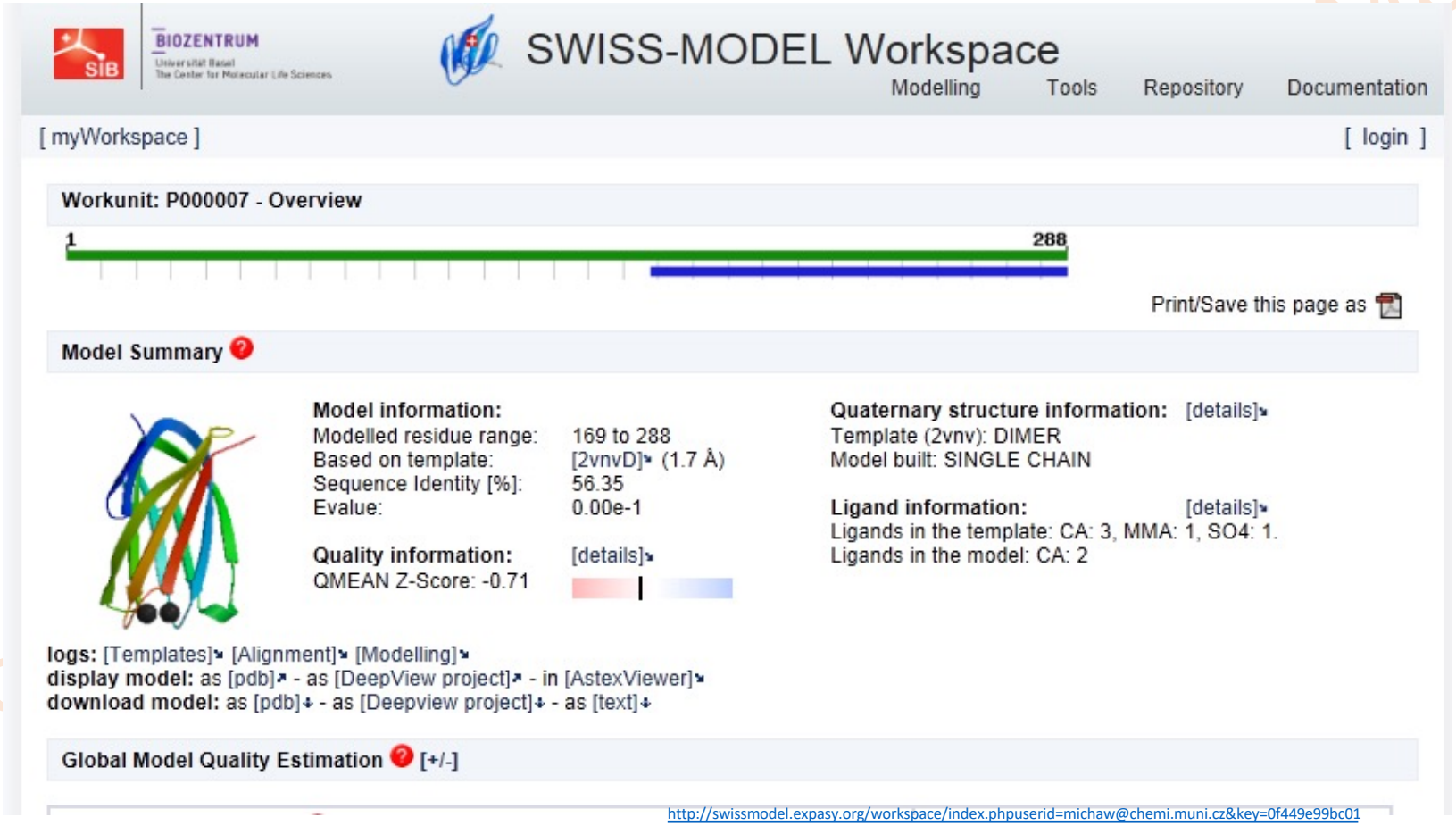
Phyre – C-term http://www.sbg.bio.ic.ac.uk/phyre2/phyre2_output/e332b1ecabb8d0a6/summary.html

