Předpověď 3D-struktury/foldu/funkce

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

- Threading - „navlékání“
- Homology modeling
- *Ab inicio* metody

Vše začíná u PDB ...

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | MyPDB

RCSB PDB PROTEIN DATA BANK

151079 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands | Go

Advanced Search | Browse by Annotations

PDB-101 | wwPDB | EMDataResource | ndb Nucleic Acid Database | Worldwide Protein Data Bank Foundation | Take the User Survey

Welcome
Deposit
Search
Visualize
Analyze
Download
Learn

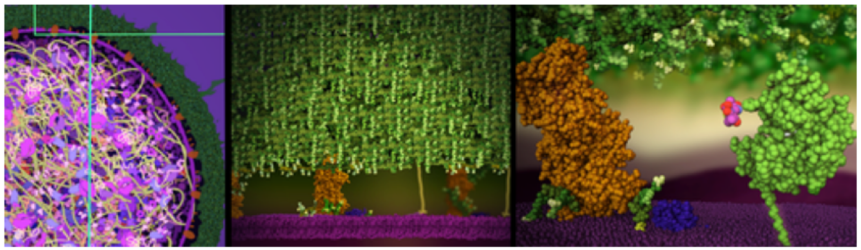
A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

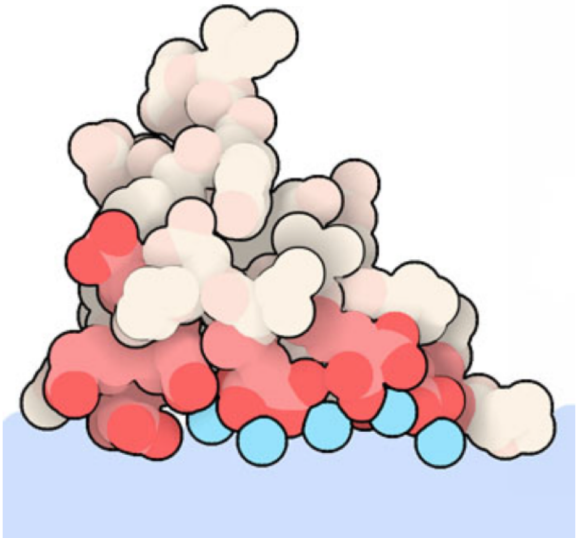
As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

New Video: Penicillin and Antibiotic Resistance



April **Molecule of the Month**



Proteins and Biominerals

Vše začíná u PDB ...

EMBL-EBI

Services | Research | Training | About us

Protein Data Bank in Europe

Bringing Structure to Biology

Search

Examples: hemoglobin, BRCA1_HUMAN

Advanced search

PDBe home | Deposition | PDBe services | PDBe training | Documentation | About PDBe | Share | Feedback

PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures. Read more about PDBe.

Featured structure

Getting on the front foot

1st April 2019

Our featured structure for April in our 2019 calendar is based around a beautiful image of a virus that causes devastation to the world's livestock population.

Read more...

Previous featured structures

News

New PDBe-KB aggregated views of protein structure

21 March, 2019

The PDB archive reaches a significant

Events

CCP-EM Spring Symposium

Nottingham University, UK

29 Apr 2019 to 1 May 2019

Bioinformatics resources for protein

Popular

- EMsearch
- PDBeFold
- PDBePISA
- PDBeChem
- Sequence search
- PDBe REST API
- EM resources
- NMR resources
- EMPIAR
- Coordinate Server
- PDB Component Library
- News
- Events
- Training
- Contact us

Latest archive statistics

As of 17 April 2019 the PDB contains 151081 entries (latest PDB entries, chemistry, biology) and EMDB contains 8002 entries (latest map releases, latest header releases, latest updates).

Tweets by @PDBeurope

Protein Data Bank Retweeted

Databases of Protein Folds

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

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

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

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

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

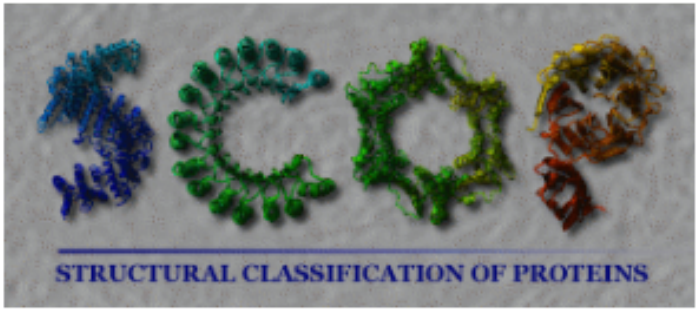
- Class - Architecture - Topology - Homologous Superfamily
- Manual classification at Architecture level
- Automated topology classification using SSAP (Orengo & Taylor)

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

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

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

SCOP **S**tructural **C**lassification **o**f **P**roteins
(http://scop.mrc-lmb.cam.ac.uk/scop)



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

38221 PDB Entries. 1 Literature Reference. 110800 Domains. (excluding nucleic acids and theoretical models).
Folds, superfamilies, and families statistics here.
New folds superfamilies families.
List of obsolete entries and their replacements.

Authors. Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. scop@mrc-lmb.cam.ac.uk
Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]
Recent changes are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [PDF],
Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [PDF], and
Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucl. Acid Res.* advance access, doi:10.1093/nar/gkm993. [PDF].

Access methods

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

The **SCOP** database, created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. http://scop.mrc-lmb.cam.ac.uk/scop

Family: *Clear evolutionarily relationship*

Proteins clustered together into families are clearly evolutionarily related. Generally, this means that pairwise residue identities between the proteins are 30% and greater. *However, in some cases similar functions and structures provide definitive evidence of common descent in the absense of high sequence identity; for example, many globins form a family though some members have sequence identities of only 15%.*

Superfamily: *Probable common evolutionary origin*

Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies. *For example, actin, the ATPase domain of the heat shock protein, and hexakinase together form a superfamily*.

Fold: *Major structural similarity*

Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. *Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.*

SCOP: Root: scop - Mozilla Firefox

Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda

http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html

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

Root: scop

Classes:

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

---

Enter search key: Search

News

November,2013
During the development of SCOP2, we have identified a new, previously unrecognised type of alpha-alpha superhelix. Unlike other alpha-alpha superhelices..
More...

January,2014
SCOP2 article in NAR is published
More...

January, 2014
The structure of the month
More...

Welcome to SCOP2!

Citation

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

Description of the SCOP2 database

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

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

Quick introduction on how to browse, search and download

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

Search Browser

Search

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

Search Graph

Search

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

http://scop2.mrc-lmb.cam.ac.uk/front.html

SCOPe | Browse | Stats & History | Downloads | Help | Search SCOPe

Welcome to SCOPe!

SCOPe (Structural Classification of Proteins — extended) is a database developed at the Berkeley Lab and UC Berkeley to extend the development and maintenance of SCOP. SCOP was conceived at the MRC Laboratory of Molecular Biology, and developed in collaboration with researchers in Berkeley. Work on SCOP (version 1) concluded in June 2009 with the release of SCOP 1.75.

SCOPe classifies many newer structures through a combination of automation and manual curation, and corrects some errors in SCOP, aiming to have the same accuracy as the hand-curated SCOP releases. **SCOPe** also incorporates and updates the ASTRAL database.

About SCOPe | Stats & Prior Releases

News

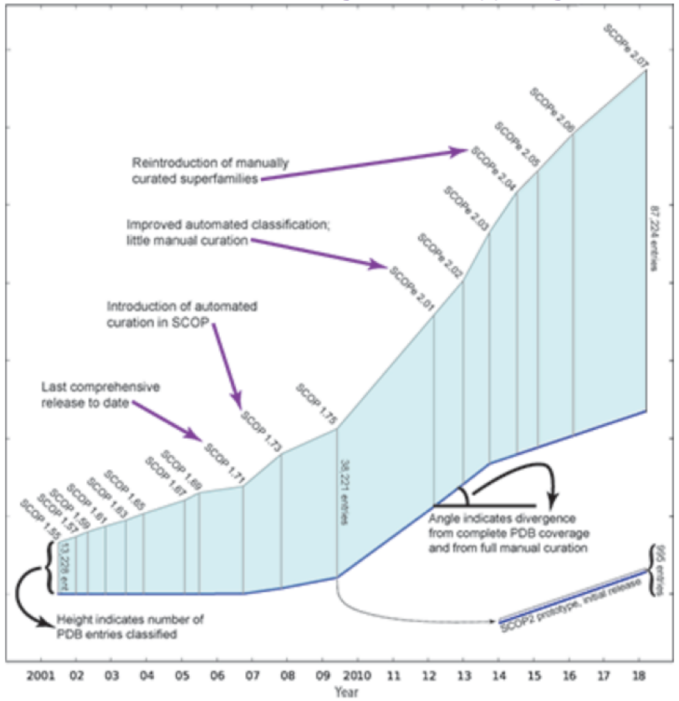
2019-04-11: New PDB entries were added in a periodic update; for more info on these updates, see the online documentation.

2019-03-05: We added an additional archive of PDB-style coordinate files for domains that were inadvertently omitted from our coordinate file archives.

2018-11-30: We published a paper describing updates to **SCOPe**, focusing on our findings from classifying large structres. [PDF].

2018-03-02: **SCOPe 2.07-stable** has been released, with nearly 10,000 new PDB entries added since the last stable release. Click either the About or Stats & History links for more details on what's new!

Click for information about changes to SCOP(e) design and size.



Classes in SCOPe 2.07:

1. a: All alpha proteins [46456] (289 folds)
2. b: All beta proteins [48724] (178 folds)
3. c: Alpha and beta proteins (a/b) [51349] (148 folds)
4. d: Alpha and beta proteins (a+b) [53931] (388 folds)

Classes in SCOPe 2.07:

1. a: All alpha proteins [46456] (289 folds)
2. b: All beta proteins [48724] (178 folds)
3. c: Alpha and beta proteins (a/b) [51349] (148 folds)
4. d: Alpha and beta proteins (a+b) [53931] (388 folds)
5. e: Multi-domain proteins (alpha and beta) [56572] (71 folds)
6. f: Membrane and cell surface proteins and peptides [56835] (60 folds)
7. g: Small proteins [56992] (98 folds)
8. h: Coiled coil proteins [57942] (7 folds)
9. i: Low resolution protein structures [58117] (25 folds)
10. j: Peptides [58231] (148 folds)
11. k: Designed proteins [58788] (44 folds)
12. l: Artifacts [310555] (1 fold)

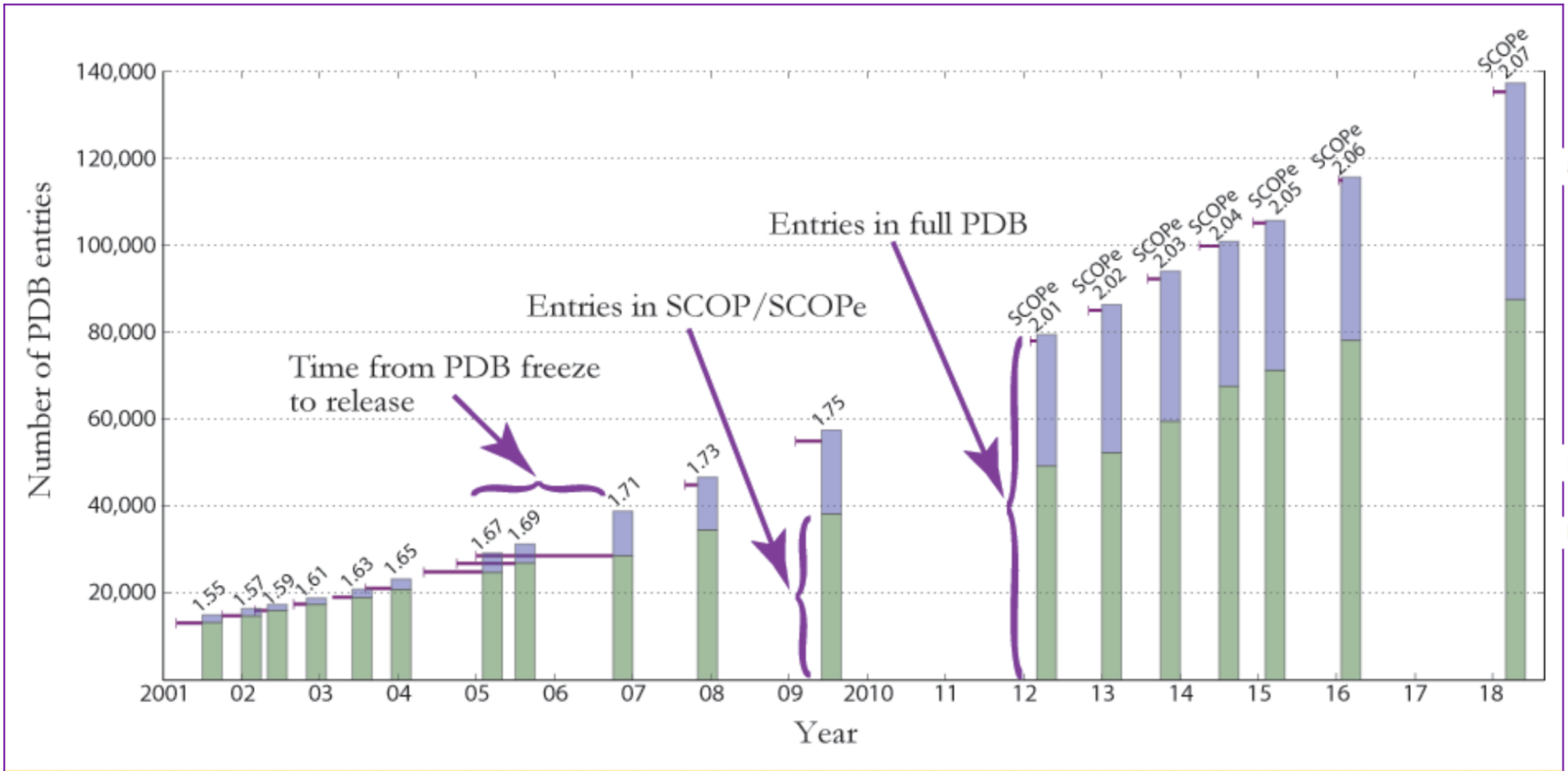
SCOPe: Structural Classification of Proteins — extended. Release 2.07 (updated 2019-04-11, stable release March 2018)

scope@compbio.berkeley.edu

Changes to SCOP(e) design and size

Applies to: SCOP version 1.55 through current release | References: 1-4,7

The figure below shows the number of structures in the PDB and SCOP(e) at the time of each SCOP(e) release. Extended horizontal lines start at the freeze date for each SCOP(e) release and show the number of PDB entries available on that date. (The "freeze date" is the last date for PDB entries to be released and still classified in a given SCOP(e) release. Prior to SCOP 1.73, all protein structures available on the freeze date were manually classified.)



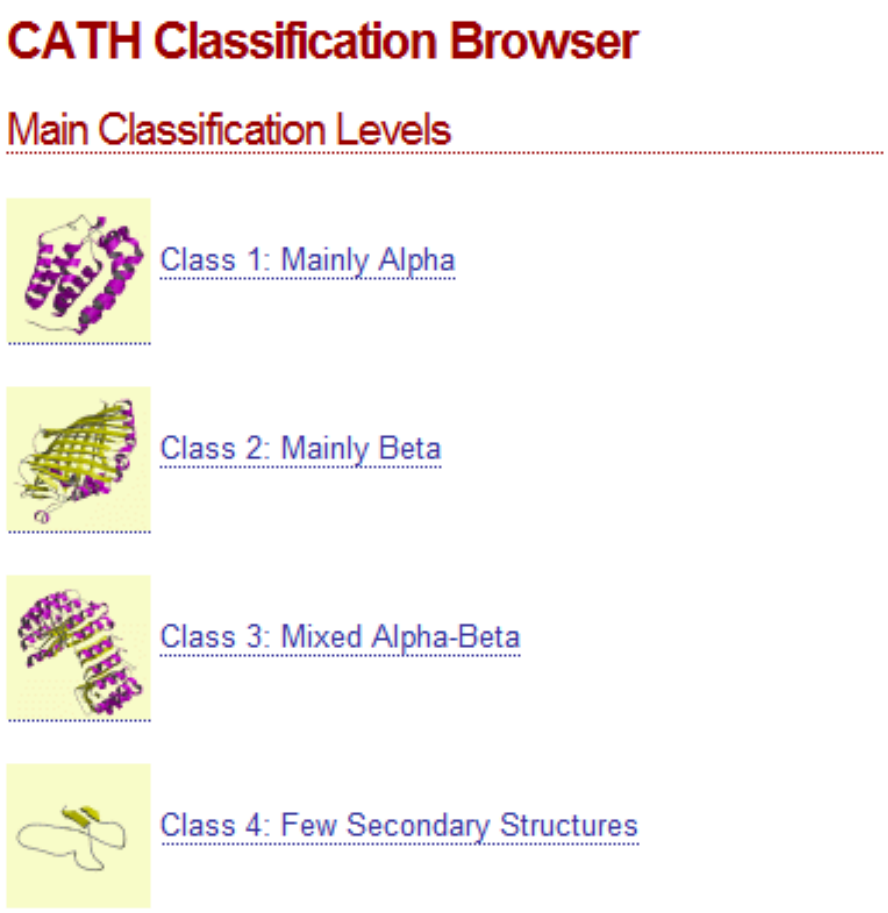
Timeline of SCOP(e) releases

Note that since releases beyond SCOP 1.71 are not comprehensive, not all structurally characterized protein families and folds from the PDB are classified in these releases. Therefore, we caution against using later releases to (for example) analyze the rate at which new folds are being discovered.

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

CATH is a hierarchical classification of protein **domain** structures, which clusters proteins at four major levels: Class (C), Architecture (A), Topology (T) and Homologous superfamily (H). The boundaries and assignments for each protein domain are determined using a combination of automated and manual procedures which include computational techniques, empirical and statistical evidence, literature review and expert analysis

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



Class

Similar secondary structure content

All α, all β,αβ, alternating α/β,…

Fold (Architecture)

Major structural similarity

SSE's in similar arrangement

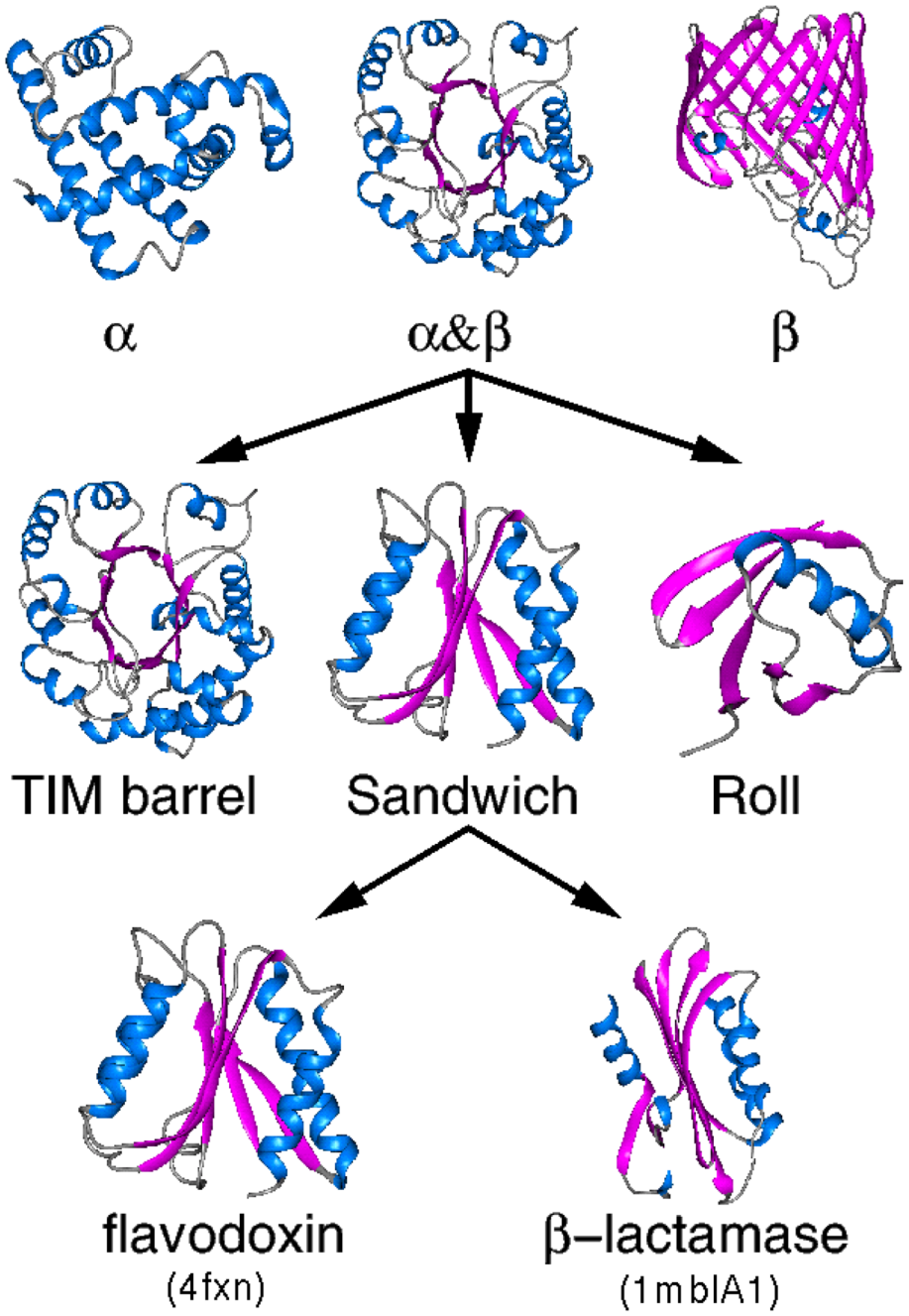
Superfamily (Topology)

Probable common ancestry

HMM family membership

Family

Clear evolutionary relationship



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

26 million protein domains classified into 2,738 superfamilies

Browse » Search » Download » Take the Tour »

What is CATH?

