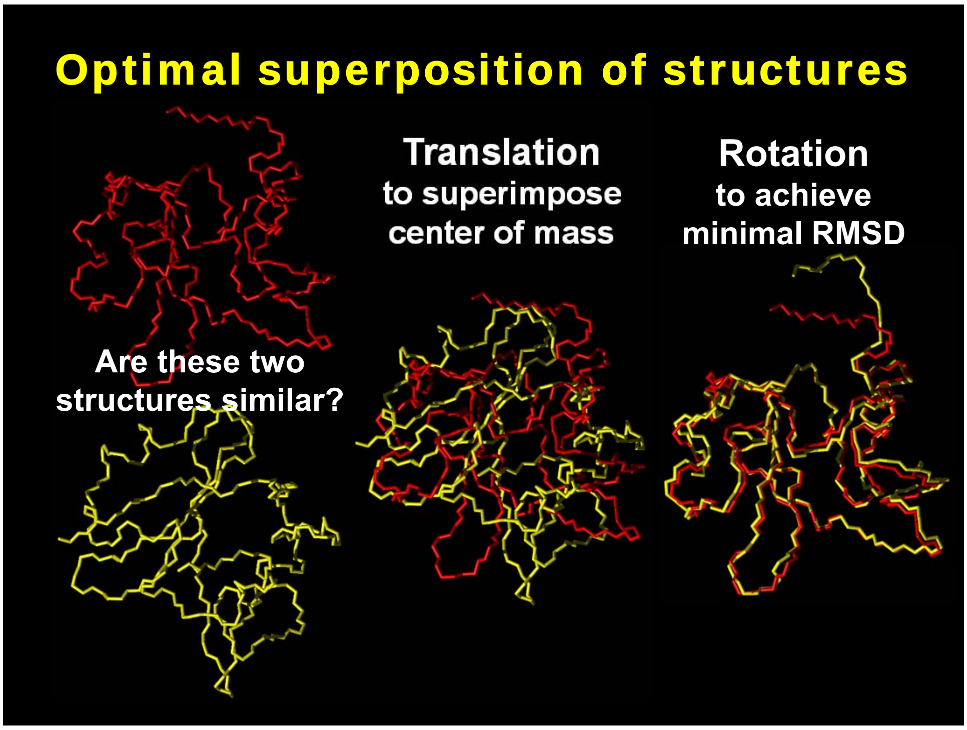
Structure Analysis Tools
Structure modelling

Wintersemester 2011/12

Peter Güntert

Structure comparison

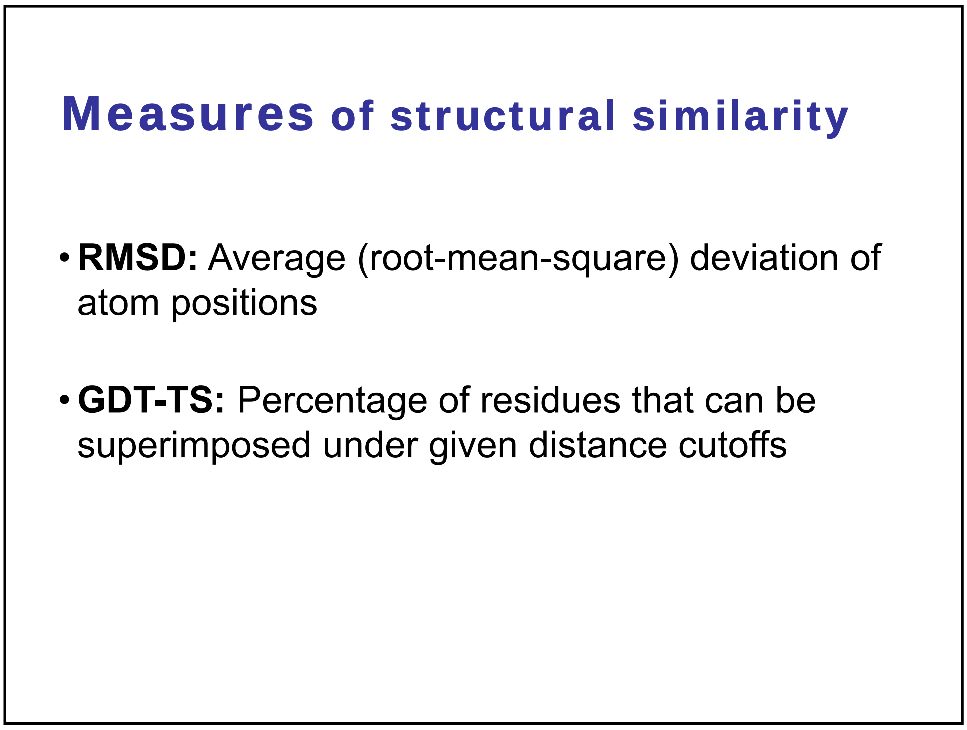


Optimal superposition of structures

Are these two structures similar?