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c2xr4A_	Alignment		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	c2vnvC_	Alignment		100.0	62	**PDB header:**sugar-binding protein **Chain:** C: **PDB Molecule:**bcla; **PDBTitle:** crystal structure of bcla lectin from burkholderia2 cenocepacia in complex with alpha-methyl-mannoside at 1.73 angstrom resolution
3	d1uzva_	Alignment		100.0	30	**Fold:**Calcium-mediated lectin **Superfamily:**Calcium-mediated lectin **Family:**Calcium-mediated lectin

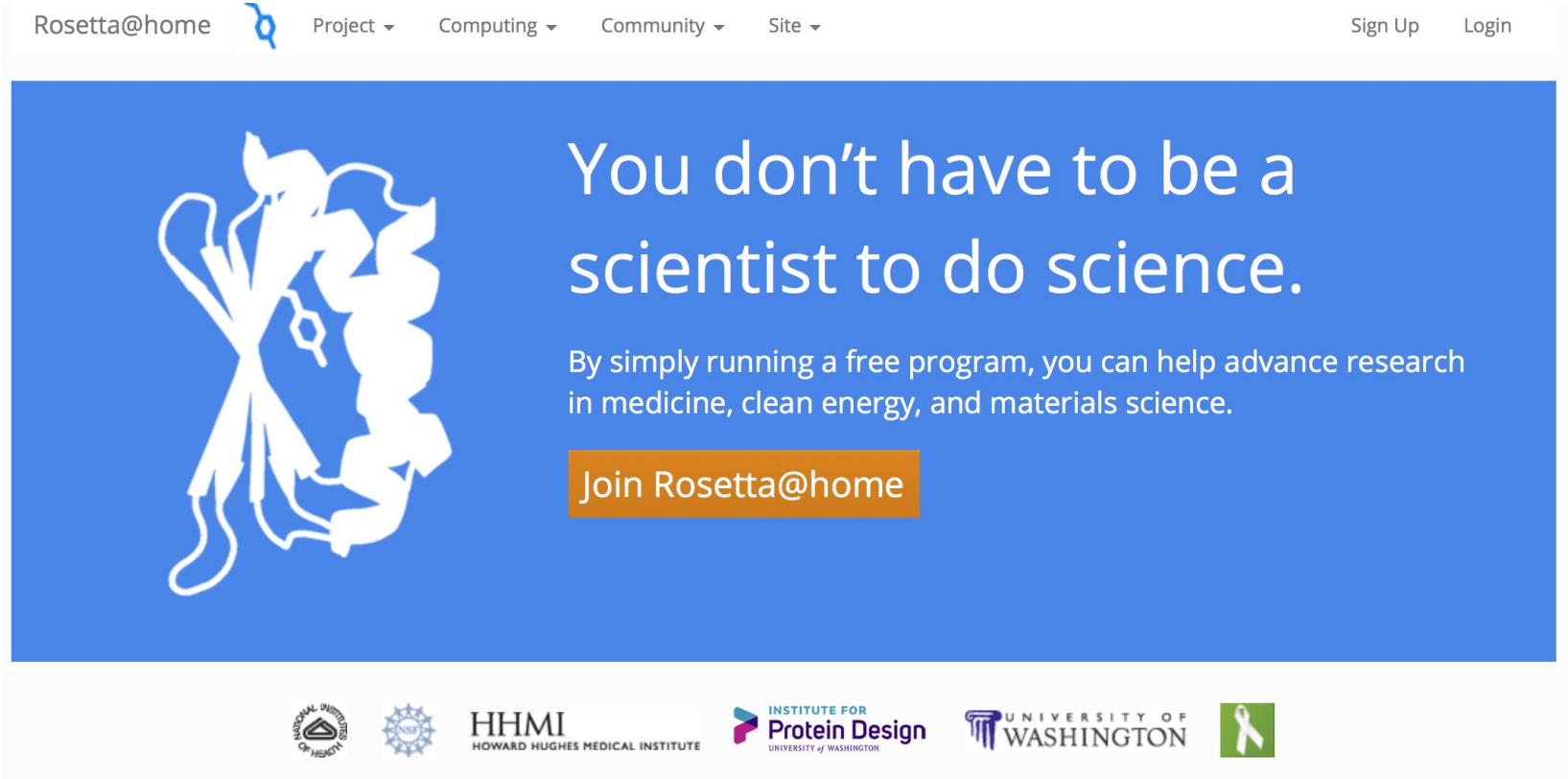
Phyre – n-term http://www.sbg.bio.ic.ac.uk/phyre2/phyre2_output/e332b1ecabb8d0a6/summary.html