CATH is a classification of protein structures downloaded from the Protein Data Bank. We group protein domains into superfamilies when there is sufficient evidence they have diverged from a common ancestor.

- Search CATH by text, ID or keyword
- Search CATH by protein sequence (FASTA)
- Search CATH by PDB structure
- Browse CATH Hierarchy
- CATH Release Statistics
- CATH Tutorials

Example pages

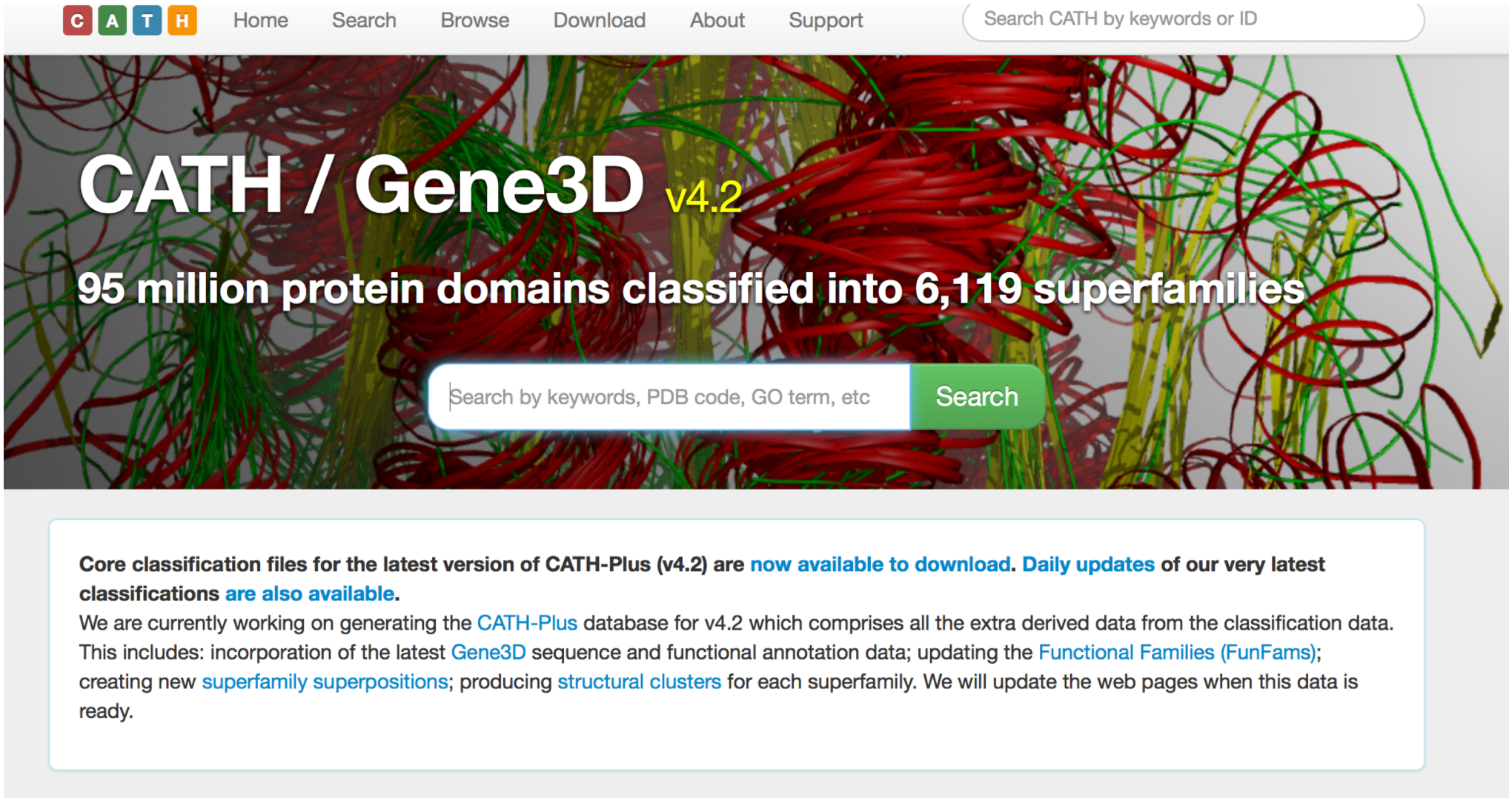
- PDB "2bop"
- Domain "1cukA01"
- Relatives of "1cukA01"
- Superfamily "HUPs"
- Functional Family
- FunFam Alignment
- Search for "enolase"
- Superfamily Comparison

Latest Release Statistics

CATH v4.0 based on PDB dated March 26, 2013	
235,858	CATH Domains
2,738	CATH Superfamilies
69,058	Annotated PDBs

Gene3D v12 released March 18, 2012	
6,131	Cellular Genomes
21,662,155	Protein Sequences
25,615,754	CATH Domain Predictions

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



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

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

What is CATH-Gene3D?

CATH is a classification of protein structures downloaded from the Protein Data Bank. We group protein domains into superfamilies when there is sufficient evidence they have diverged from a common ancestor.

- Search CATH by text, ID or keyword
- Search CATH by protein sequence
- Search CATH by PDB structure
- Browse CATH Hierarchy
- CATH Release Statistics
- CATH Tutorials

Gene3D uses the information in CATH to predict the locations of structural domains on millions of protein sequences available in public databases. This allows us to include additional annotations to the CATH-Gene3D database such as functional information and active site residues.

- Go to Gene3D
- Compare Genomes
- Download Gene3D Data
- Learn how Gene3D is created

If you have any questions, comments or suggestions please get in touch via Twitter, ask a question in our online forum or visit our support page.

Latest Release Statistics

Info

	CATH-Plus 4.2.0	CATH (daily snapshot)
PDB Release	17-05-2017	5 days ago
Domains	434857	464208
Superfamilies	6119	6892
Annotated PDBs	131091	138047

	Gene3D v16
Protein Sequences	52,073,853
CATH Domain Predictions	95,665,487

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

What is CATH-Gene3D?

CATH is a classification of protein structures downloaded from the Protein Data Bank. We group protein domains into superfamilies when there is sufficient evidence they have diverged from a common ancestor.

- Search CATH by text, ID or keyword
- Search CATH by protein sequence
- Search CATH by PDB structure
- Browse CATH Hierarchy
- CATH Release Statistics
- CATH Tutorials

Gene3D uses the information in CATH to predict the locations of structural domains on millions of protein sequences available in public databases. This allows us to include additional annotations to the CATH-Gene3D database such as functional information and active site residues.

- Go to Gene3D
- Compare Genomes
- Download Gene3D Data
- Learn how Gene3D is created

Latest Release Statistics

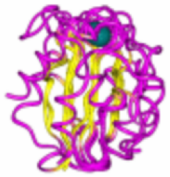
Info

	CATH v4.1	CATH-B
PDB Release	01-01-2015	17 days ago
Domains	308999	436020
Superfamilies	2737	6344
Annotated PDBs	108378	128287

	Gene3D v14
Cellular Genomes	19,471
Protein Sequences	43,387,462
CATH Domain Predictions	53,479,436

Superfamily 1.75

HMM library and genome assignments server



Search SUPERFAMILY

Home

SEARCH

Keyword search

Sequence search

BROWSE

Organisms

- Taxonomy
- Statistics

SCOP

- Hierarchy

Ontologies

- GO
- EC
- Phenotype

TOOLS

Compare genomes

Phylogenetic trees

Web services

Downloads

ABOUT

Description

Display a menu

Publications

SUPERFAMILY

G+1 +38 Recommend this on Google

Follow @SUPERFAMILY

SUPERFAMILY is a database of structural and functional annotation for all proteins and genomes.

The SUPERFAMILY annotation is based on a collection of **hidden Markov models**, which represent structural protein domains at the SCOP superfamily level. A superfamily groups together domains which have an evolutionary relationship. The annotation is produced by scanning protein sequences from over **2,478 completely sequenced genomes** against the hidden Markov models.

For each **protein** you can:

- Submit sequences for SCOP classification
- View domain organisation, sequence alignments and protein sequence details

For each **genome** you can:

- Examine superfamily assignments, phylogenetic trees, domain organisation lists and networks
- Check for over- and under-represented superfamilies within a genome

For each **superfamily** you can:

- Inspect SCOP classification, functional annotation, Gene Ontology annotation, InterPro abstract and genome assignments
- Explore taxonomic distribution of a superfamily across the tree of life

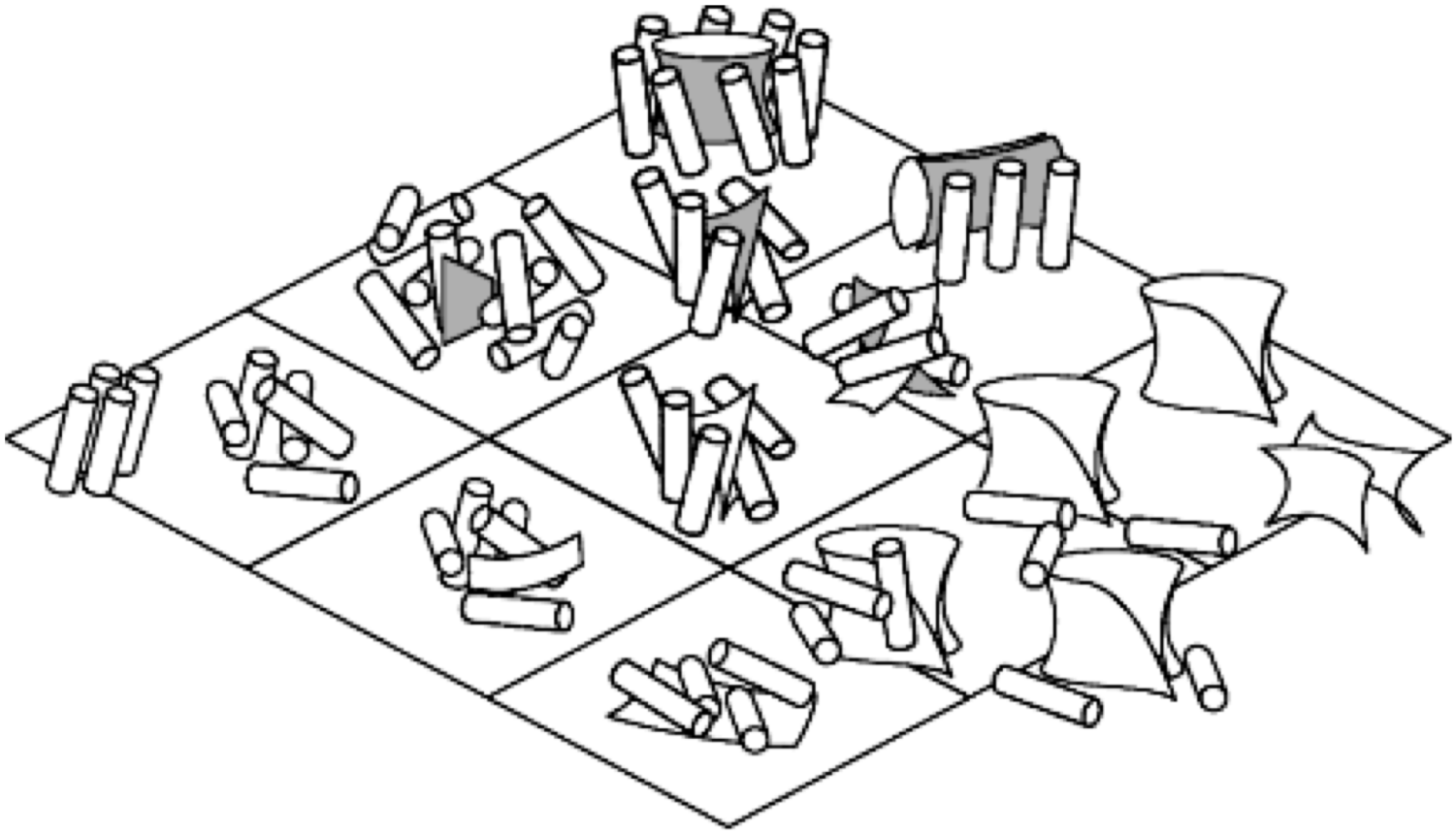
All annotation, models and the database dump are freely available for download to everyone. Description cont.

Jump to [SUPERFAMILY description · Recent news]

http://supfam.org/SUPERFAMILY/cgi-bin/gen_list.cgigenome=Hs


SUPERFAMILY Under Maintenance
Back on Tuesday 23rd April
Thank you for your patience

Structural classes of proteins

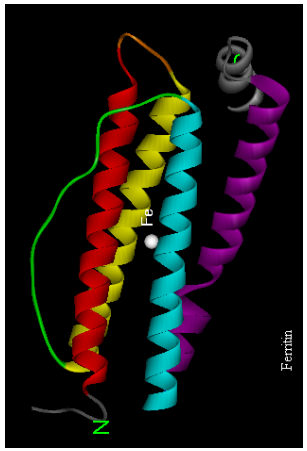


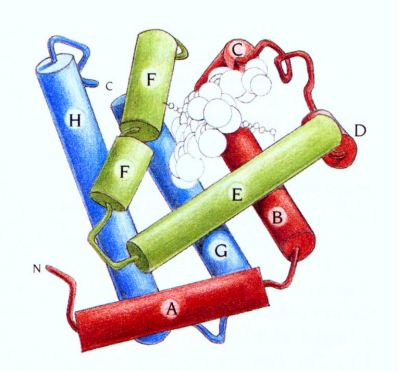
Others:
Multi-domain, membrane and cell surface, small proteins, peptides and fragments, designed proteins, ..

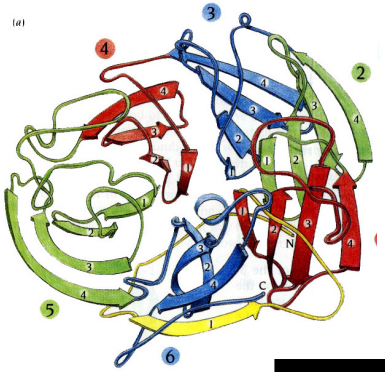
Folds/Architectures

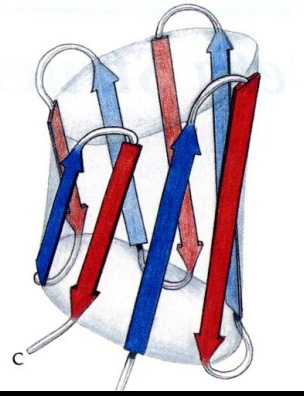
Mainly α
- Bundle
- Non-Bundle

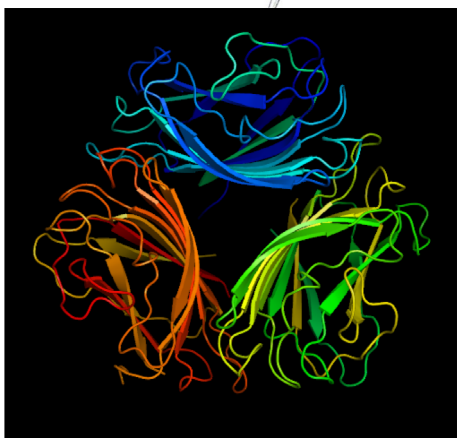
Mainly β
- Single sheet
- Roll
- Barrel
- Clam
- Sandwich
- Prism
- 4/6/7/8 Propeller
- Solenoid







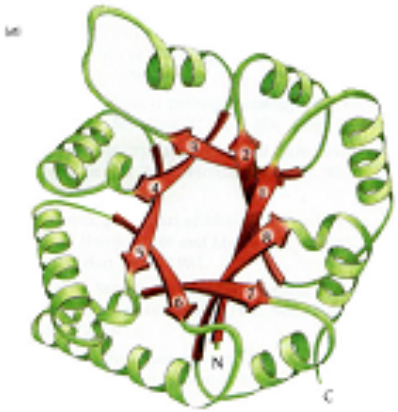




α/β and α+β

Closed
- Barrel
- Roll, ...

Open
- Sandwich
- Clam, ...



Fold versus topology!