Translation to superimpose center of mass

Rotation to achieve minimal RMSD



Measures of structural similarity

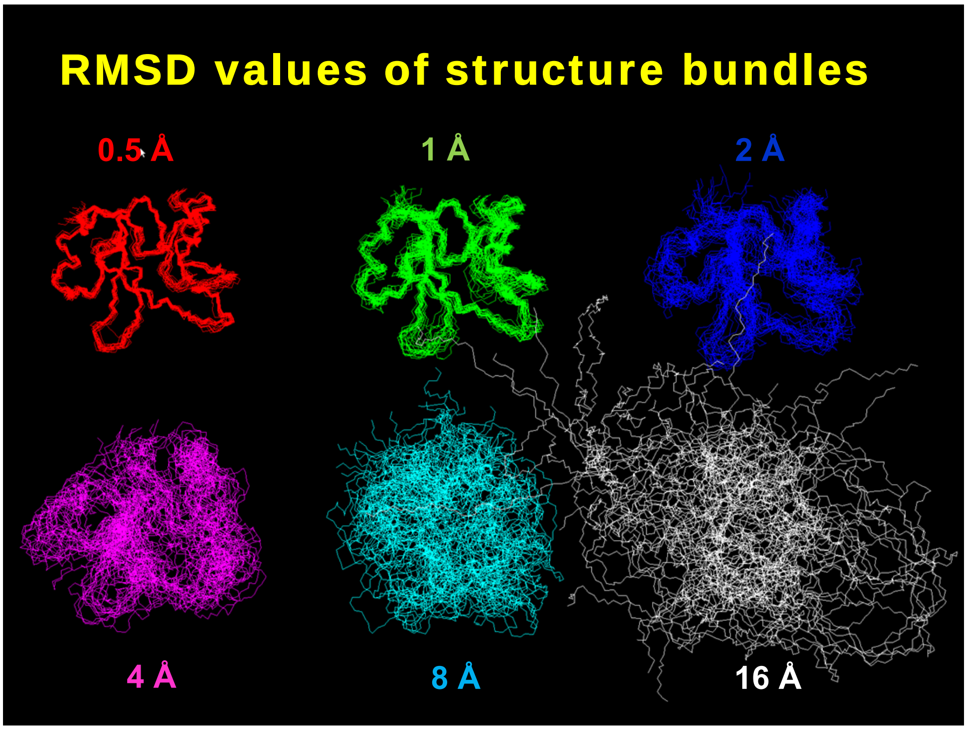
- **RMSD:** Average (root-mean-square) deviation of atom positions
- **GDT-TS:** Percentage of residues that can be superimposed under given distance cutoffs

RMSD (root-mean-square deviation)

- Zwei Strukturen mit *n* Atomen und Koordinaten $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ und $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$

$$RMSD = \min_{R,\vec{t}} \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\vec{x}_i - R\vec{y}_i - \vec{t}\right|^2}$$

- Minimum über alle Rotationen *R* und Translationen ***t*** → optimale Überlagerung



RMSD values of structure bundles

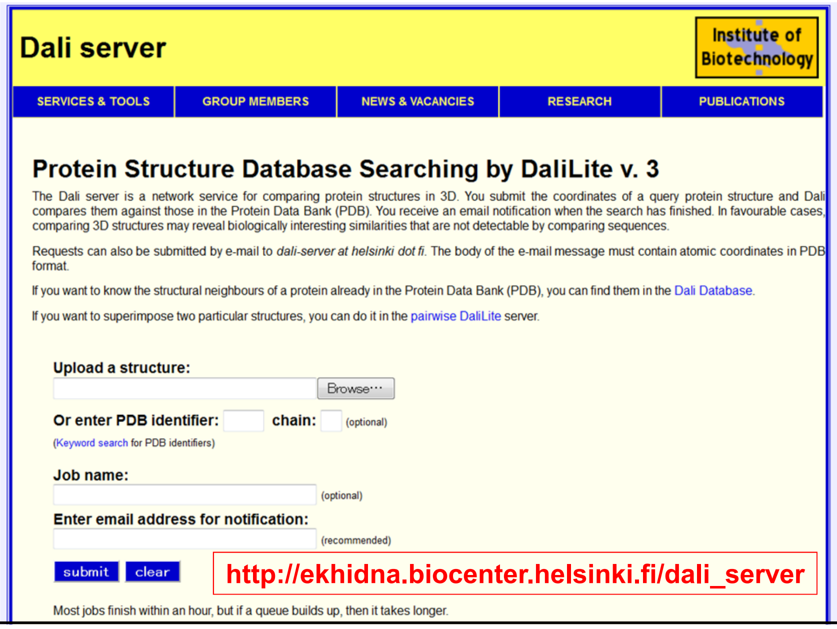
GDT_TS

- The GDT (“global distance test”) algorithm searches for the largest (not necessarily continuous) set of residues that deviate by no more than a specified distance cutoff.
- Results are reported as the percentage of residues under a given distance cutoff.
- A popular measure is the “GDT total score”,

$$GDT\_TS = (P_1 + P_2 + P_4 + P_8)/4,$$

where P_d is the fraction of residues that can be superimposed under a distance cutoff of d Å, which reduces the dependence on the choice of the cutoff by averaging over four different distance cutoff values.

DALI: structure similarity search



Protein Structure Database Searching by DaliLite v. 3

The Dali server is a network service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the Protein Data Bank (PDB). You receive an email notification when the search has finished. In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences.