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c1sddB_	Alignment		83.7	9	**PDB header:**blood clotting **Chain:** B: **PDB Molecule:**coagulation factor v; **PDBTitle:** crystal structure of bovine factor vai
2	c3cdzB_	Alignment		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	Alignment		68.0	13	**Fold:**Cupredoxin-like **Superfamily:**Cupredoxins **Family:**Multidomain cupredoxins

Swissprot – whole protein



http://swissmodel.expasy.org/workspace/index.phpuserid=michaw@chemi.muni.cz&key=0f449e99bc0176edfa75fba19b2d96e4&func=workspace_modelling&prjid=P000007



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!

The Science Behind Foldit

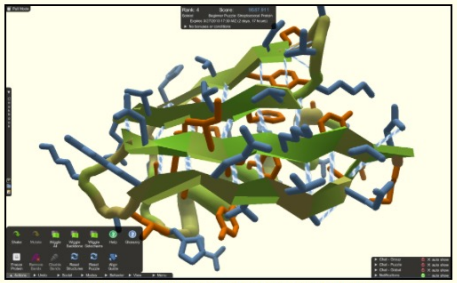
Foldit is a revolutionary crowdsourcing computer game enabling *you* to contribute to important scientific research. This page describes the science behind Foldit and how your playing can help.

Page Contents:

http://fold.it/portal/

What is protein folding?

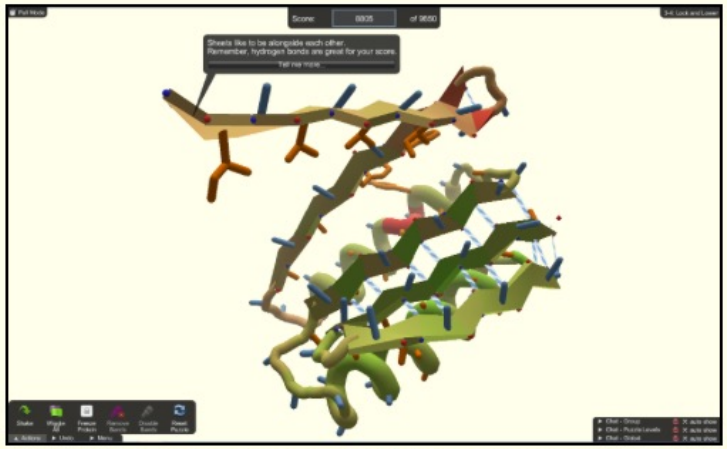
What is a protein? Proteins are the workhorses in every cell of every living thing. Your body is made up of trillions of cells, of all different kinds: muscle cells, brain cells, blood cells, and more. Inside those cells, proteins are allowing your body to do what it does: break down food to power your 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