Up-and-down β barrel

Jelly Roll Motif

Immunoglobulin Fold

Předpověď 3D-struktury/ foldu

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

- Threading - „navlékání“
- Homology modeling
- *Ab inicio* metody

Metody pro predikci funkce

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

- Analýza 2D struktu
  identifikuje nekte

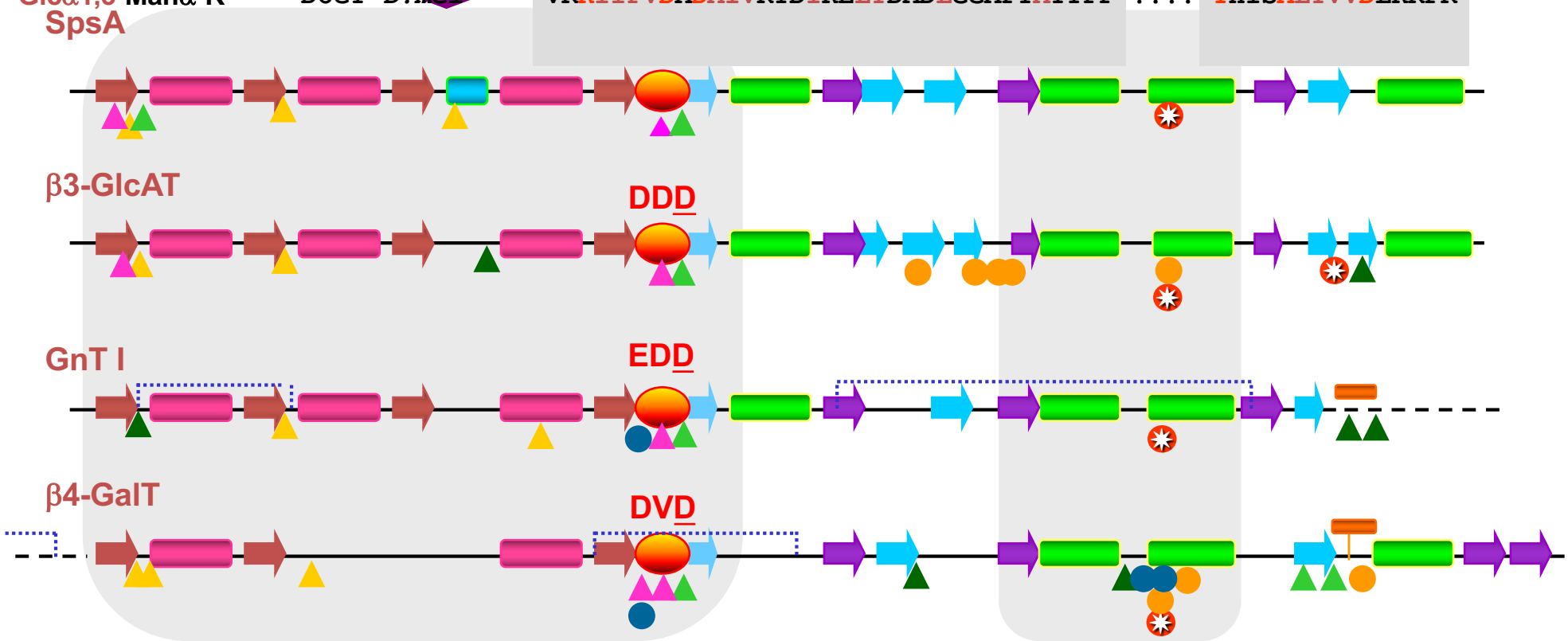
Galα1,4-Galβ-R	LgtC *N.men*	CDKVLYLDIDVLVRDSLTPLWDTDLGDNWLGACID	....	YFNAGVLLINLKKWR
Glcα1,3-Glcα-R	RfaI *E.coli*	APKVLYLDADIICQGTIEPLINFSFPDDKVAMVVT	....	YFNSGFLLINTAQWA
Galα1,3-Glcα-R	RfaI *S.typh*	QIKVLYLDADIACKGSIQELIDLNFAENEIAAVVA	....	YFNAGFILIXIPLWT
Glcα1,2-Glcα-R	RfaJ *E.coli*	LDRLLYLDADVVCKGDISQLLHLGLN-GAVAAVVK	....	YFNSGVVYLDLKKWA
Galα1,6-Manα-R	LpcA *R.leg*	IERLLYLDADVLAVSPVDELFTRNFQGKALAAVDD	.....	YFNAGVLLFDWSACR
Glcα1,3-Manα-R	DUGT *D.mel*	VRKIIFVDADAIVRTDIKELYDMDLGGAPYAYTPF	....	YHISALYVVDLKRFR

Manthing

« Rossmann-

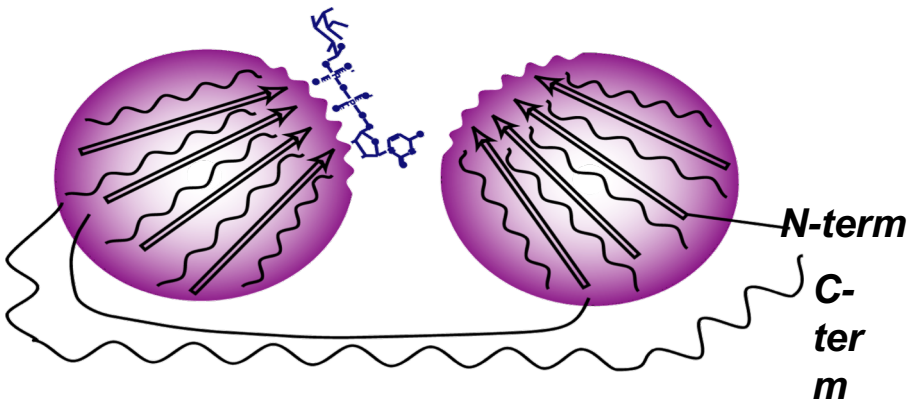
(1[illegible]0-12[illegible]

or re



Dvě pozorované topologie 3D struktur glykosyltransferas

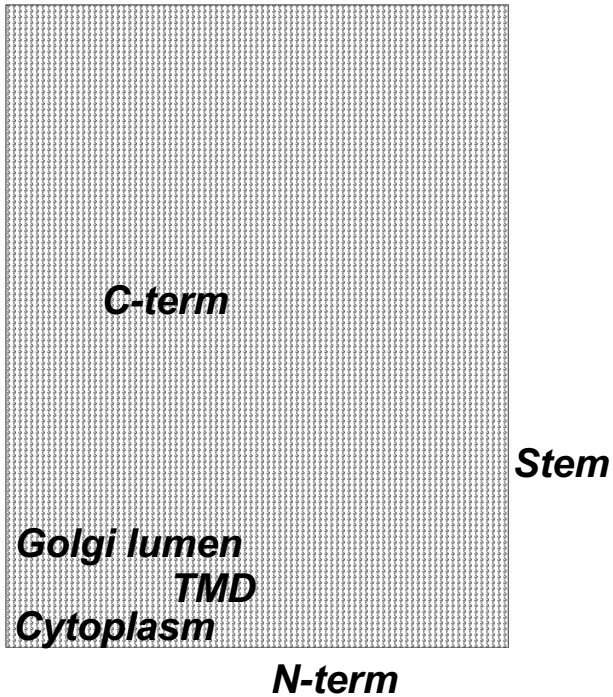
BGT-fold



(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



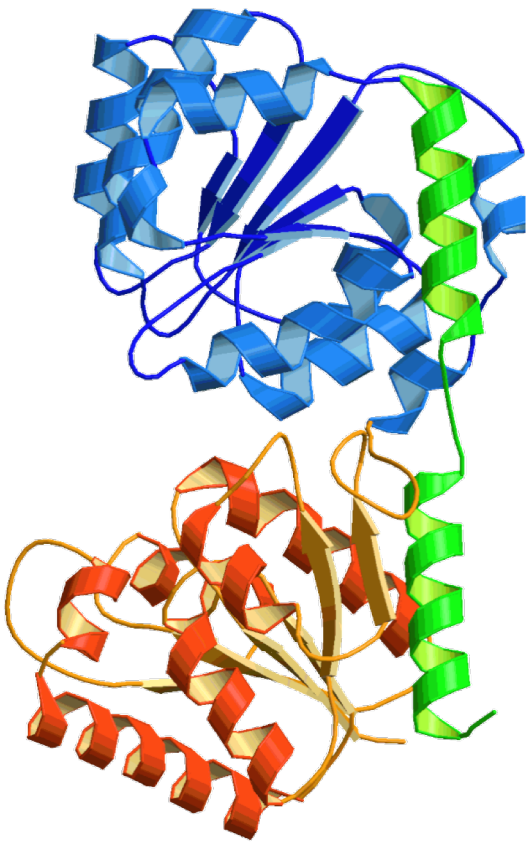
(Procaryotes)

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

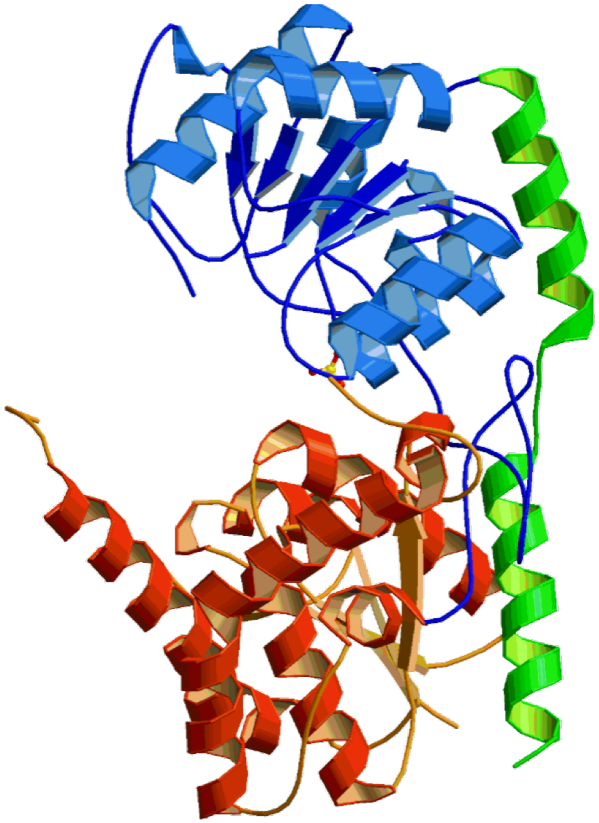
(Eucaryotes)

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

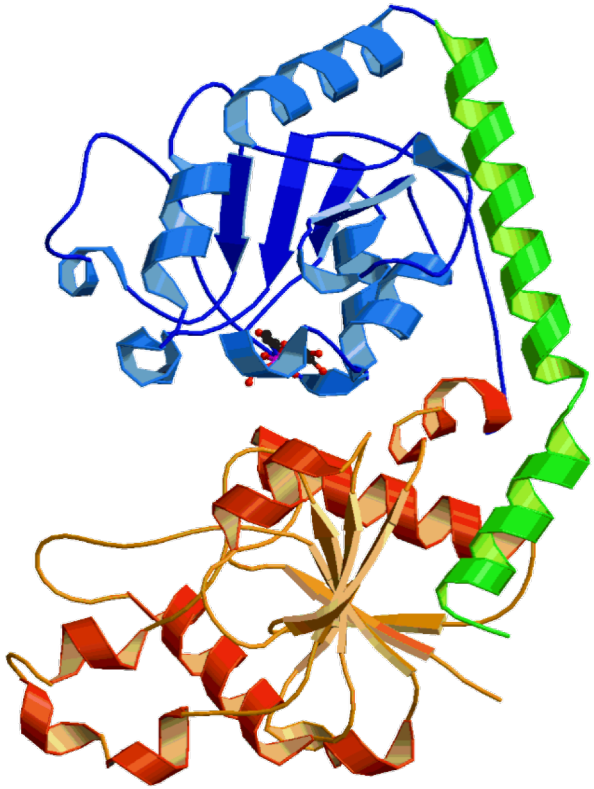
Nadrodina s BGT foldem



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



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

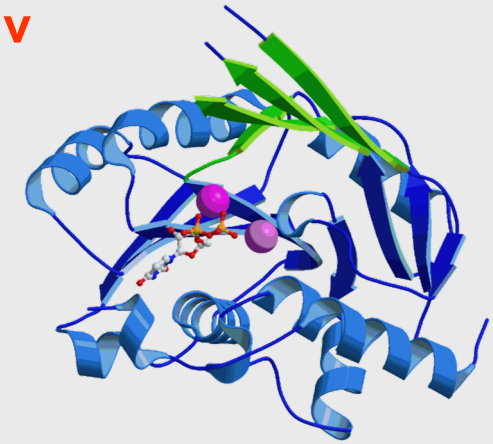


BGT (β-GlcT)
n.c.
Phage T4
Vrielink *et al*, 1994

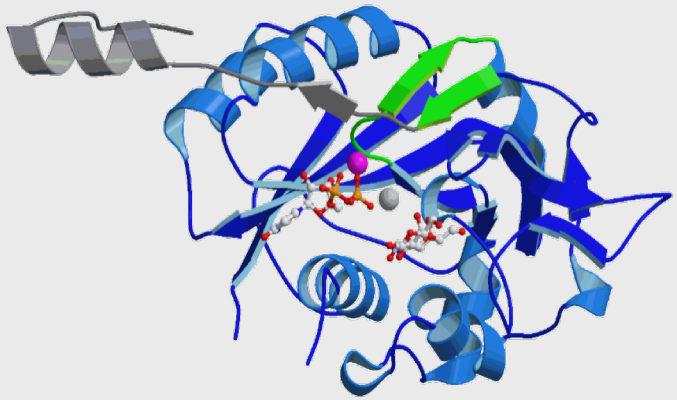
Nadrodina s SpsA foldem

Společná NBD

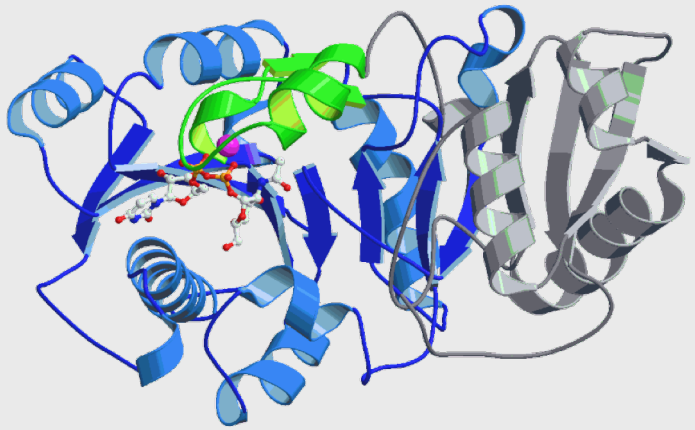
Inv



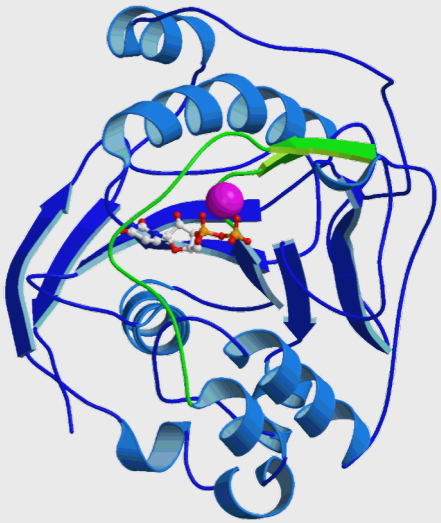
SpsA [GT2]
Charnok *et al*, 1999, 2001



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

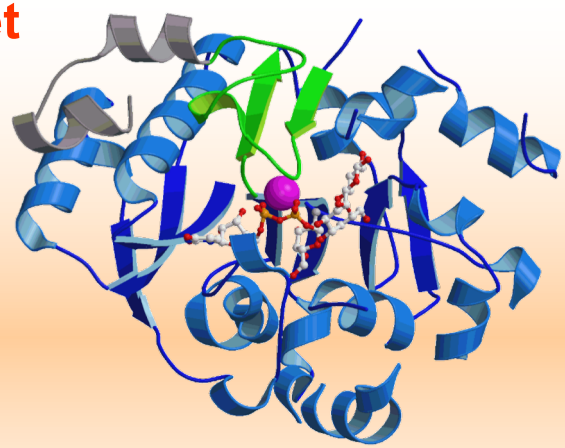


Rabbit GnT I [GT13]
Ünligil *et al*, 2000

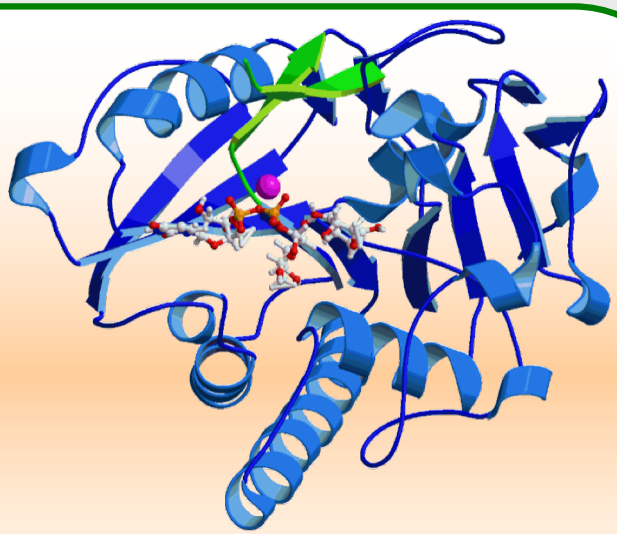


Bovine β4-GalT [GT7]
Gastinel *et al*, 1999
Ramakrishnan *et al*, 2001, 2002

Ret



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



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

Threading

- „navlékání“ = rozpoznání a přiřazení proteinového foldu aminokyselinové sekvenci

- sekvence je porovnávána s databází existujících foldů (3D profilů) a na jejich základě jsou konstruovány 3D- modely

- 3D profil - každému reziduu v 3D struktuře je přiřazena environmentální proměnná (obsah polárních atomů v postranním řetězci, skrytá plocha, sekundární elementy, apod.) vycházející z předpokladu, že okolí rezidua je více konzervováno než aminokyselina samotná.

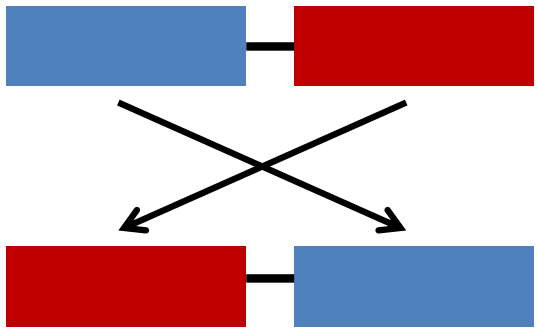
- Reziduum může být také popsáno pomocí svých interakcí