Requests can also be submitted by e-mail to *dali-server at helsinki dot fi*. The body of the e-mail message must contain atomic coordinates in PDB format.

If you want to know the structural neighbours of a protein already in the Protein Data Bank (PDB), you can find them in the Dali Database.

If you want to superimpose two particular structures, you can do it in the pairwise DaliLite server.

Upload a structure:

Or enter PDB identifier: **chain:** (optional)

(Keyword search for PDB identifiers)

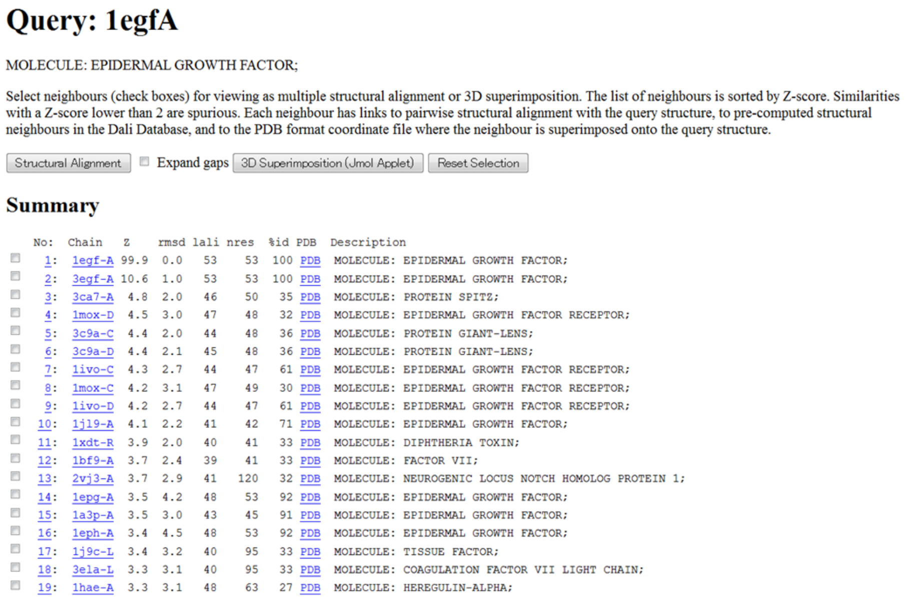
Job name: (optional)

Enter email address for notification: (recommended)

http://ekhidna.biocenter.helsinki.fi/dali_server

Most jobs finish within an hour, but if a queue builds up, then it takes longer.

DALI: Example result



Query: 1egfA

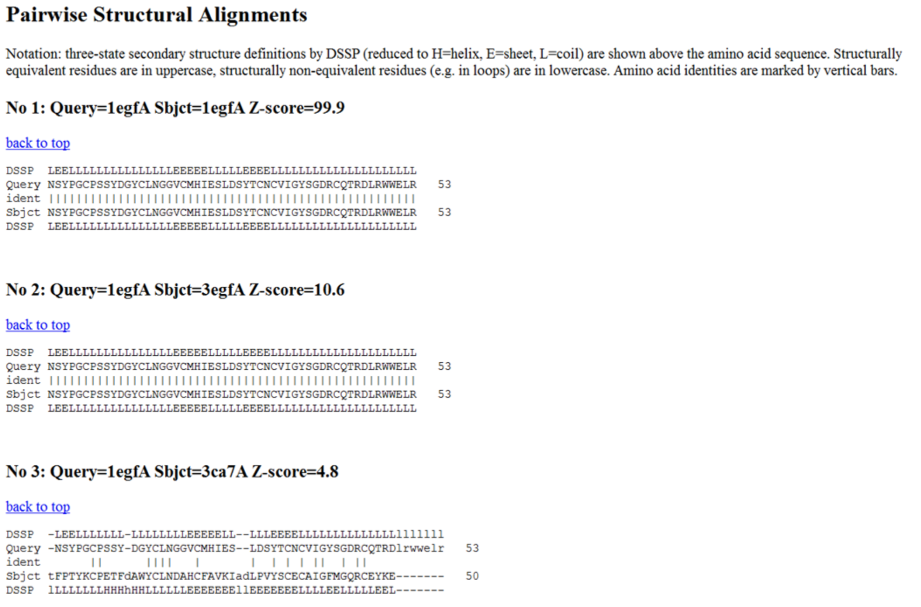
MOLECULE: EPIDERMAL GROWTH FACTOR;

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, to pre-computed structural neighbours in the Dali Database, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