Folded up Streptococcal Protein Puzzle
(+) Enlarge This Image

https://fold.it/portal/" consists of a long chain of

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

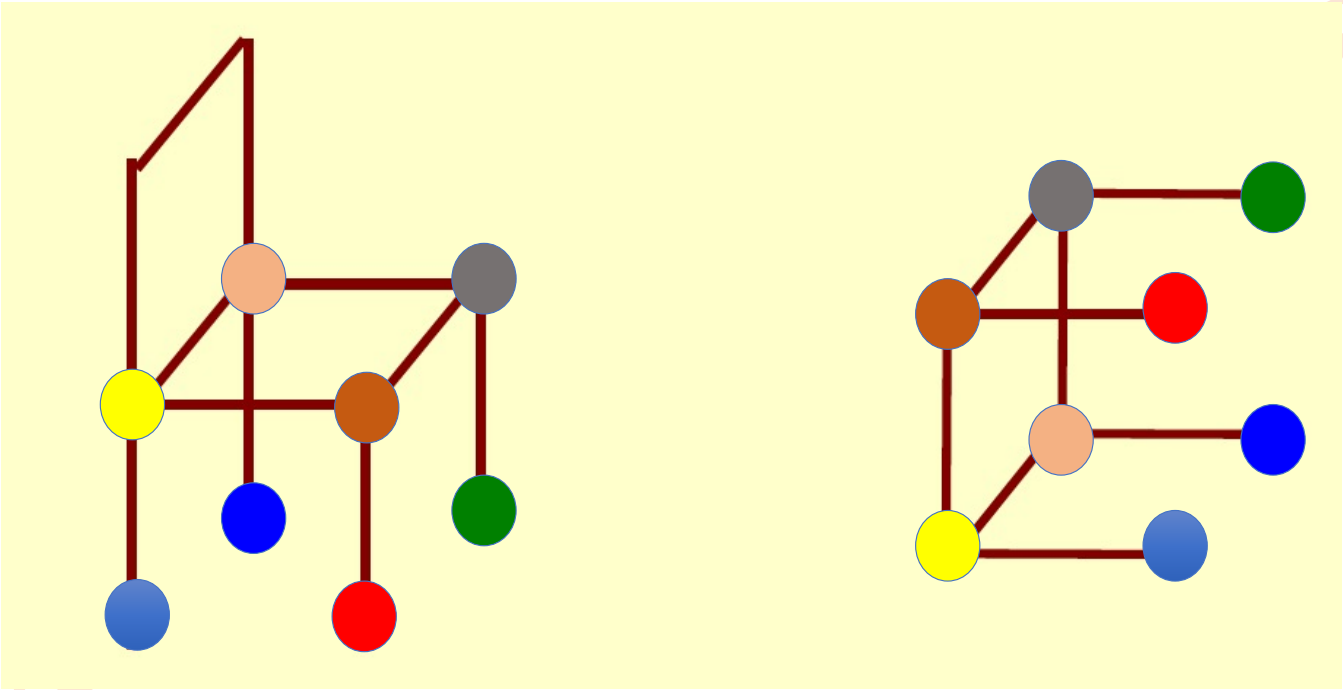
What other good stuff am I contributing to by playing?

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?

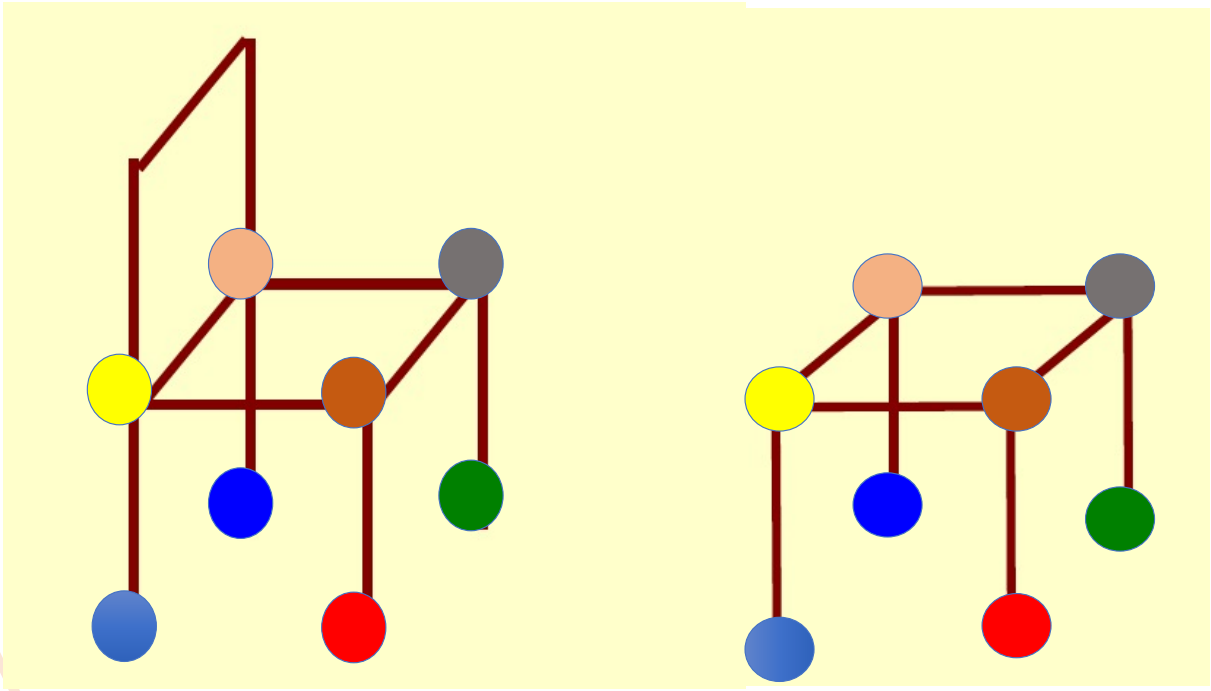
We're collecting data to find out if humans' pattern-recognition and puzzle-solving abilities make them more efficient than existing computer programs at pattern-folding tasks. If this turns out to be true, we can then teach human strategies to computers and fold proteins faster than ever!