- Výsledná kvalita modelu shoda je popsána pomocí Z-skóre nebo energie

- U 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
```

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

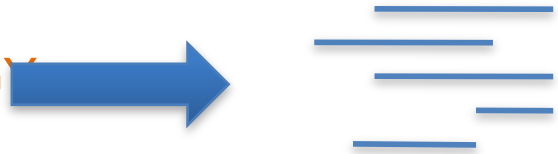
https://zhanglab.ccmb.med.umich.edu/I-TASSER/

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

ARDLVIPMIYCGHGY

User sequence



Homologous sequences

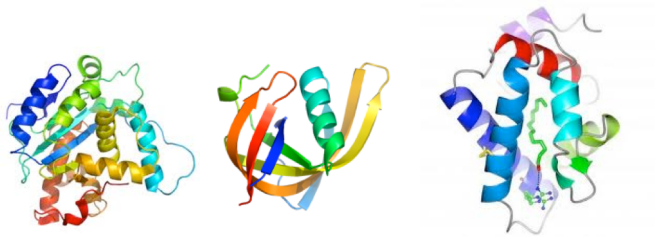
Search the 10 million known sequences for homologues using PSI-Blast.

Phyre2

Capture the mutational propensities at each position in the protein

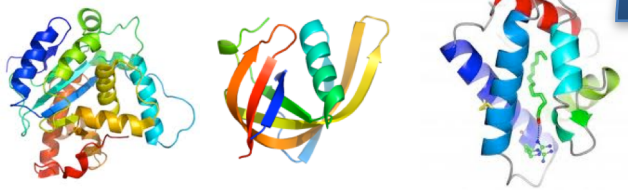
An evolutionary fingerprint

Phyre2

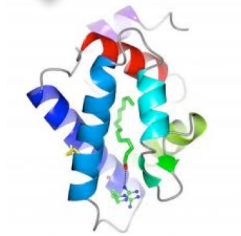


~ 65,000 known 3D structures

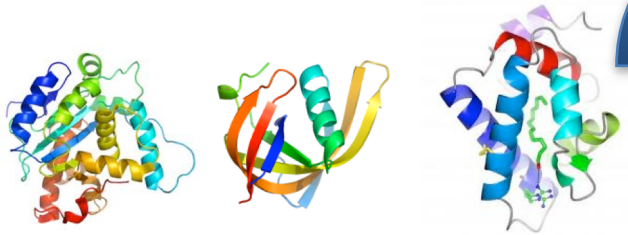
Phyre2



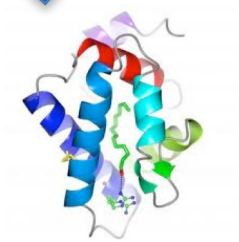
~ 65,000 known 3D structures



Phyre2



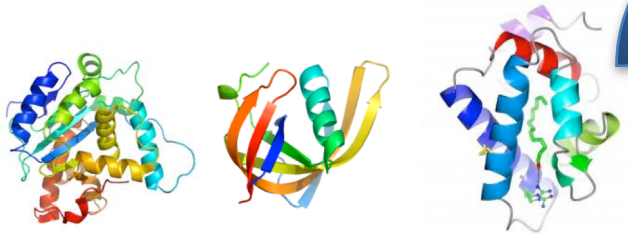
~ 65,000 known 3D structures



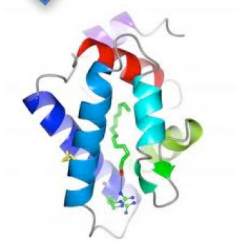
Extract sequence

HAPTLVRDC…….

Phyre2



~ 65,000 known 3D structures

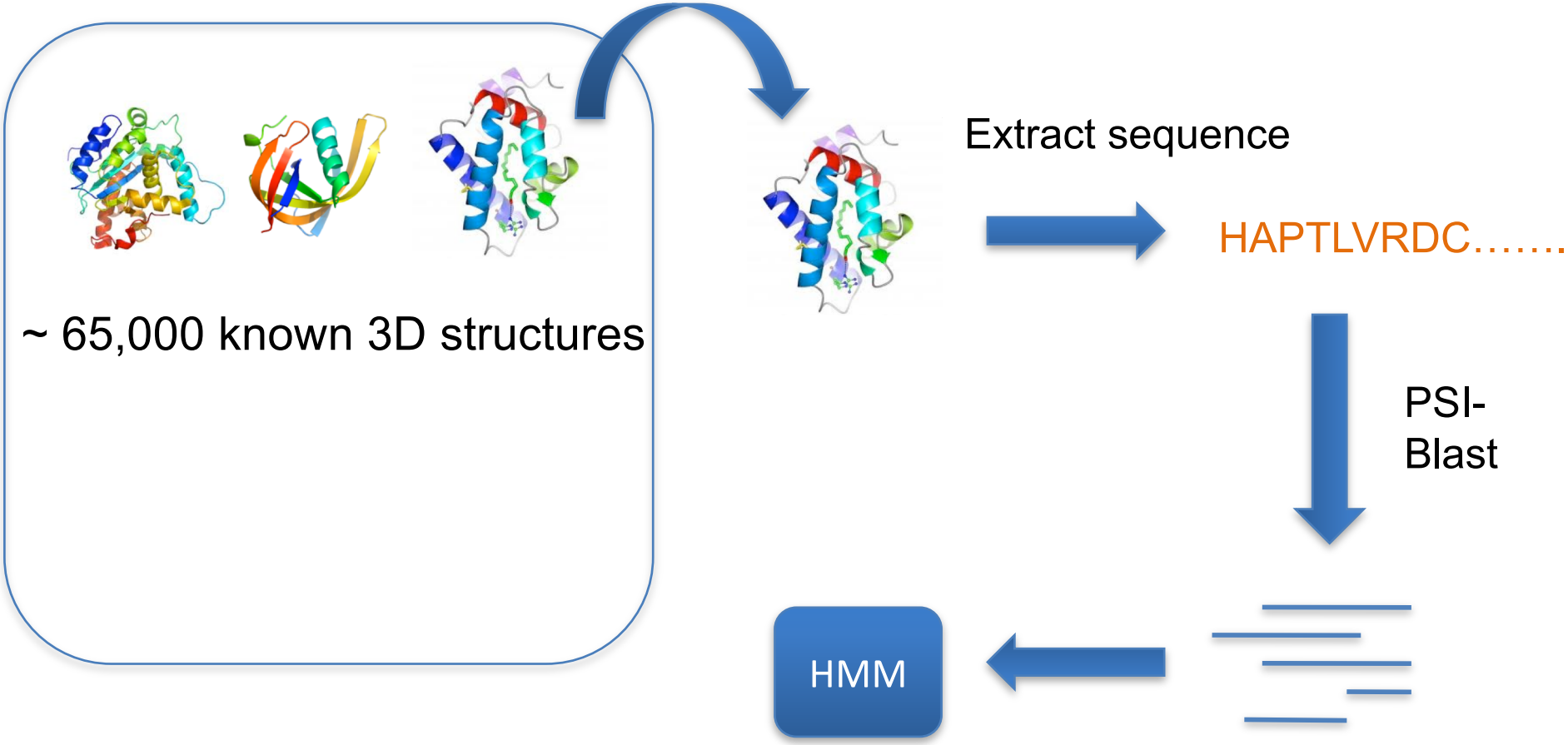


Extract sequence

HAPTLVRDC…….

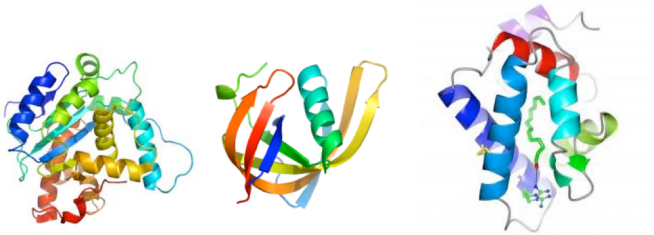
PSI-
Blast

Phyre2



Hidden Markov model
for sequence of KNOWN structure

Phyre2

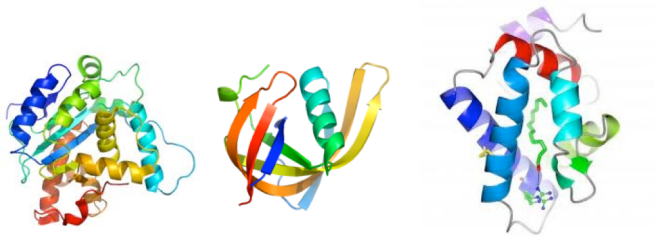


~ 65,000 known 3D structures

HMM HMM HMM

~ 65,000 hidden Markov models

Phyre2



~ 65,000 known 3D structures

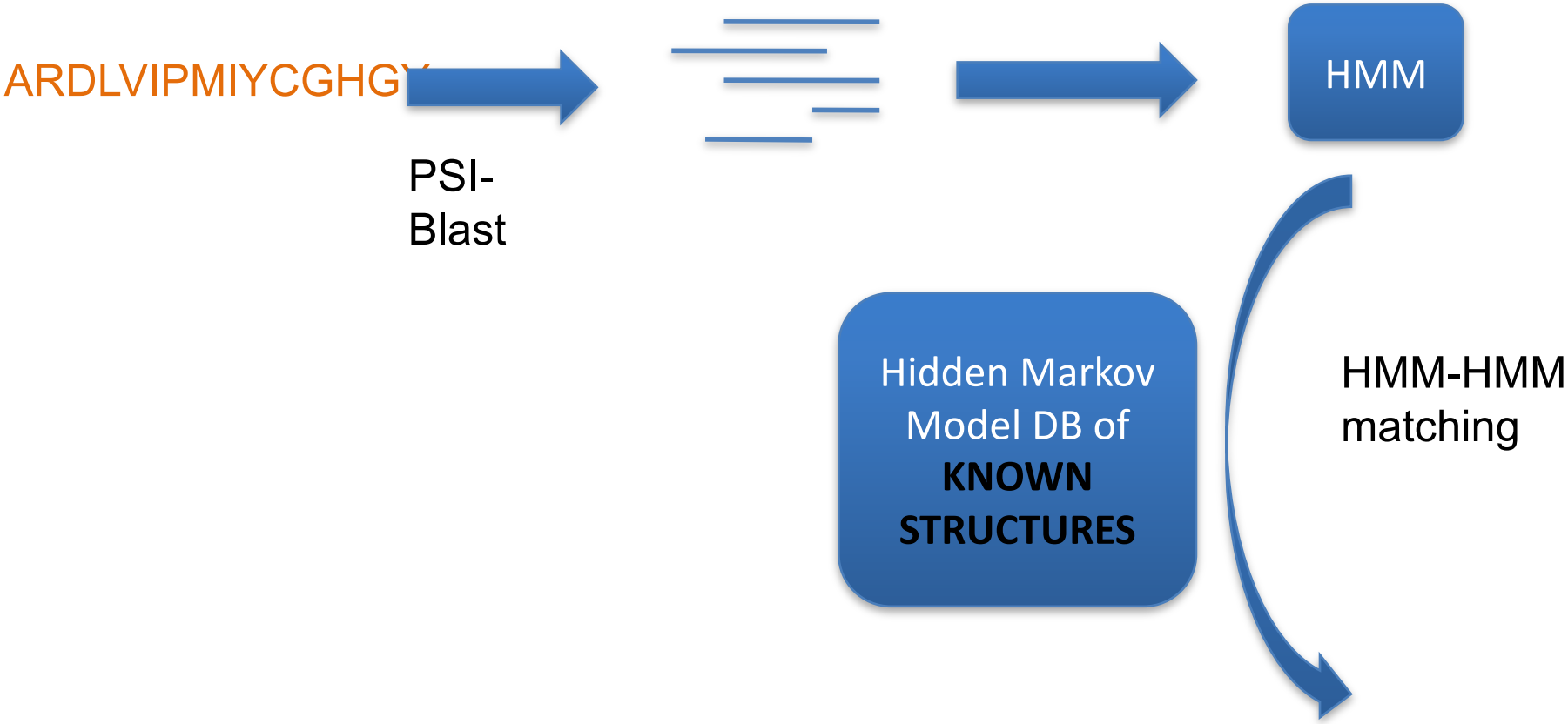
Hidden Markov Model Database of **KNOWN STRUCTURES**

Phyre2

Capture the mutational propensities at each position in the protein

An evolutionary fingerprint

Phyre2

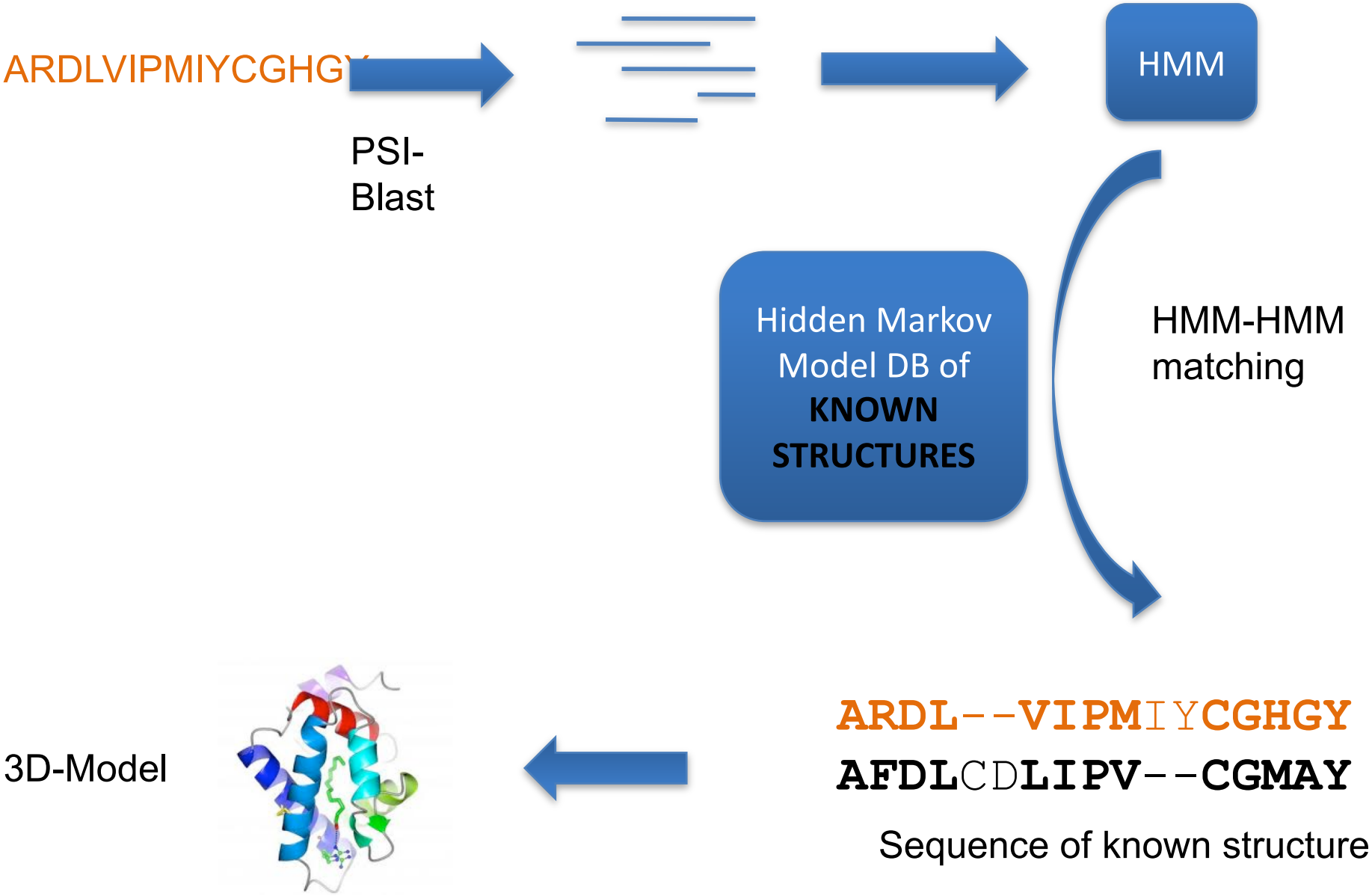


Alignments of user sequence to known structures ranked by confidence.

```
ARDL--VIPMIYCGHGY
AFDLCDLIPV--CGMAY
```

Sequence of known structure

Phyre2


ARDLVIPMIYCGHGY
PSI-
Blast
HMM
Hidden Markov
Model DB of
KNOWN
STRUCTURES
HMM-HMM
matching
ARDL--VIPMIYCGHGY
AFDLCDLIPV--CGMAY
Sequence of known structure
3D-Model

Phyre2

ARDLVIPMIYCGHGY

PSI-
Blast

HMM

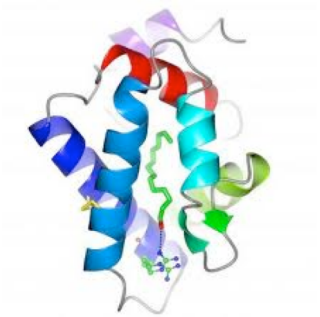
Hidden Markov
Model DB of
KNOWN
STRUCTURES

HMM-HMM
matching

Very powerful –
able to reliably detect extremely remote homology

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

3D-Model



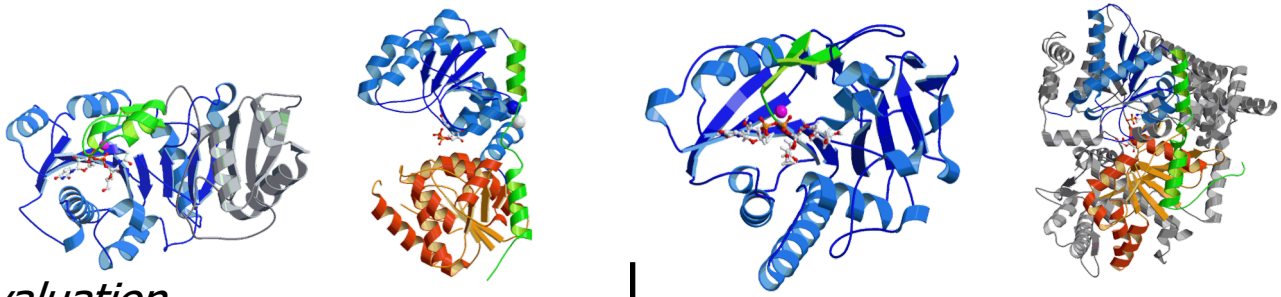
ARDL--VIPMIYCGHGY
AFDLCDLIPV--CGMAY

Sequence of known structure

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

results for job synt____ - Mozilla Firefox

vy Zobrazit Historie Záložky Nástroje Nápověda

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

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

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

To predict functional residues and GO classification, try ConFunc

cognition

nments	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/g phosphorylase
	d1rzua (length:477) **14% i.d.**		6.1e-36	100 %	n/a	UDP-Glycosyltransferase/glycogen phosphorylase	UDP-Glycosyltransferase/g 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

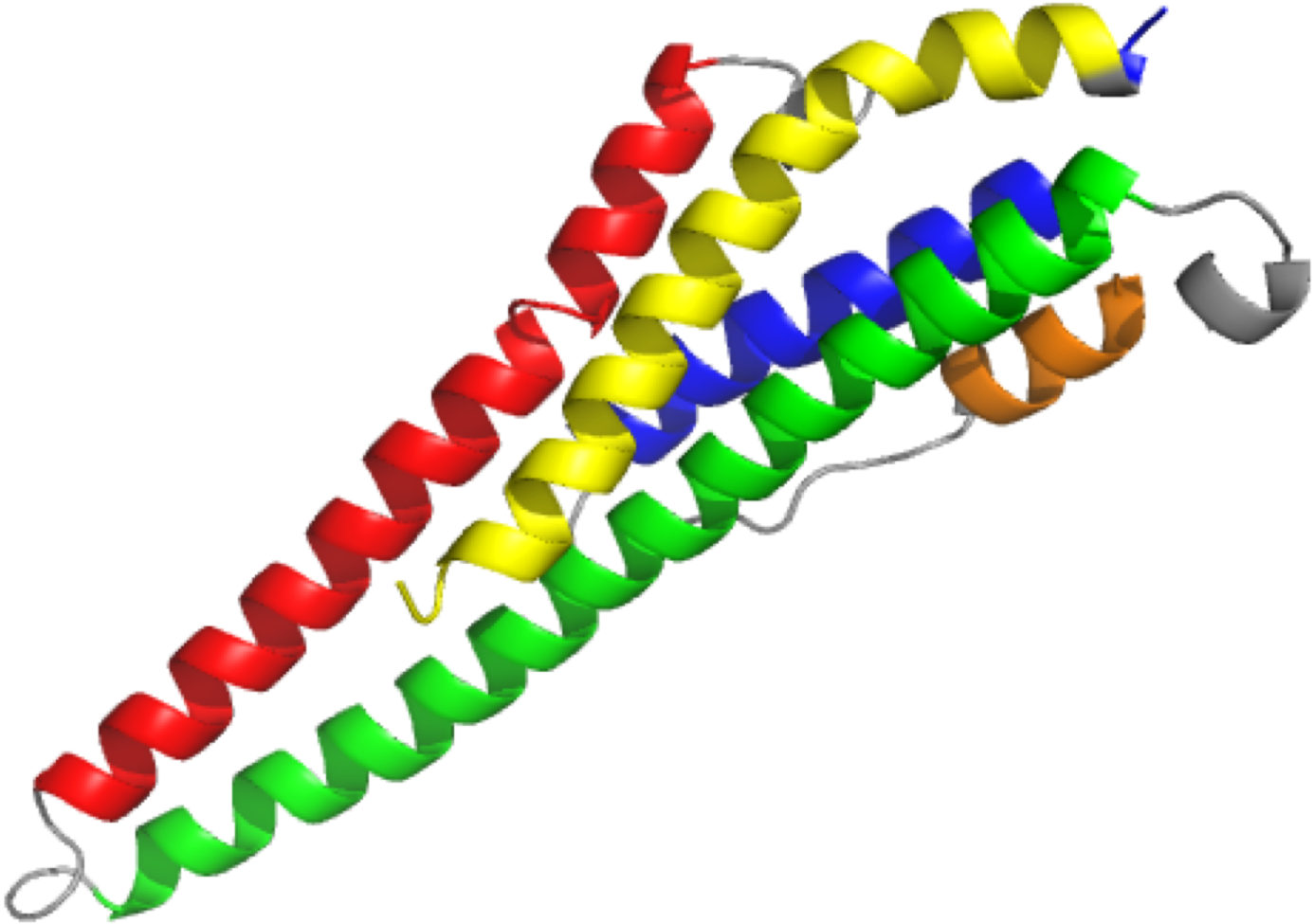
Nejnavštěvovanější Jak začít Přehled zpráv http://www.ncbi.nlm.... http://www.glycoscie... CHMI Radar Departme...

Fold Recognition

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

AB2L structure overview

Structure: 4 helical bundle



Top model

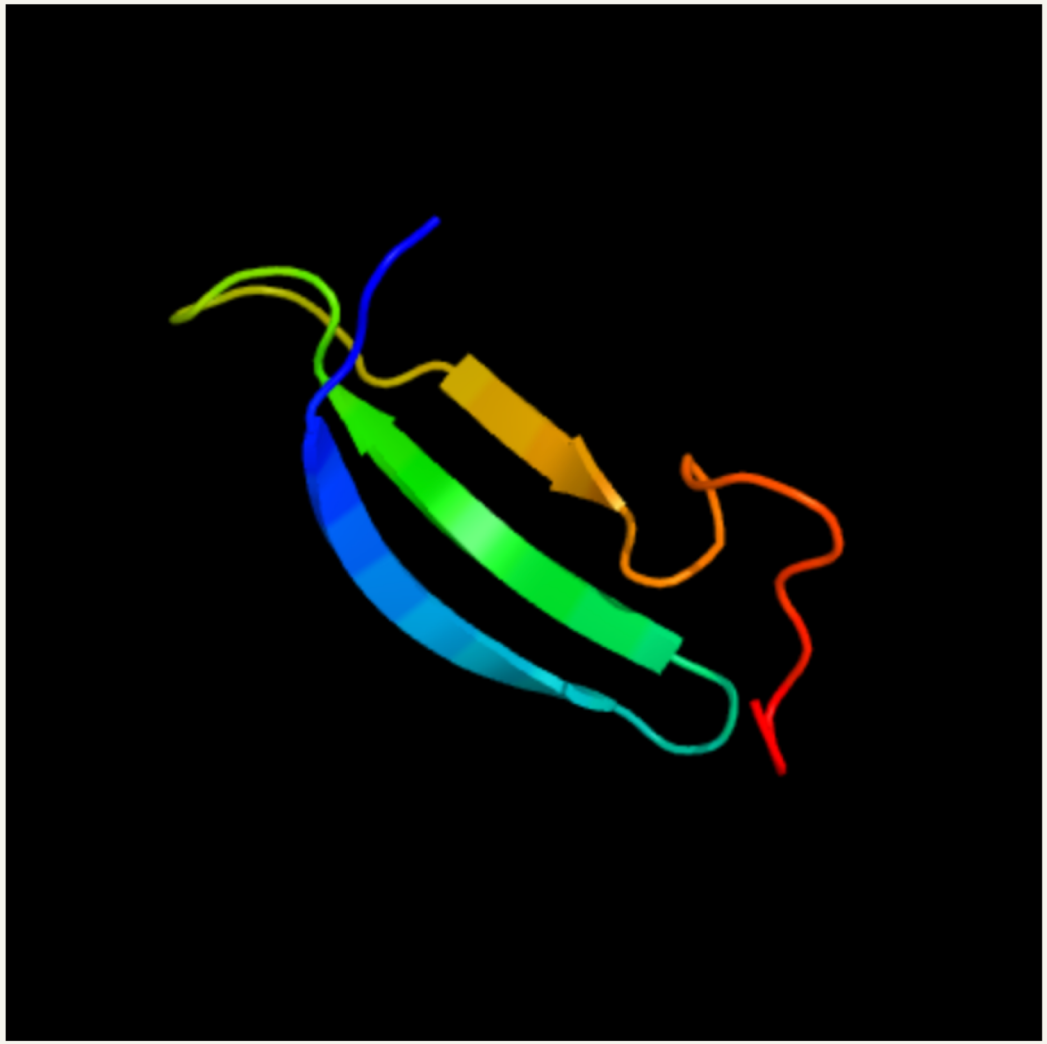


Image coloured by rainbow N → C terminus

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.

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

Prozkoumání možností a principů fungování I-TASSERu bude domácím úkolem

Homology modeling

•přiložení cílové sekvence se sekvencí homologního proteinu se známou 3D strukturou

• extrakce uhlíkové páteře ze struktury templátu a umístění postranních řetězců

•modelování otoček a smyček

• minimalizace energie

•validace modelované struktury

MODELLER

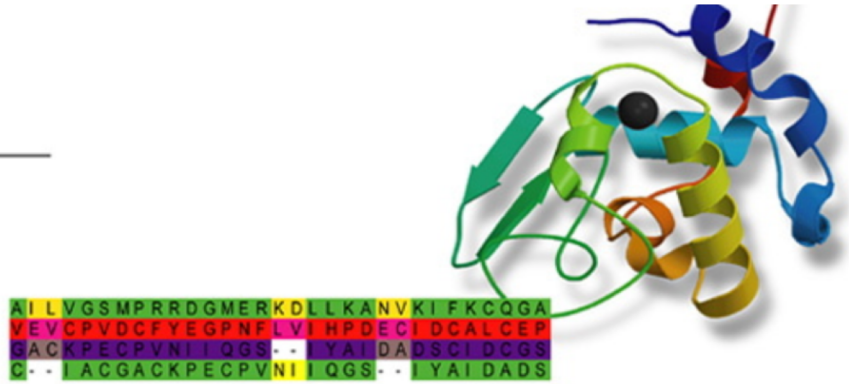
Mostly used program in academic environment for serious homology modeling

SWISS-MODEL

An automated knowledge-based protein modelling server

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



About MODELLER

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (3,4), and can perform many additional tasks, including de novo modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc. MODELLER is available for download for most Unix/Linux systems, Windows, and Mac.

Several graphical interfaces to MODELLER are commercially available. There are also many other resources and people using Modeller in graphical or web interfaces or other frameworks.

1. B. Webb, A. Sali. Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics 54, John Wiley & Sons, Inc., 5.6.1-5.6.37, 2016.
2. M.A. Marti-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.
3. A. Sali & T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815, 1993.
4. A. Fiser, R.K. Do, & A. Sali. Modeling of loops in protein structures, Protein Science 9. 1753-1773, 2000.

The current release of Modeller is **9.21**, which was released on Dec 11th, 2018. Modeller is currently maintained by Ben Webb.

BIOZENTRUM
University of Basel
The Center for Molecular Life Sciences

SWISS-MODEL

Modelling | Repository | Tools | Documentation | Log in | Create Acc

Start a New Modelling Project

Target Sequence:
(Format must be FASTA, Clustal, plain string, or a valid UniProtKB AC)

Paste your target sequence(s) or UniProtKB AC here

+ Upload Target Sequence File... | Validate

Project Title: Untitled Project

Email: Optional

Search For Templates | Build Model

By using the SWISS-MODEL server, you agree to comply with the following terms of use and to cite the corresponding articles.

Supported Inputs

- Sequence(s)
- Target-Template Alignment
- User Template
- DeepView Project

You are currently not logged in - to take advantage of the workspace, please log in or create an account.

(There is no requirement to create an account to use any part of SWISS-MODEL, however you will gain the benefit of seeing a list of your previous modelling projects here.)

MODELLER

Mostly used program in academic environment for serious homology modeling

SWISS-MODEL

An automated knowledge-based protein modelling server