Summary

	No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
☐	1:	1egf-A	99.9	0.0	53	53	100	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
☐	2:	3egf-A	10.6	1.0	53	53	100	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
☐	3:	3ca7-A	4.8	2.0	46	50	35	PDB	MOLECULE: PROTEIN SPITZ;
☐	4:	1mox-D	4.5	3.0	47	48	32	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
☐	5:	3c9a-C	4.4	2.0	44	48	36	PDB	MOLECULE: PROTEIN GIANT-LENS;
☐	6:	3c9a-D	4.4	2.1	45	48	36	PDB	MOLECULE: PROTEIN GIANT-LENS;
☐	7:	1ivo-C	4.3	2.7	44	47	61	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
☐	8:	1mox-C	4.2	3.1	47	49	30	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
☐	9:	1ivo-D	4.2	2.7	44	47	61	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
☐	10:	1jl9-A	4.1	2.2	41	42	71	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
☐	11:	1xdt-R	3.9	2.0	40	41	33	PDB	MOLECULE: DIPHTHERIA TOXIN;
☐	12:	1bf9-A	3.7	2.4	39	41	33	PDB	MOLECULE: FACTOR VII;
☐	13:	2vj3-A	3.7	2.9	41	120	32	PDB	MOLECULE: NEUROGENIC LOCUS NOTCH HOMOLOG PROTEIN 1;
☐	14:	1epg-A	3.5	4.2	48	53	92	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
☐	15:	1a3p-A	3.5	3.0	43	45	91	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
☐	16:	1eph-A	3.4	4.5	48	53	92	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
☐	17:	1j9c-L	3.4	3.2	40	95	33	PDB	MOLECULE: TISSUE FACTOR;
☐	18:	3ela-L	3.3	3.1	40	95	33	PDB	MOLECULE: COAGULATION FACTOR VII LIGHT CHAIN;
☐	19:	1hae-A	3.3	3.1	48	63	27	PDB	MOLECULE: HEREGULIN-ALPHA;

DALI: Example result



Pairwise Structural Alignments

Notation: three-state secondary structure definitions by DSSP (reduced to H=helix, E=sheet, L=coil) are shown above the amino acid sequence. Structurally equivalent residues are in uppercase, structurally non-equivalent residues (e.g. in loops) are in lowercase. Amino acid identities are marked by vertical bars.

No 1: Query=1egfA Sbjct=1egfA Z-score=99.9

back to top

```
DSSP  LEELLLLLLLLLLLLLLLEEEEELLLLEEEEELLLLLLLLLLLLLLLLLLLLL
Query NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR   53
ident |||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR   53
DSSP  LEELLLLLLLLLLLLLLLEEEEELLLLEEEEELLLLLLLLLLLLLLLLLLLLL
```

No 2: Query=1egfA Sbjct=3egfA Z-score=10.6

back to top

```
DSSP  LEELLLLLLLLLLLLLLLEEEEELLLLEEEEELLLLLLLLLLLLLLLLLLLLL
Query NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR   53
ident |||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR   53
DSSP  LEELLLLLLLLLLLLLLLEEEEELLLLEEEEELLLLLLLLLLLLLLLLLLLLL
```

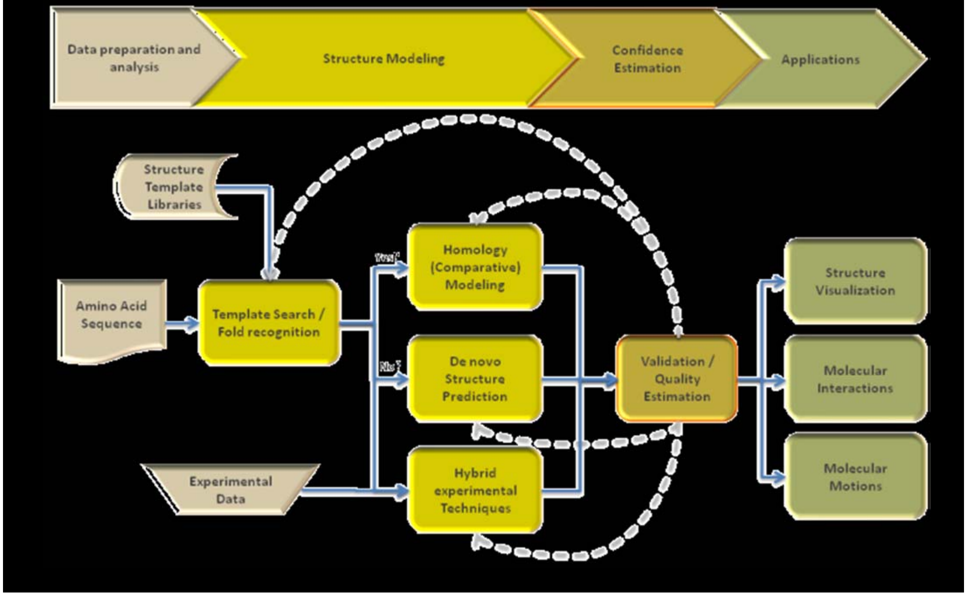
No 3: Query=1egfA Sbjct=3ca7A Z-score=4.8

back to top

```
DSSP  -LEELLLLLLL-LLLLLLLEEEEEELL--LLLEEEELLLLLLLLLLLLLLlllllll
Query -NSYPGCPSSY-DGYCLNGGVCMHIES--LDSYTCNCVIGYSGDRCQTRDlrwwelr   53
ident       ||      ||||   |       |  | | | ||  | ||
Sbjct tFPTYKCPETFdAWYCLNDAHCFAVKIadLPVYSCECAIGFMGQRCEYKE-------   50
DSSP  lLLLLLLLHHHhHHLLLLLLEEEEEEEllEEEEEEELLLLEELLLLLEEL-------
```