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

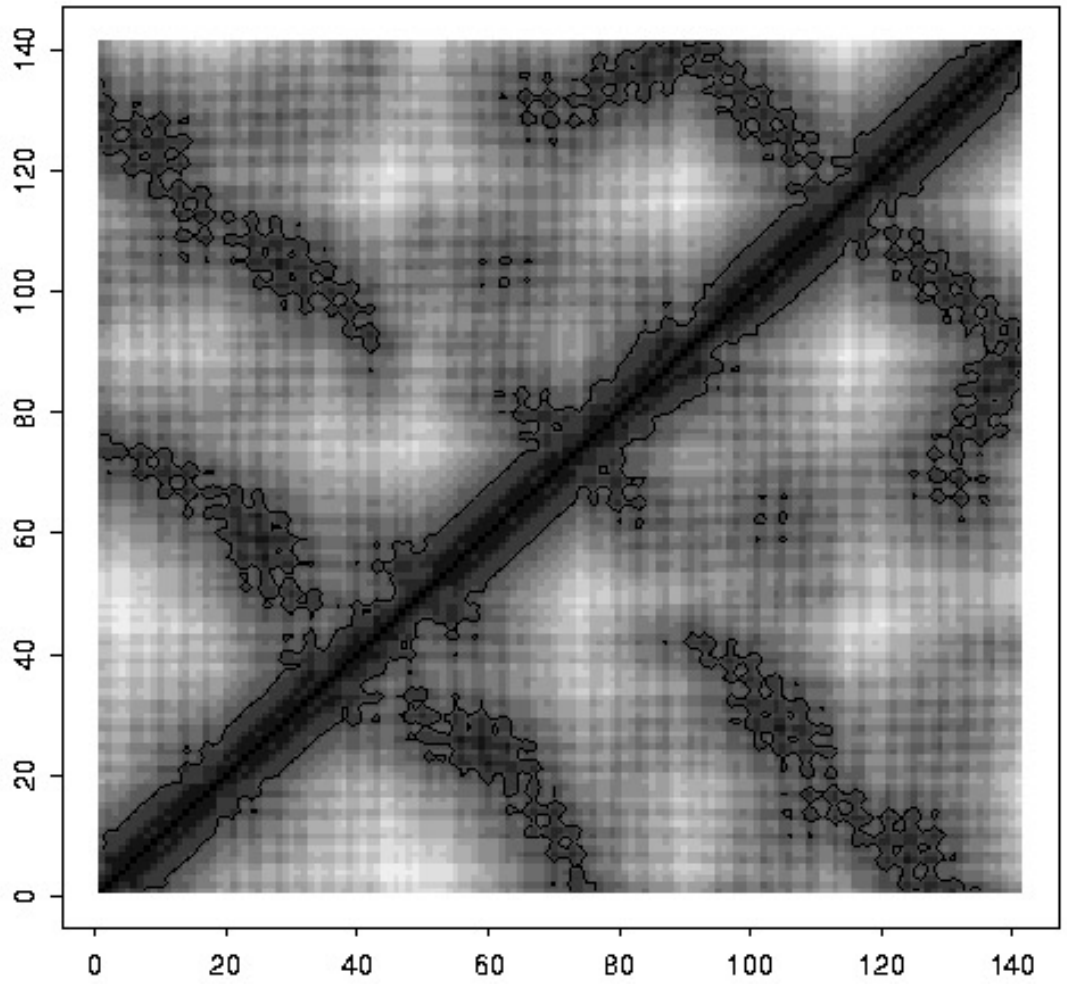
Both

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

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

http://fatcat.godziklab.org/fatcat/

DALI



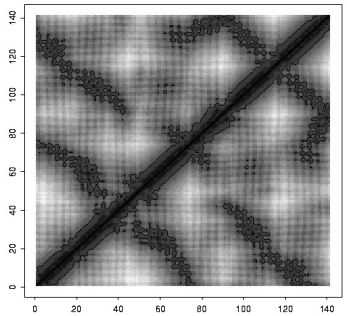
An intra-molecular distance plot for myoglobin