```
- Start SMR-Pipeline in automated mode on BC2-cluster at Thu May  2 08:51:47 2013


- Start BLAST for highly similar template structure identification
- No suitable templates found!

- Run HHSearch to detect remotely related template structures
- Unfortunately, we could not identify useful template structures

- For troubleshooting, please see our article in Nature Protocols:

- Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J. and Schwede, T. (2009).
Protein structure homology modelling using SWISS-MODEL Workspace.  Nature Protocols, 4, 1.
```

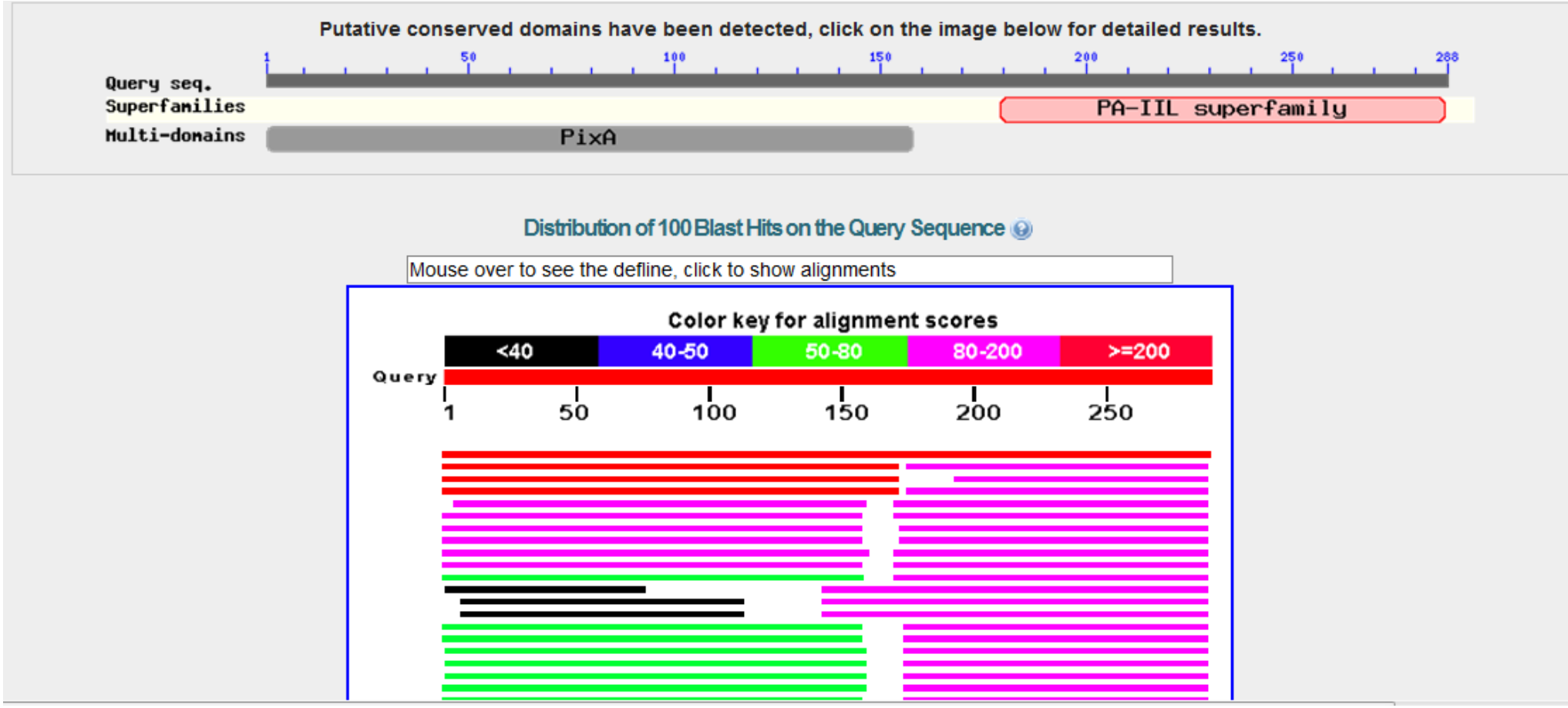
Ale!

! pozor na domény !

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

Prohledávání databází známých aminokyselinových sekvencí

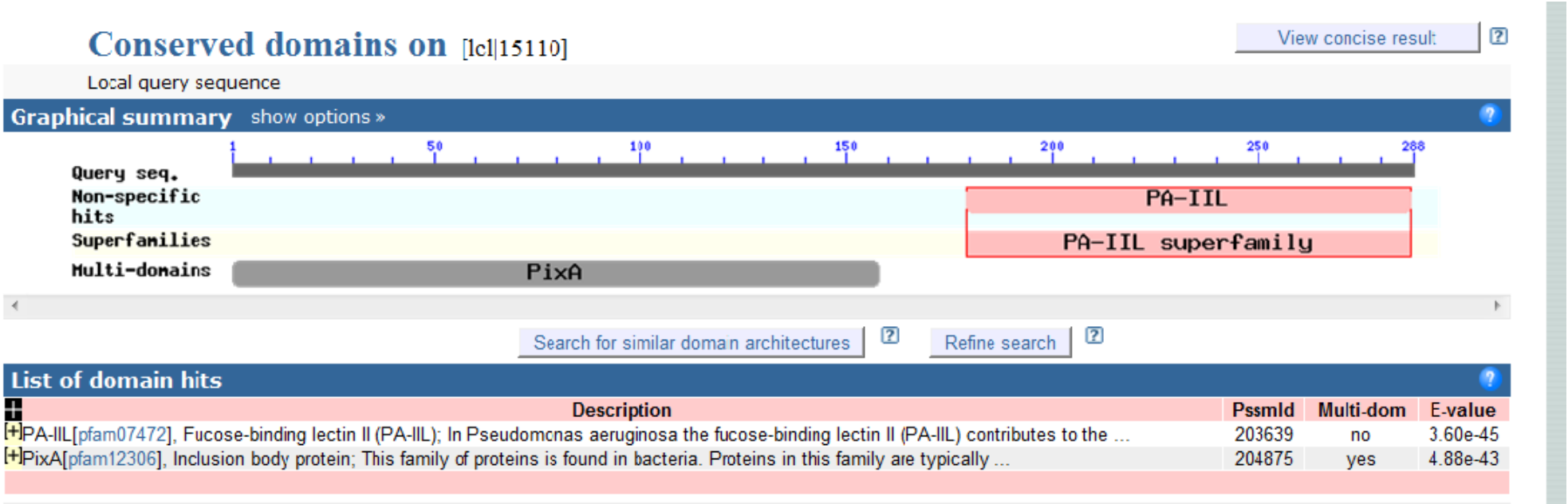
- **celý protein**



NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

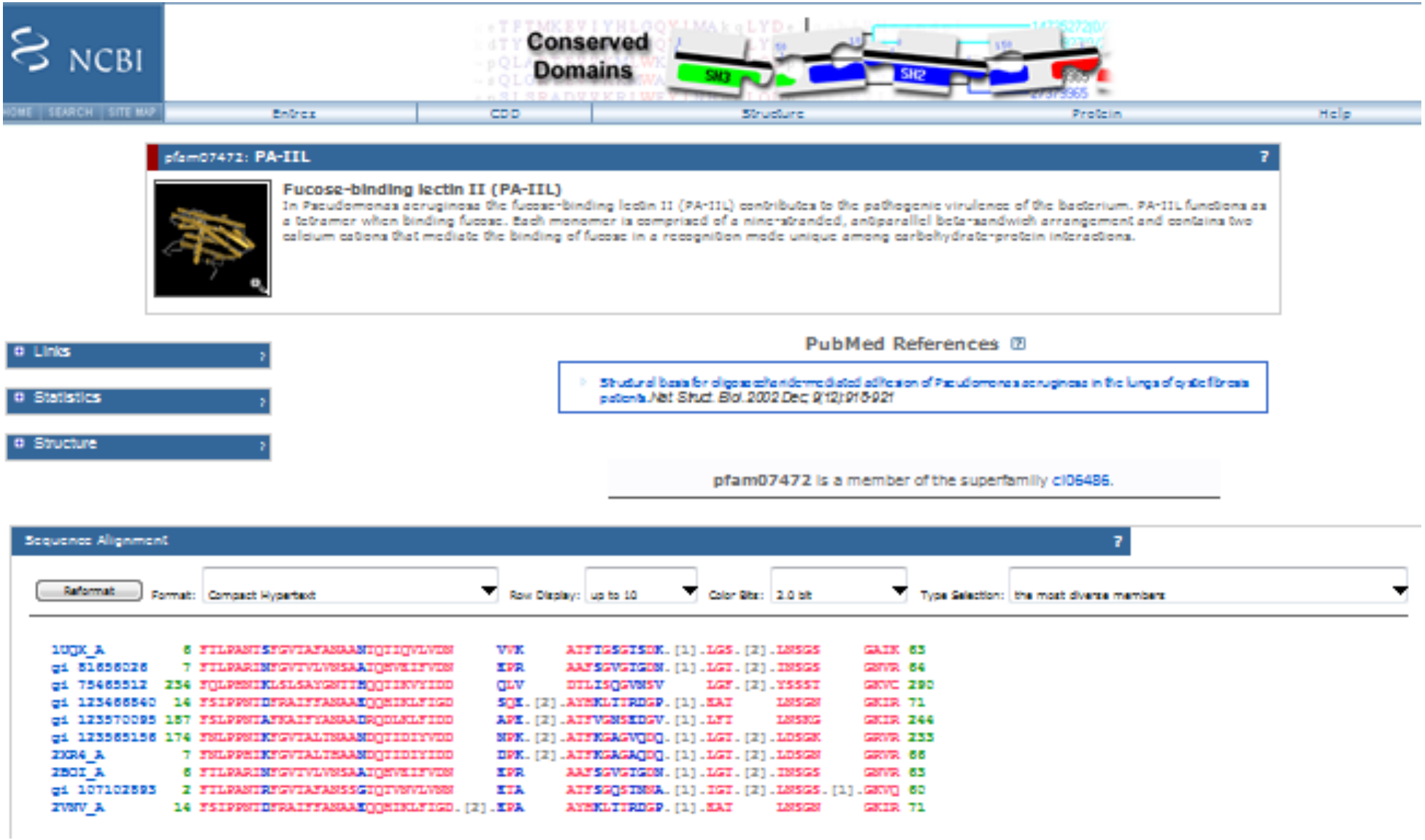
- **celý protein**



NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

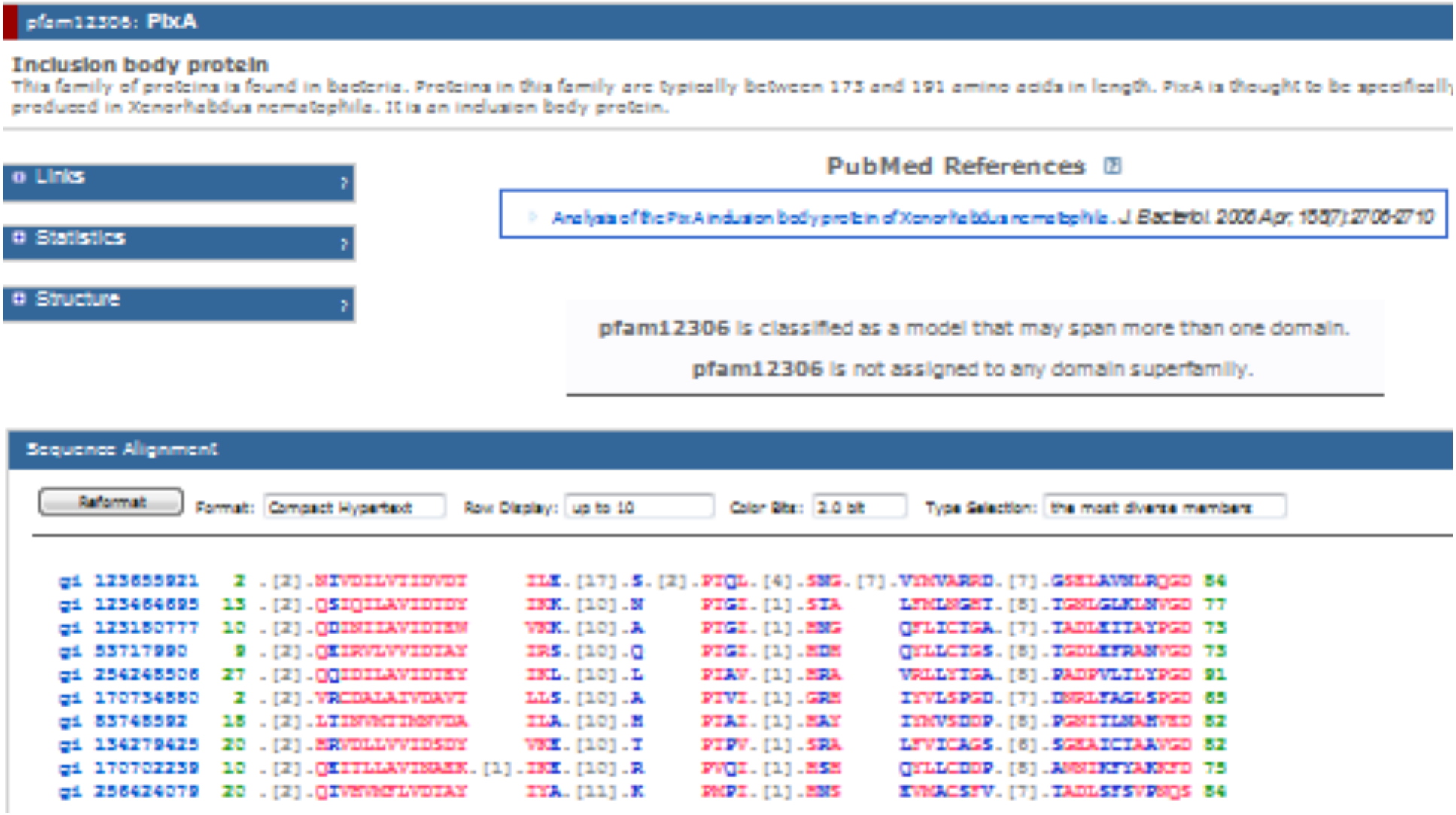
- **celý protein**



NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

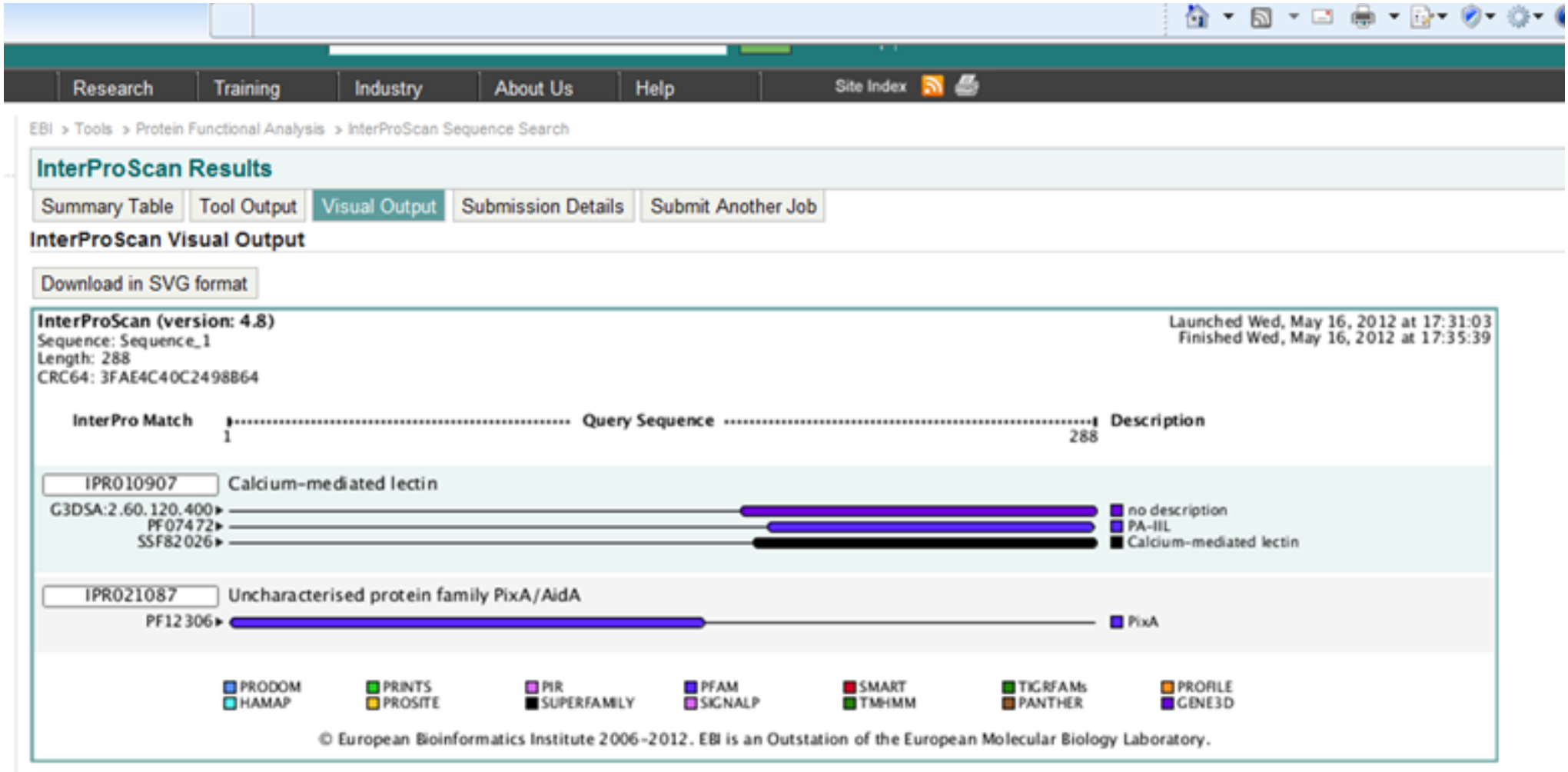
- **celý protein**



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/



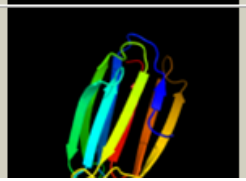
Proč potřebujeme predikci domén

- Prohledávání sekvenčních databází bez predikce domén může být neúspěšné

- Automatická predikce struktury se zaměří jen na nejlépe „definovanou“ část

- ….

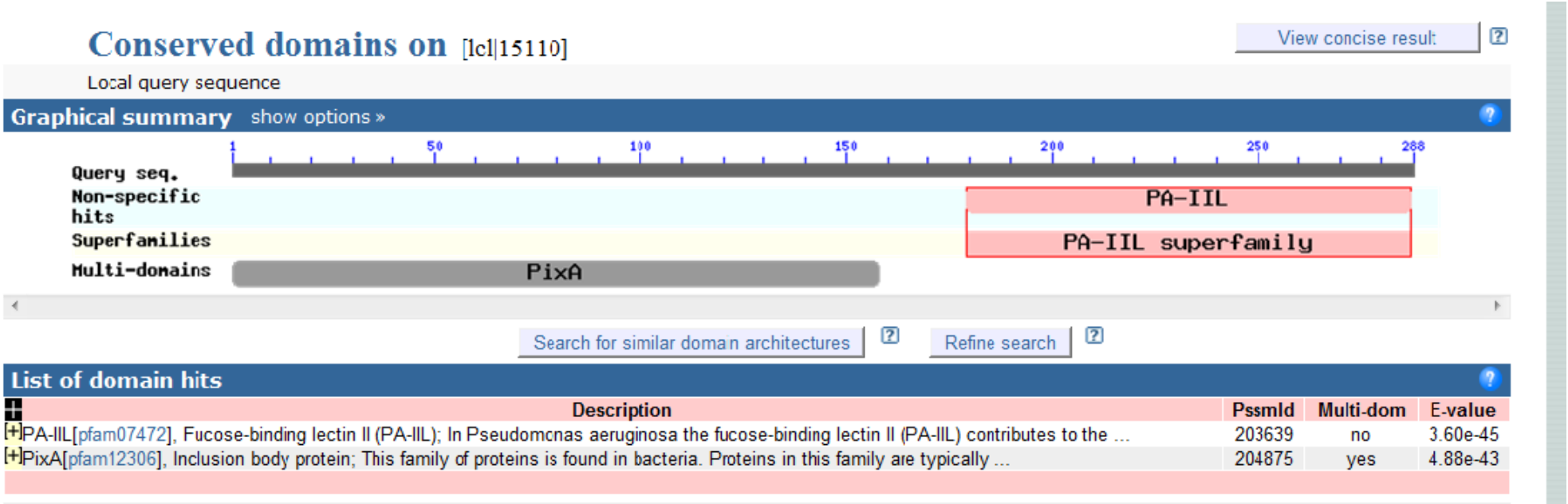
Phyre – whole protein http://www.sbg.bio.ic.ac.uk/phyre2/phyre2_output/a132b051273537c4/summary.htm

#	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

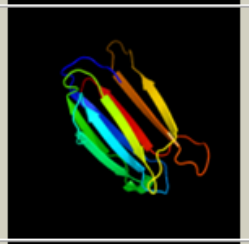
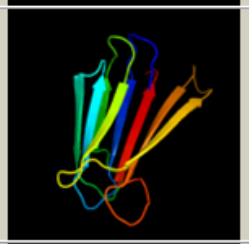
NCBI – Blast

Prohledávání databází známých aminokyselinových sekvencí

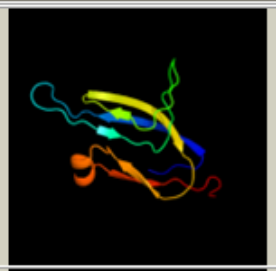
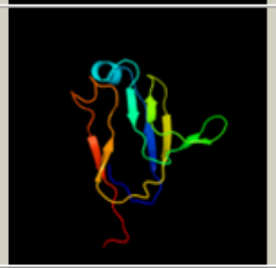
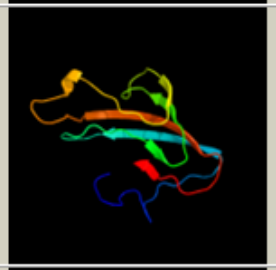
- **celý protein**



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 chair **PDBTitle:** crystal structure of human factor viii
3	d1kbva2	Alignment		68.0	13	**Fold:**Cupredoxin-like **Superfamily:**Cupredoxins **Family:**Multidomain cupredoxins

Swissprot – whole protein

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SWISS-MODEL Workspace

Modelling Tools Repository Documentation

[myWorkspace] [login]

Workunit: P000007 - Overview

1 288

Print/Save this page as

Model Summary

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

Quality information: [details]
QMEAN Z-Score: -0.71

Quaternary structure information: [details]
Template (2vnv): 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=0f449e99bc0176edfa75fba19b2d96e4&func=workspace_modelling&prjid=P000007

Swissprot - only N terminal part

Computation of this workunit has stopped.

Please see the following log report for details:

```
Started: Thu May 17 15:21:24 2012 (sms_automode_2011)
Reading user input sequence


 - Start SMR-Pipeline in automated mode on BC2-cluster at Thu May 17 13:21:24 2012


 - Start BLAST for highly similar template structure identification
 - No suitable templates found!

 - Run HHSearch to detect remotely related template structures
 - Unfortunately, we could not identify useful template structures

 - For troubleshooting, please see our article in Nature Protocols:

 - Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J. and Schwede, T. (2009).
Protein structure homology modelling using SWISS-MODEL Workspace.  Nature Protocols, 4, 1.

 - Workspace Pipeline parameter
   Cut-off parameters to model the target based on a BLAST target-template alignment
    Evalue :                                      0.0001
    Minimum Template size (aa) for ranking :      25
    Minimum Sequence identity :                   60
   Cut-off parameters to model the target based on a HHSearch target-template alignment
    Evalue :                                      0.0001
    Probability :                                 50
    MAC :                                         0.3
   Parameters for model selection
    Minimal number of uncovered target
     residues after BLAST to run HHSEARCH :       50
    Minimal number of uncovered target
     residues to model an additional template :   25
 - Finish SMR-Pipeline in automated mode on BC2-cluster at Thu May 17 13:35:44 2012
```

Prediction of protein structure from scratch

ab initio approaches

De novo modelling with Rossetta

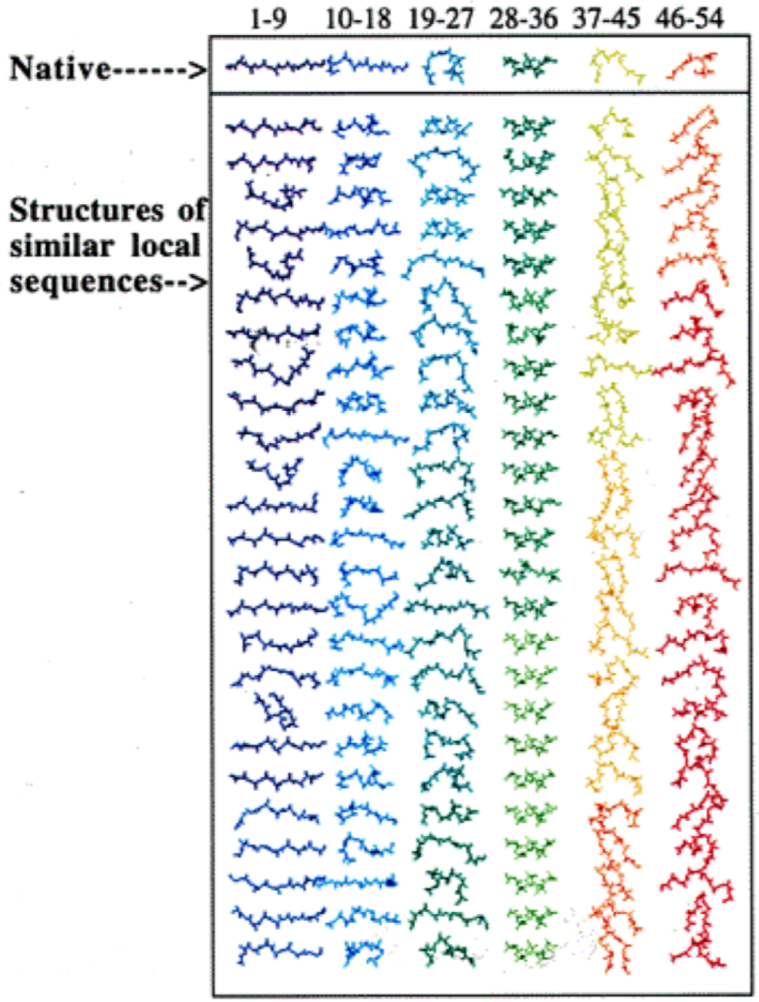
(David Baker lab, Univ. of Washington)

•• 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

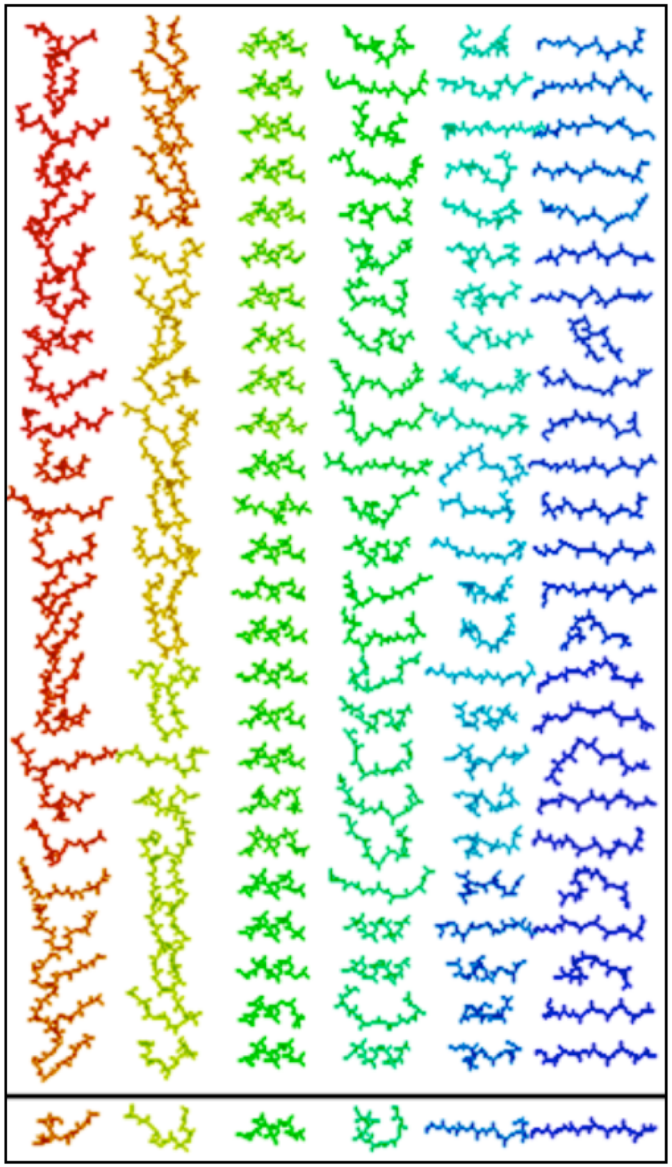
De novo modelling with Rossetta

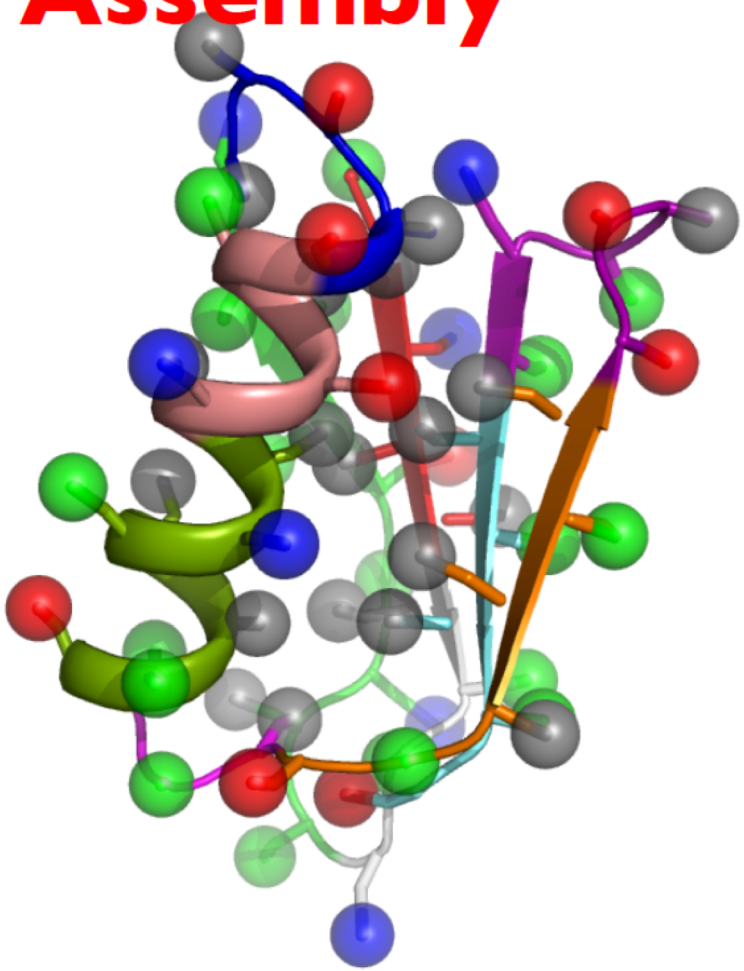
- fragments are selected from known structures
- the window-fragment matches are calculated using