Structure modelling

Structure modelling



Data preparation and analysis
Structure Modeling
Confidence Estimation
Applications
Structure Template Libraries
Amino Acid Sequence
Template Search / Fold recognition
Yes?
No?
Homology (Comparative) Modeling
De novo Structure Prediction
Hybrid experimental Techniques
Experimental Data
Validation / Quality Estimation
Structure Visualization
Molecular Interactions
Molecular Motions

Sequence identity → Structural similarity

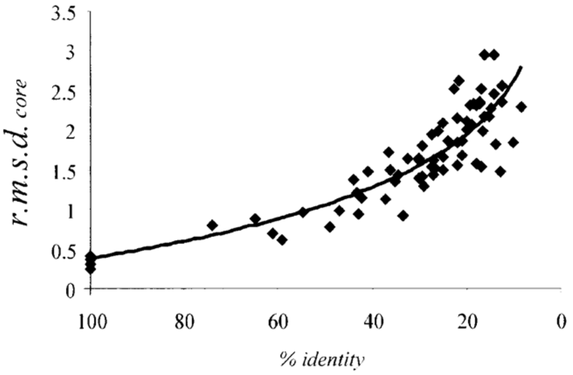


Figure 1.25 Relationships between sequence identity and structural similarity.

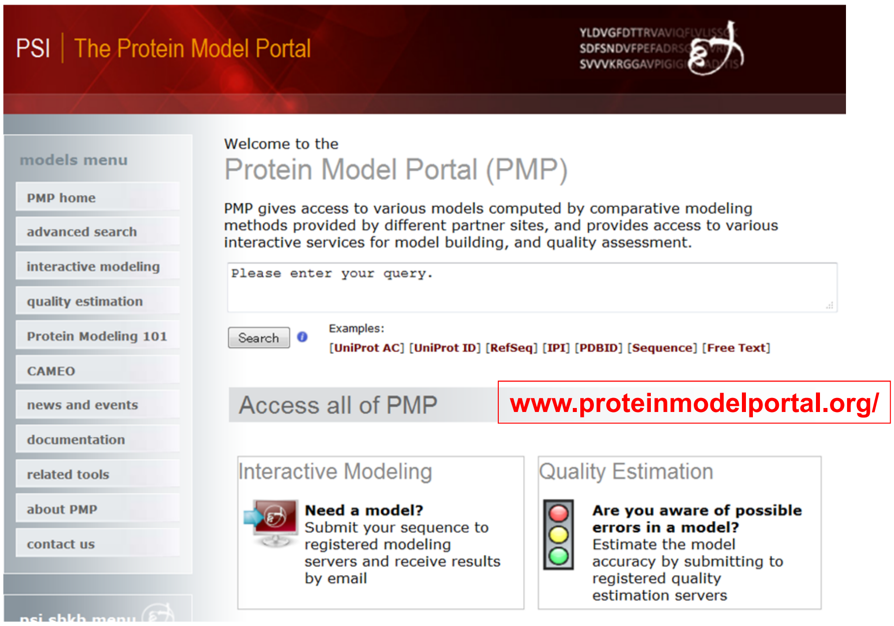
BUT:
Structural similarity ↛ Sequence identity

Methods for protein structure prediction

Methods are distinguished according to the relationship between the target protein(s) and proteins of known structure:

- **Comparative modelling**: A clear evolutionary relationship between the target and a protein of known structure can be easily detected from the sequence.
- **Fold recognition:** The structure of the target turns out to be related to that of a protein of known structure although the relationship is difficult, or impossible, to detect from the sequences.
- **New fold prediction:** Neither the sequence nor the structure of the target protein are similar to that of a known protein.

PSI Protein Model portal (PMP)



Welcome to the

Protein Model Portal (PMP)

PMP gives access to various models computed by comparative modeling methods provided by different partner sites, and provides access to various interactive services for model building, and quality assessment.

Please enter your query.

Examples: [UniProt AC] [UniProt ID] [RefSeq] [IPI] [PDBID] [Sequence] [Free Text]