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

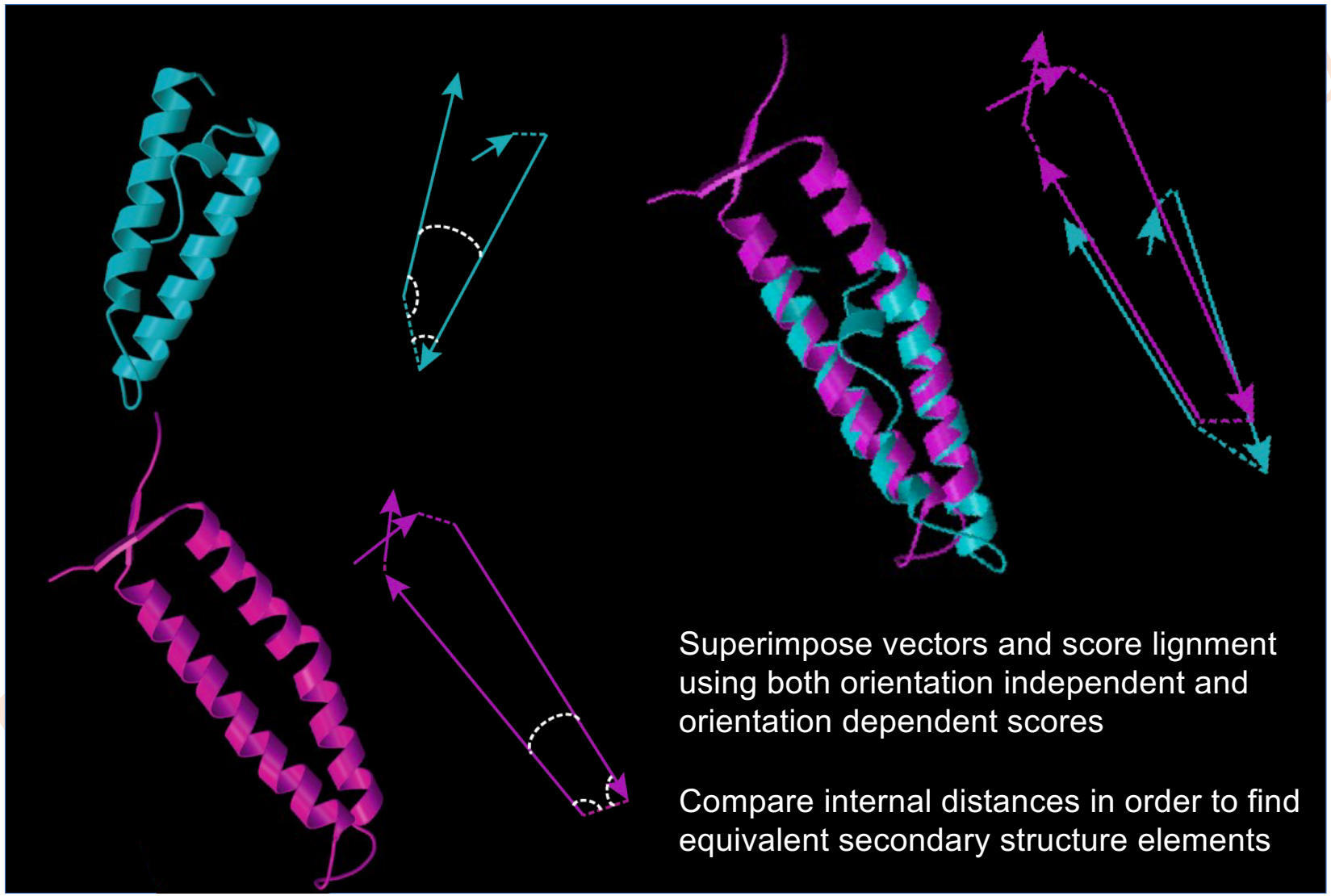


VAST – Vector Alignment Search Tool

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


Superimpose vectors and score lignment using both orientation independent and orientation dependent scores
Compare internal distances in order to find equivalent secondary structure elements

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