  - PSI-BLAST to build a profile model of the sequence
  - the predicted secondary structure of the sequence



De novo Modeling with Rosetta

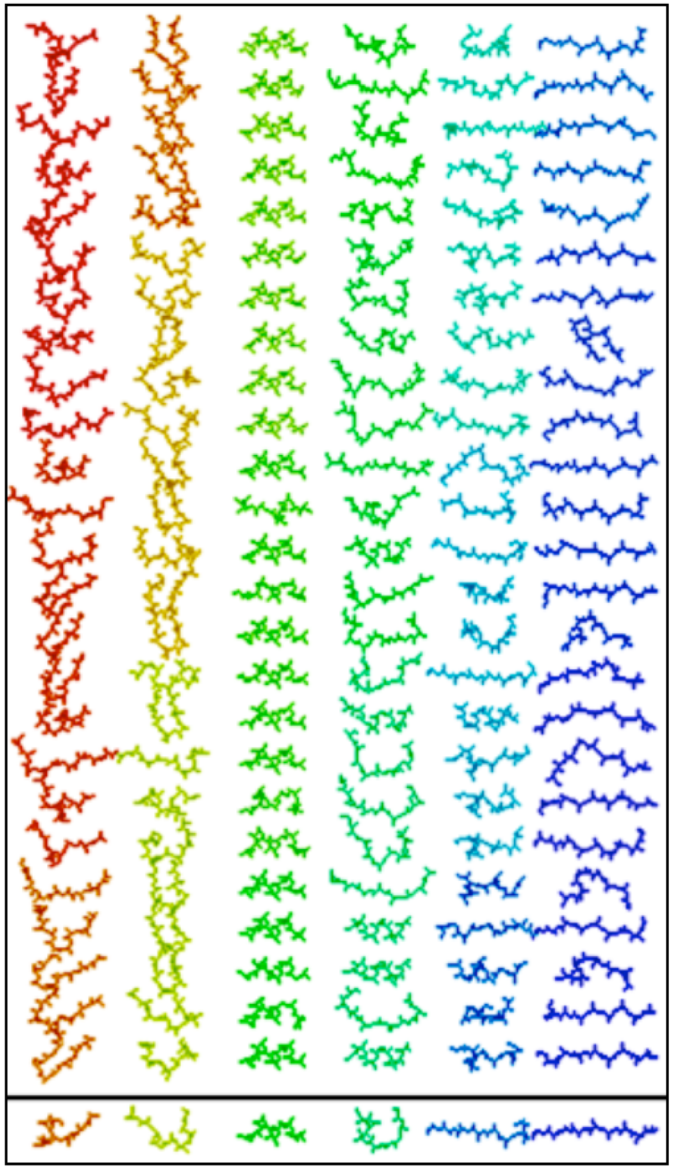
Stage I. Fragment Assembly

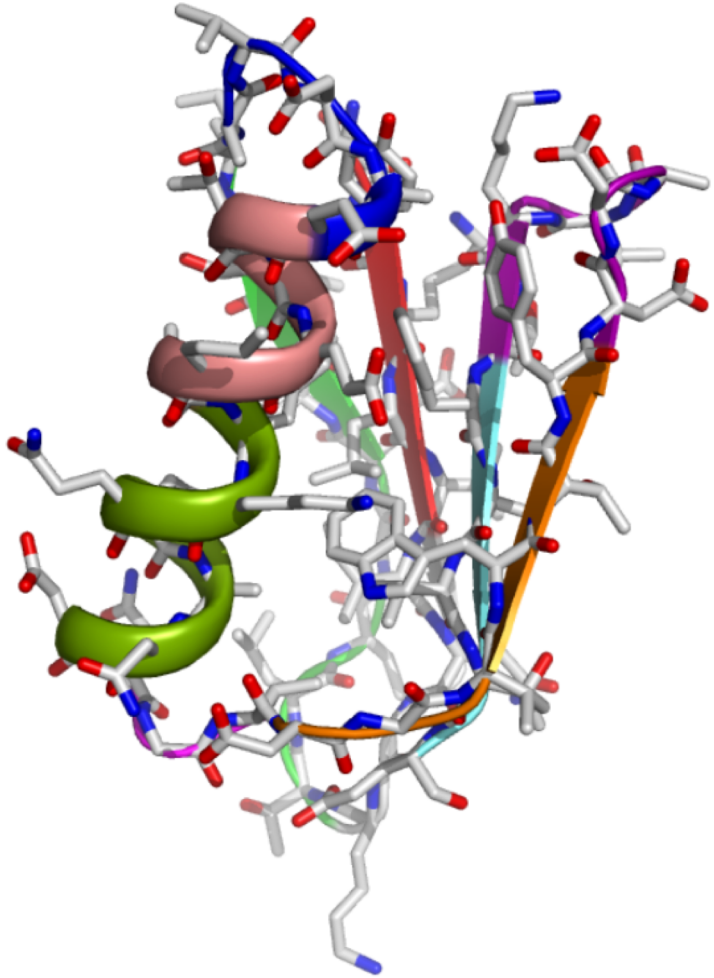




De novo Modeling with Rosetta

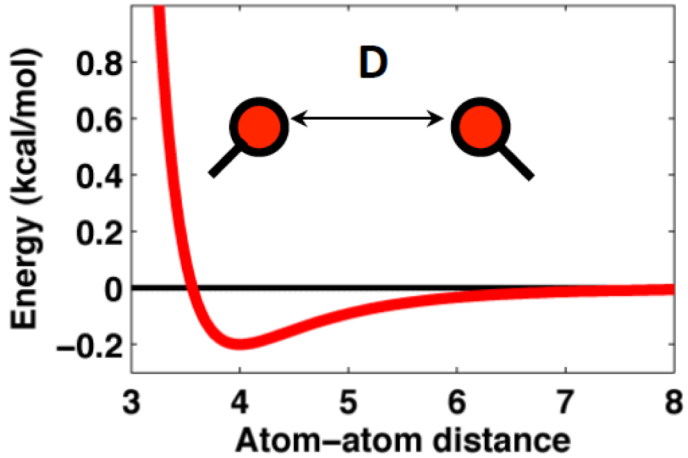
Stage II. All-atom refinement



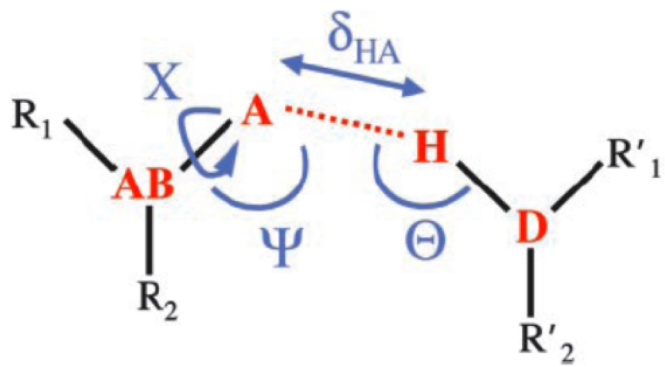


Ingredients of a high resolution potential

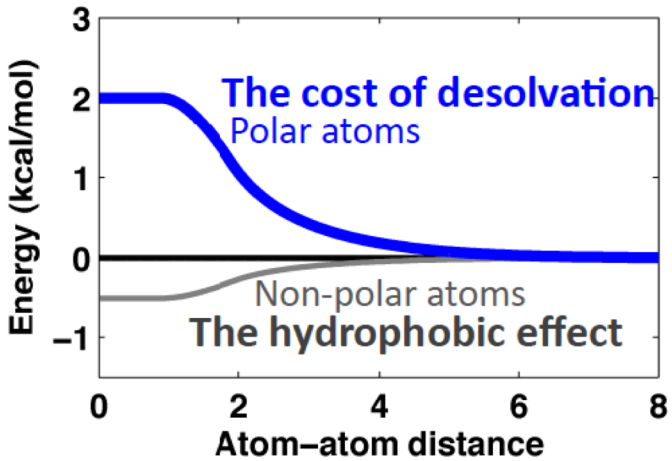
1. Van der waals packing



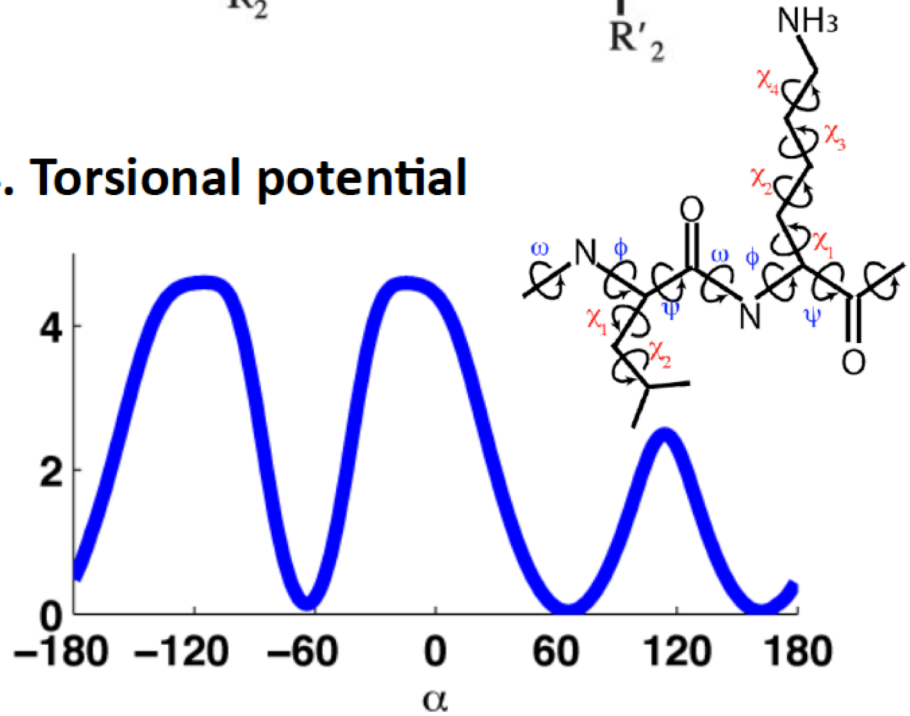
2. Hydrogen bonds



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.

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.

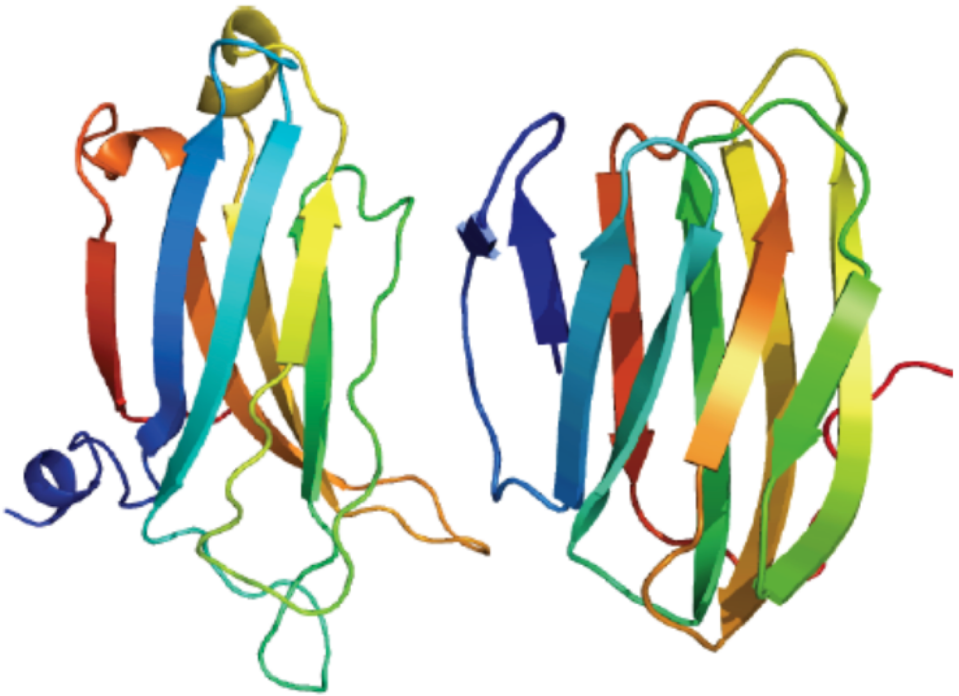
Cut and paste your sequence (in FASTA format, less than 200 AA. Example input

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

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

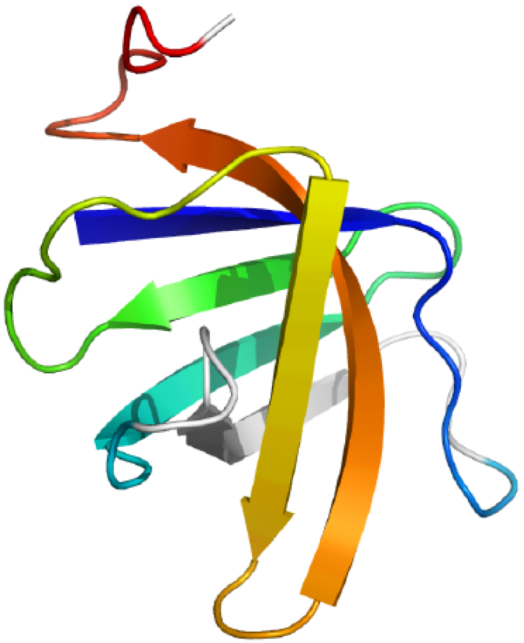
Number of groups registered	**210**
including: *expert groups*	*123*
prediction servers	*87*
Number of tertiary structure prediction targets released	**90**
(including *all-group targets*)	*(82)*
Number of hetero-multimer targets released	**13**
Number of refinement targets released	**31**
Number of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

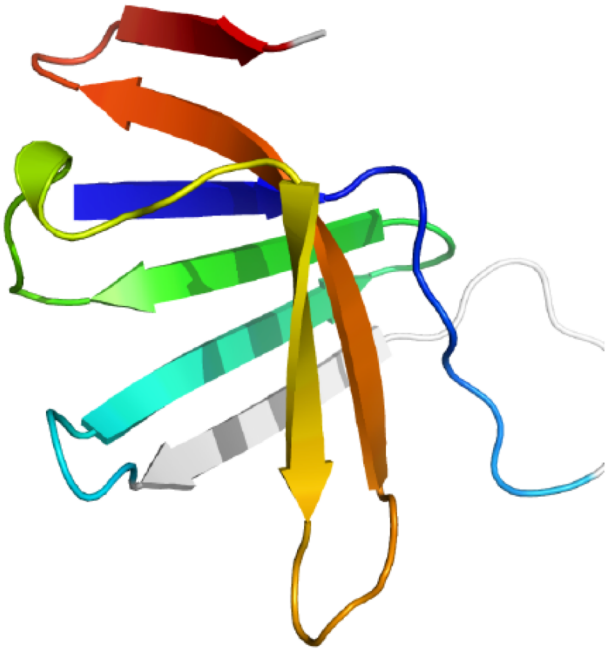
Prediction category	Number of groups/servers contributing	Number of models designated as 1	Total number of models
Tertiary structure predictions	107 / 39	7542	35982
Oligomeric predictions	40 / 9	662	2861
Data assisted predictions	24 / 5	456	2017
Residue-residue contacts	46 / 25	3914	3914
Accuracy estimation	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)



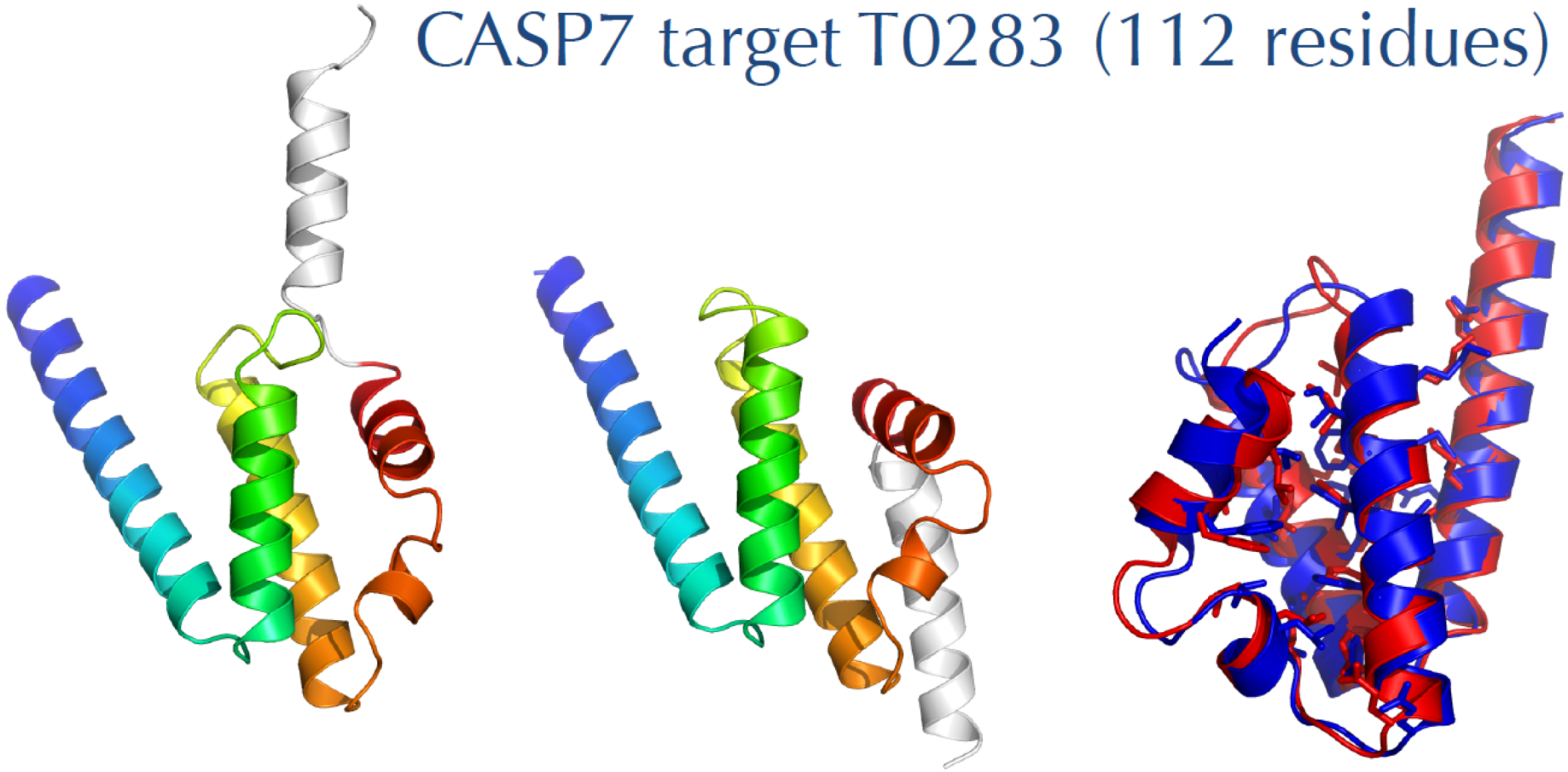


Native Model

2.0 Å over 61 residues

De novo successes: all-α

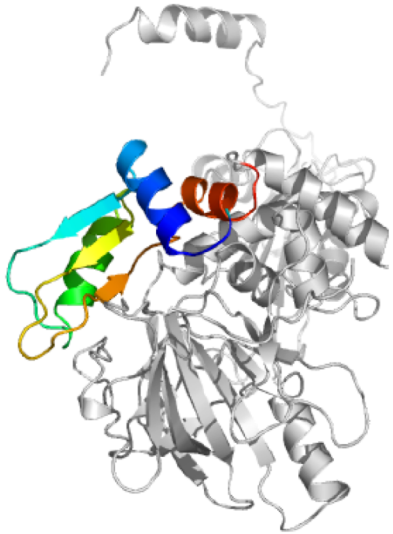
CASP7 target T0283 (112 residues)



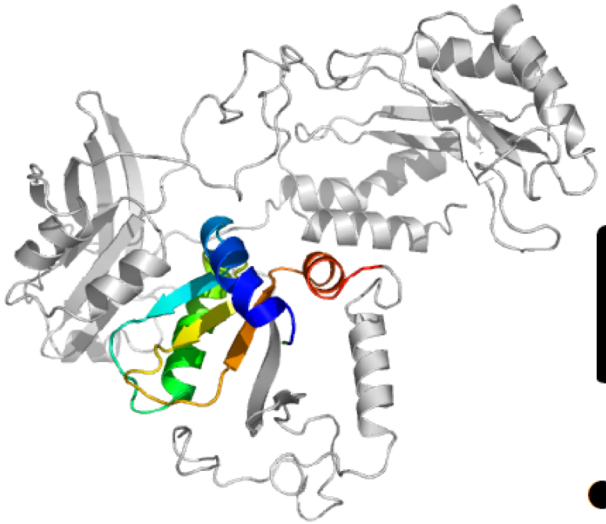