Access all of PMP

www.proteinmodelportal.org/

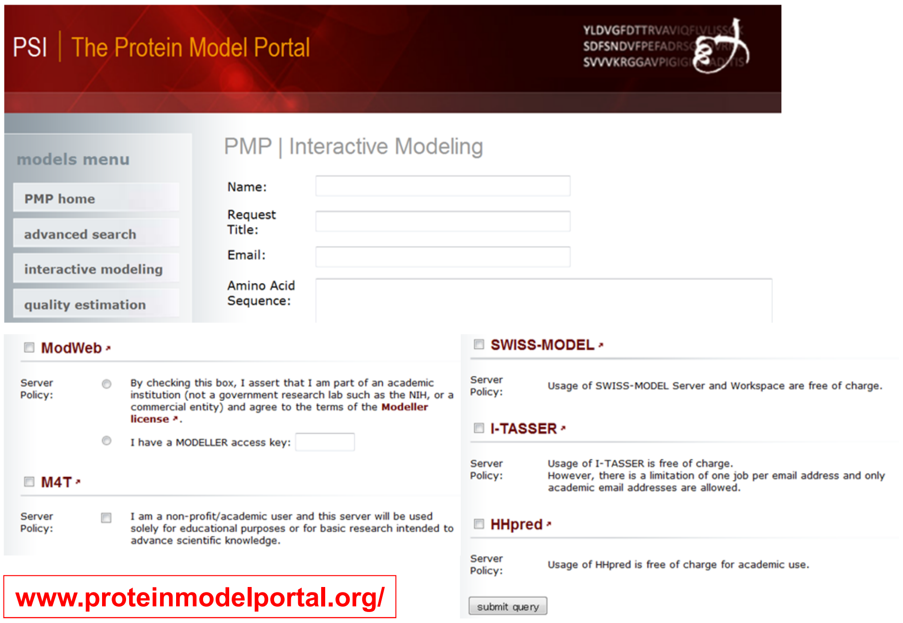
Interactive Modeling

Need a model? Submit your sequence to registered modeling servers and receive results by email

Quality Estimation

Are you aware of possible errors in a model? Estimate the model accuracy by submitting to registered quality estimation servers

PSI Protein Model portal (PMP)



PMP | Interactive Modeling

Name:

Request Title:

Email:

Amino Acid Sequence:

□ ModWeb

Server Policy: ○ By checking this box, I assert that I am part of an academic institution (not a government research lab such as the NIH, or a commercial entity) and agree to the terms of the **Modeller license**.

○ I have a MODELLER access key:

□ M4T

Server Policy: □ I am a non-profit/academic user and this server will be used solely for educational purposes or for basic research intended to advance scientific knowledge.

□ SWISS-MODEL

Server Policy: Usage of SWISS-MODEL Server and Workspace are free of charge.

□ I-TASSER

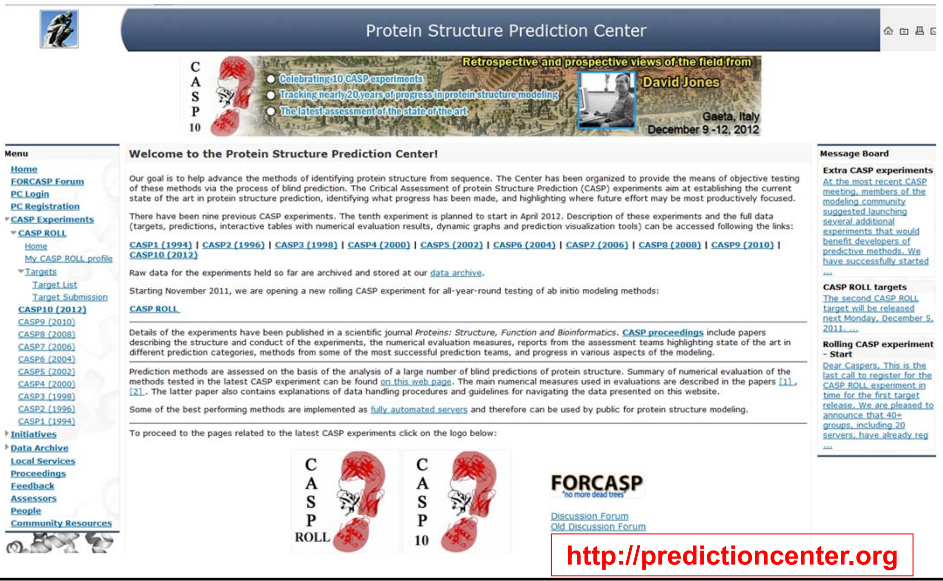
Server Policy: Usage of I-TASSER is free of charge.
However, there is a limitation of one job per email address and only academic email addresses are allowed.

□ HHpred

Server Policy: Usage of HHpred is free of charge for academic use.

www.proteinmodelportal.org/

CASP: Critical Assessment of Structure Prediction



Protein Structure Prediction Center

CASP 10

○ Celebrating 10 CASP experiments
○ Tracking nearly 20 years of progress in protein structure modeling
○ The latest assessment of the state of the art

Retrospective and prospective views of the field from David Jones

Gaeta, Italy
December 9 -12, 2012

Menu

- Home
- FORCASP Forum
- PC Login
- PC Registration
- CASP Experiments
  - CASP ROLL
    - Home
    - My CASP ROLL profile
    - Targets
      - Target List
      - Target Submission
  - CASP10 (2012)
  - CASP9 (2010)
  - CASP8 (2008)
  - CASP7 (2006)
  - CASP6 (2004)
  - CASP5 (2002)
  - CASP4 (2000)
  - CASP3 (1998)
  - CASP2 (1996)
  - CASP1 (1994)
- Initiatives
- Data Archive
- Local Services
- Proceedings
- Feedback
- Assessors
- People
- Community Resources

Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

There have been nine previous CASP experiments. The tenth experiment is planned to start in April 2012. Description of these experiments and the full data (targets, predictions, interactive tables with numerical evaluation results, dynamic graphs and prediction visualization tools) can be accessed following the links:

CASP1 (1994) | CASP2 (1996) | CASP3 (1998) | CASP4 (2000) | CASP5 (2002) | CASP6 (2004) | CASP7 (2006) | CASP8 (2008) | CASP9 (2010) | CASP10 (2012)

Raw data for the experiments held so far are archived and stored at our data archive.