Native Model

1.4 Å over 90 residues

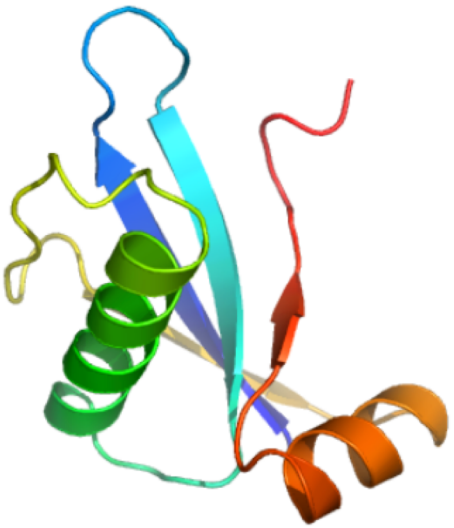
Is protein folding *solved?*

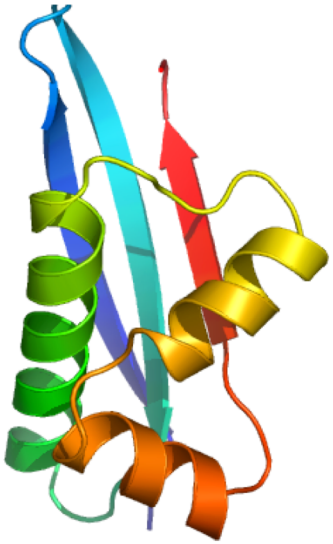


Native



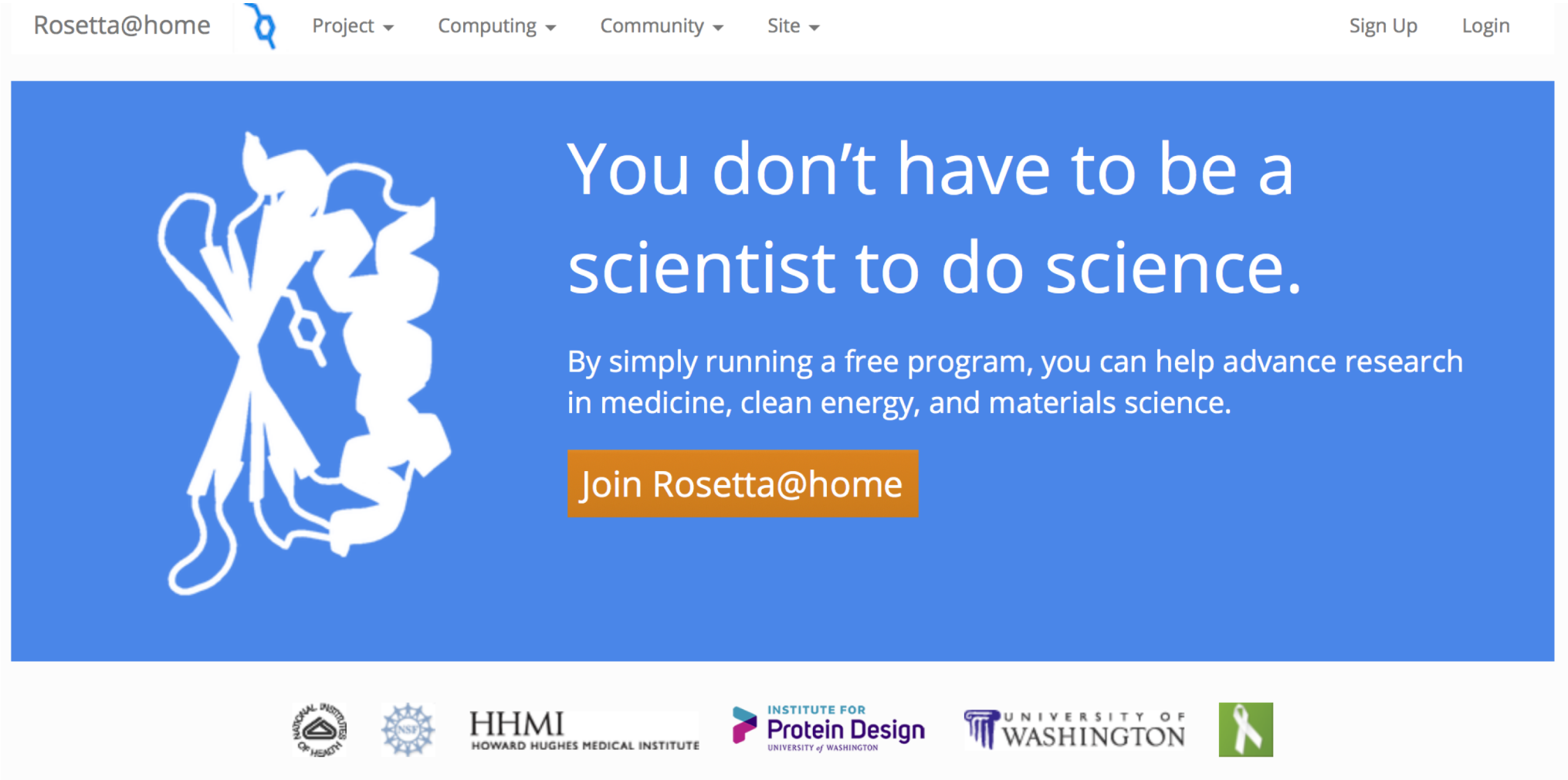
Model





NO!

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



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

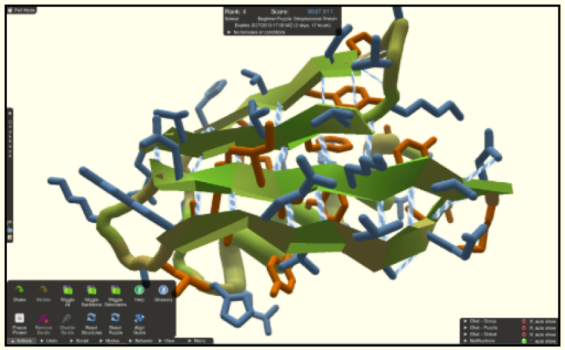
Foldit is a revolutionary crowdsourcing computer game enabling *you* to contribute to important scientific research. This page describes the science behind Foldit and how your playing can help.

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

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

Google Search — Only search fold.it

RECOMMEND FOLDIT

Send

USER LOGIN

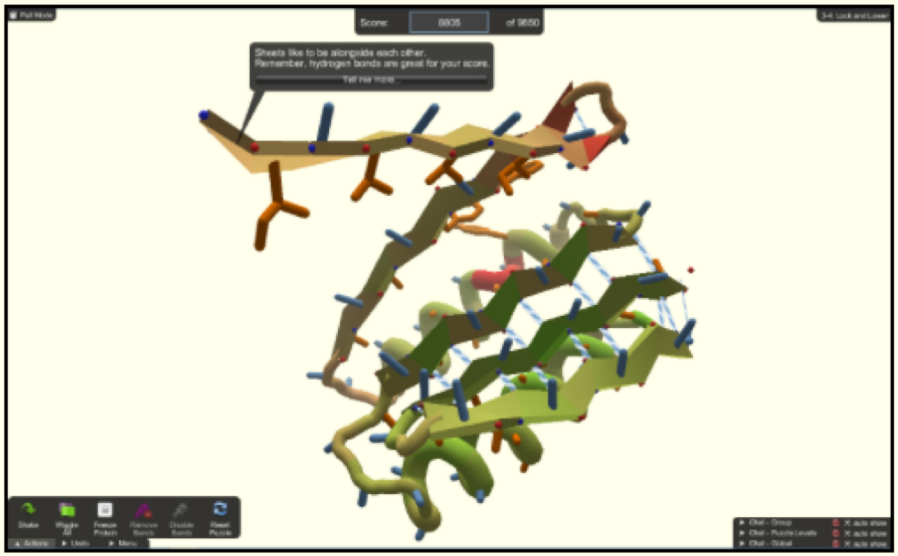
Username: *

Password: *

Log in

- Create new account
- Request new password

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!

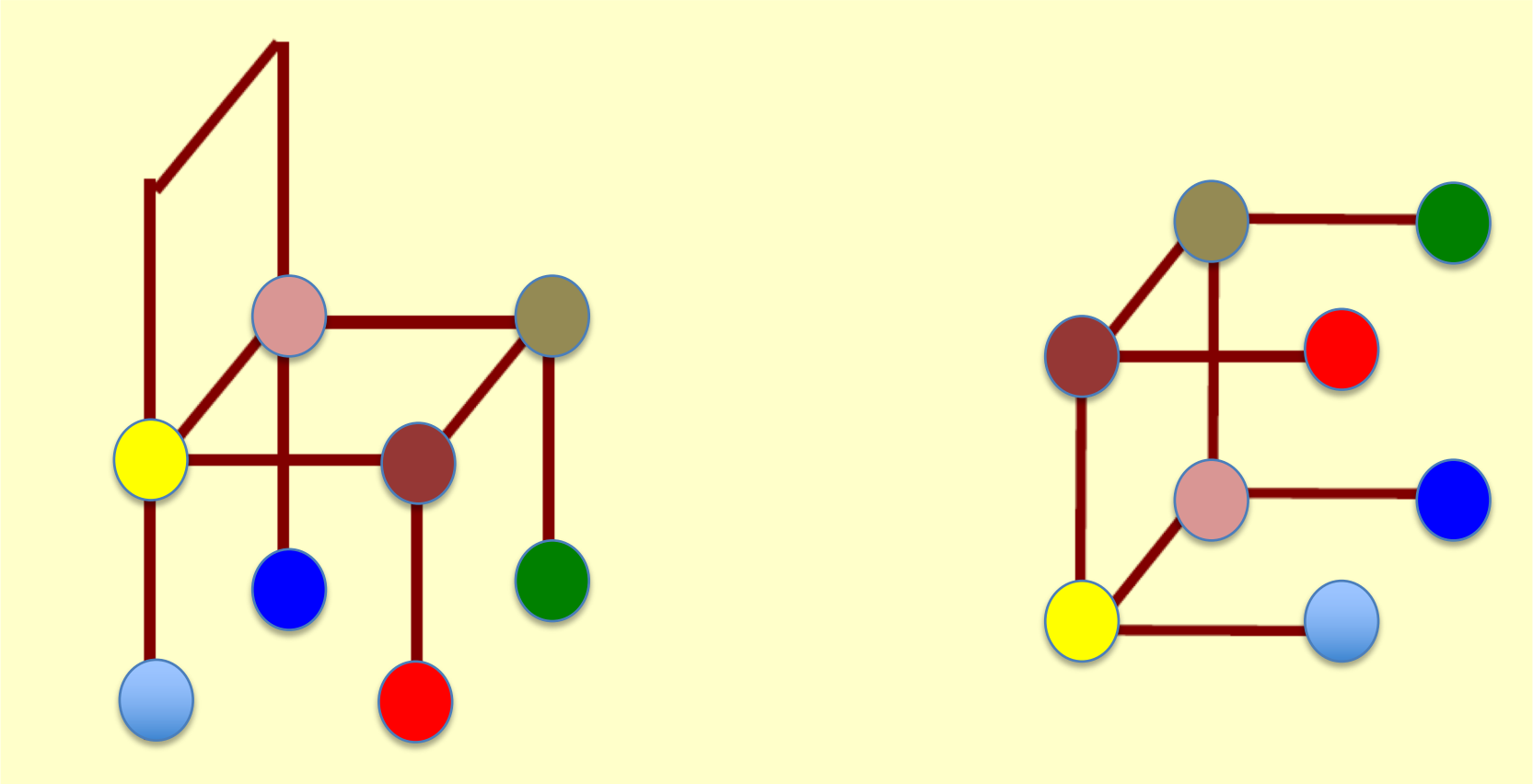
Automating Structure Classification,Fold & Function Detection

Growth of PDB demands automated techniques for classification and fold detection

Protein Structure Comparison

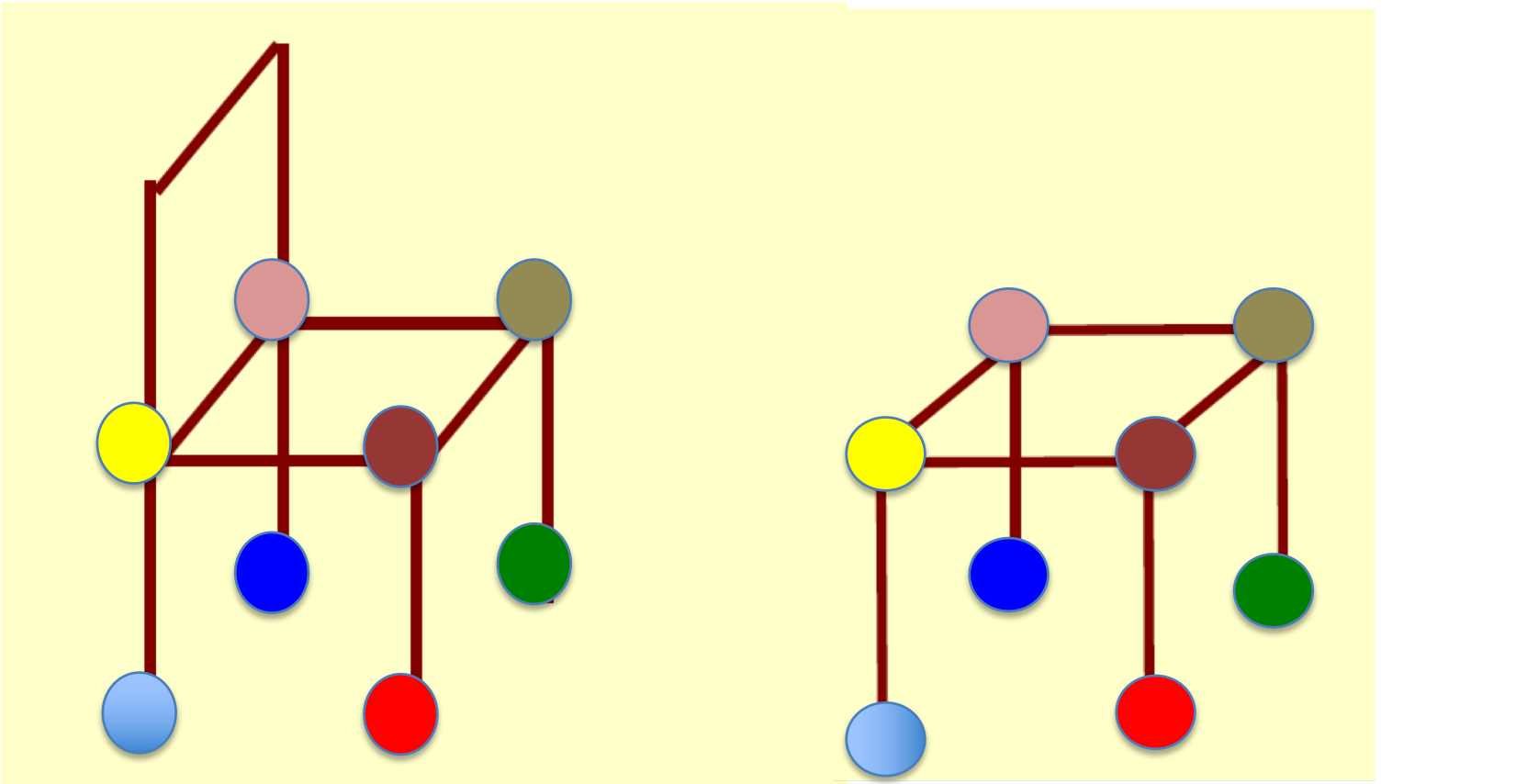
- computing structure similarity based on metrics (distances)
- identifying protein function
- understanding functional mechanism
- identifying structurally conserved regions in the protein
- finding binding sites or other functionally important regions of the protein

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

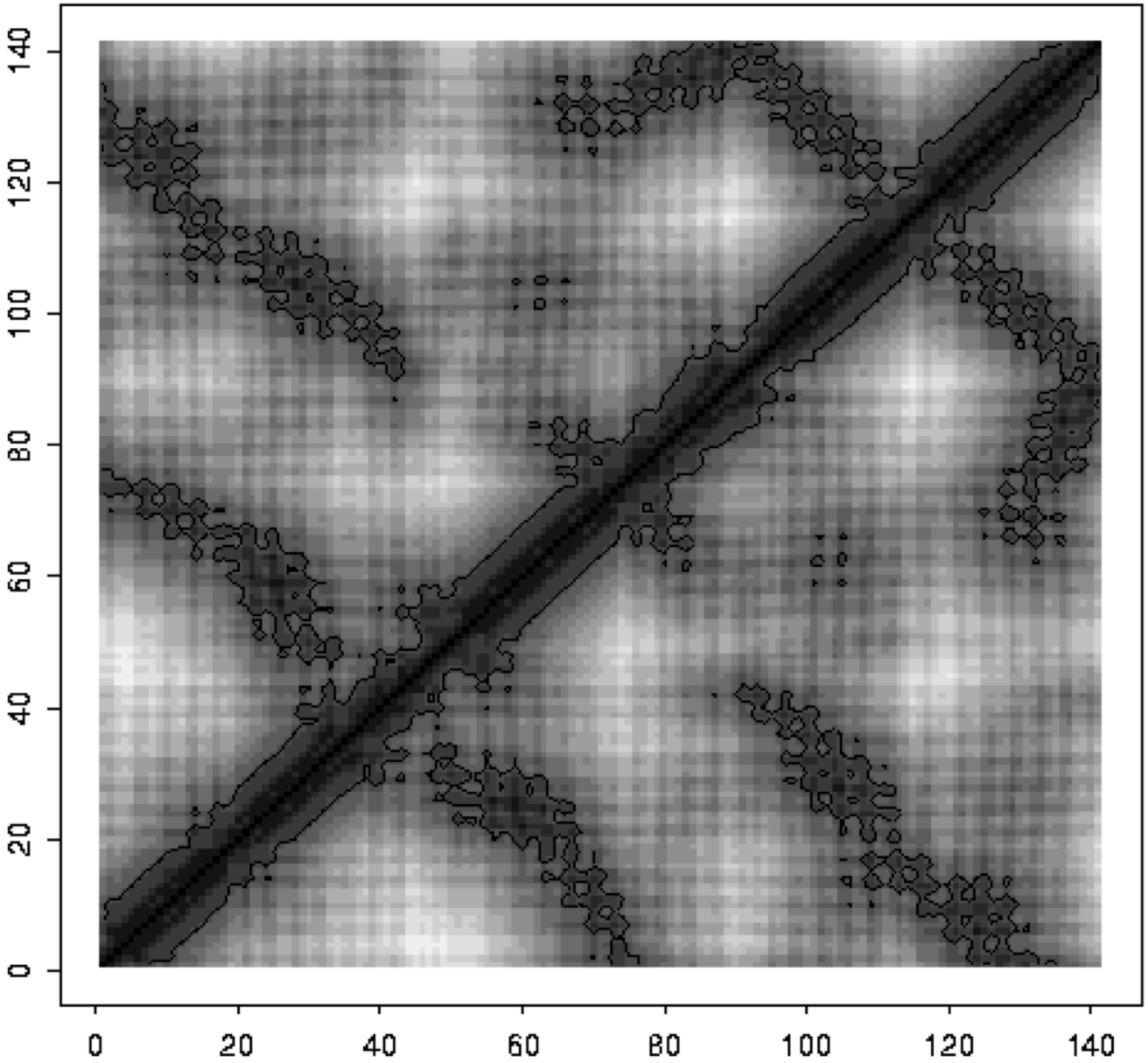
LOCK (Singh & Brutlag) LOCK2 (Ebert & Brutlag): Hierarchically uses

"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



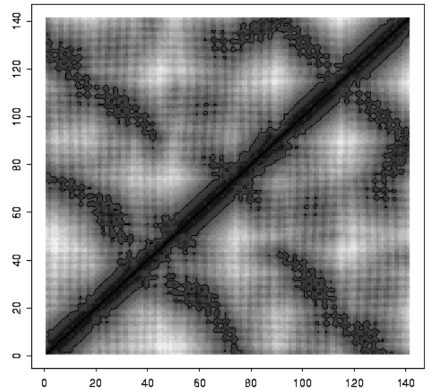
An intra-molecular distance plot for myoglobin

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



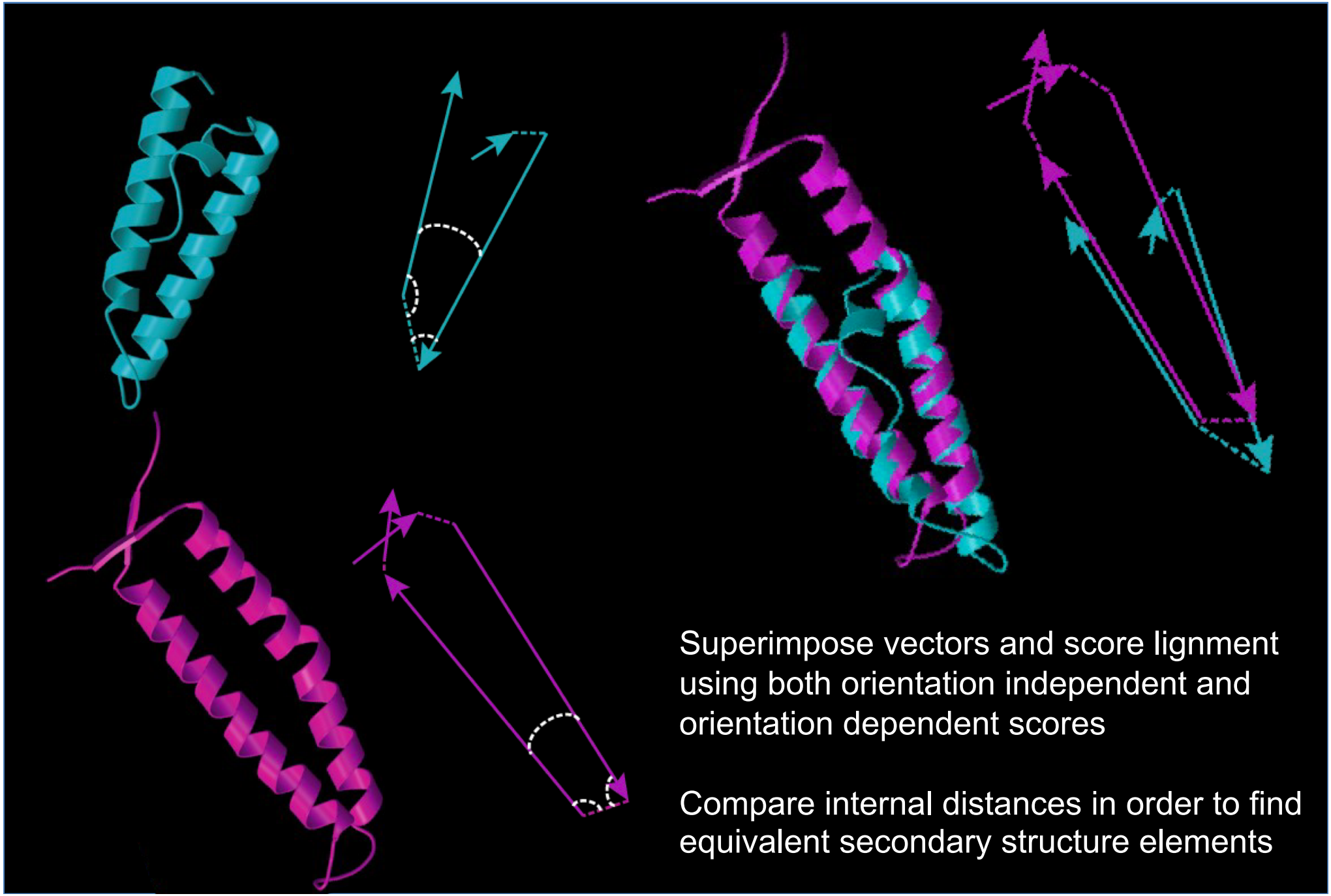
VAST – Vector Alignment Search Tool

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/