Starting November 2011, we are opening a new rolling CASP experiment for all-year-round testing of ab initio modeling methods:

CASP ROLL

Details of the experiments have been published in a scientific journal *Proteins: Structure, Function and Bioinformatics*. **CASP proceedings** include papers describing the structure and conduct of the experiments, the numerical evaluation measures, reports from the assessment teams highlighting state of the art in different prediction categories, methods from some of the most successful prediction teams, and progress in various aspects of the modeling.

Prediction methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure. Summary of numerical evaluation of the methods tested in the latest CASP experiment can be found on this web page. The main numerical measures used in evaluations are described in the papers [1], [2]. The latter paper also contains explanations of data handling procedures and guidelines for navigating the data presented on this website.

Some of the best performing methods are implemented as fully automated servers and therefore can be used by public for protein structure modeling.

To proceed to the pages related to the latest CASP experiments click on the logo below:

CASP ROLL

CASP 10

FORCASP
"no more dead trees"

Discussion Forum
Old Discussion Forum

Message Board

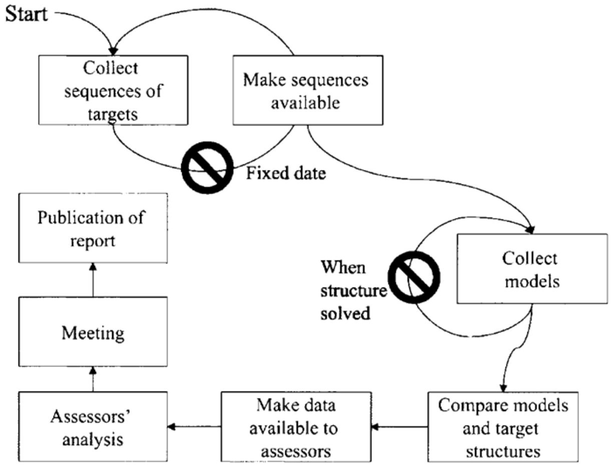
Extra CASP experiments
At the most recent CASP meeting, members of the modeling community suggested launching several additional experiments that would benefit developers of predictive methods. We have successfully started ...

CASP ROLL targets
The second CASP ROLL target will be released next Monday, December 5, 2011. ...

Rolling CASP experiment - Start
Dear Caspers, This is the last call to register for the CASP ROLL experiment in time for the first target release. We are pleased to announce that 40+ groups, including 20 servers, have already reg ...

http://predictioncenter.org

CASP: Critical Assessment of Structure Prediction



Start
Collect sequences of targets
Make sequences available
Fixed date
Publication of report
When structure solved
Collect models
Meeting
Assessors' analysis
Make data available to assessors
Compare models and target structures

Figure 2.9 The CASP experiment runs every two years. In the spring, approximately, targets are collected from experimenters working on the resolution of their structure. The sequences are made available to predictors who can submit predictions until the structure is solved. Numerical comparison of models and targets is performed by a group of scientists led by John Moult and Krzystof Fidelis. The data are then passed to thee assessors, chosen by the community on the basis of their expertise, who analyze the data and try to derive general conclusions about the state of the art in the prediction field. In approximately December of the same year, predictors, assessors, and organizers convene in a meeting to discuss the results and, later, publish the final reports in the scientific journal *Proteins: Structure, Function and Bioinformatics.*

CASP

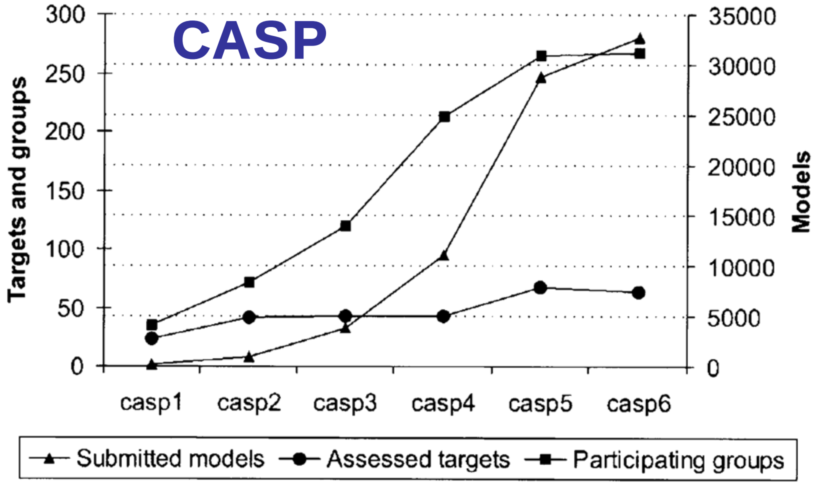


Figure 2.10 The plot shows the numbers of targets, participating groups, and models submitted to each of the editions of CASP from 1994 (CASP1) to 2004 (CASP6). All the thousands of models are publicly available on the CASP web site.

Scheme of protein structure predicition

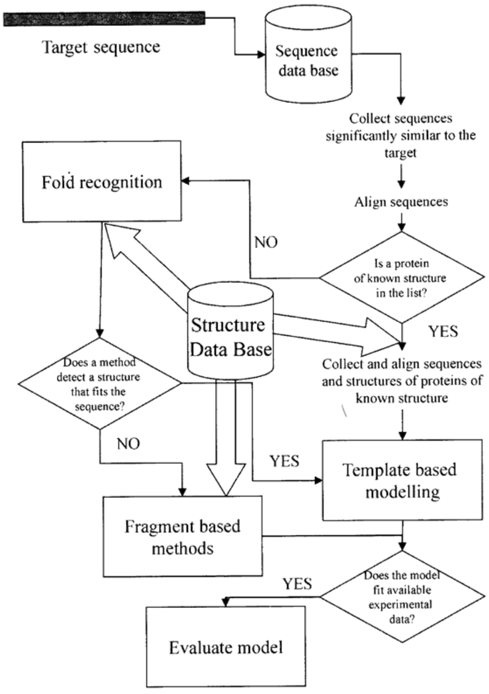

Target sequence
Sequence data base
Collect sequences significantly similar to the target
Align sequences
Is a protein of known structure in the list?
NO
YES
Fold recognition
Structure Data Base
Collect and align sequences and structures of proteins of known structure
Does a method detect a structure that fits the sequence?
NO
YES
Template based modelling
Fragment based methods
Does the model fit available experimental data?
YES
Evaluate model

Figure 4.1 A guide to protein-structure prediction. The first step is always a search in the protein sequence database. Comparative modeling should be used when a protein of known structure sharing sequence similarity with the protein under examination is present in the database. If this is not so, fold-recognition methods should be applied and, should they fail, the user should resort to new fold or fragment-based methods. Note the central role played by the structure database in all these heuristic methods.

Comparative protein structure modelling

Classical procedure for construction of a homology model

1. Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
2. If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
3. Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
4. Model the regions outside the conserved core.
5. Model the positions of the side-chains of the target.
6. Optimize the final three-dimensional structure.

Scheme of comparative modelling

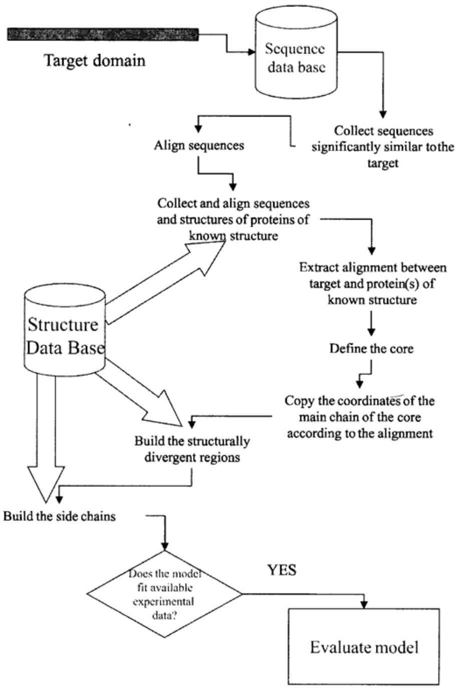
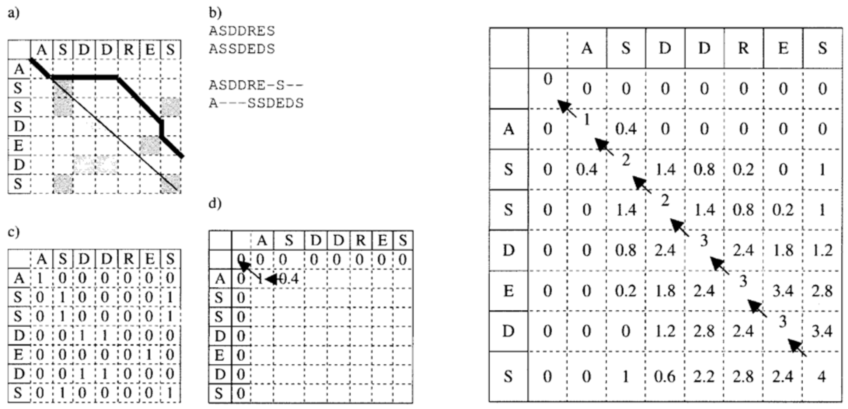

Target domain
Sequence data base
Align sequences
Collect sequences significantly similar to the target
Collect and align sequences and structures of proteins of known structure
Structure Data Base
Extract alignment between target and protein(s) of known structure
Define the core
Copy the coordinates of the main chain of the core according to the alignment
Build the structurally divergent regions
Build the side chains
Does the model fit available experimental data?
YES
Evaluate model

Figure 4.2 Schematic diagram of a typical comparative modeling procedure. The protein of interest should first be split into its domains. For each domain, sequences similar to the target sequences should be collected using a database search tool such as FASTA, BLAST, or PSI-BLAST. The sequences retrieved should be realigned using a multiple sequence alignment program (for example CLUSTAL or T-COFFEE). The implied alignment between the target protein and the protein(s) of known structure will form the basis of construction of the model. This can proceed by first building the main chain of the core regions, then the main chain of the structurally divergent regions, and, finally, the side-chains. The final evaluation of the model should take into account any available information on the protein of interest.

Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.

Needleman-Wunsch alignment algorithm



ASDDRES
ASSDEDS

ASDDRE-S--
A---SSDEDS

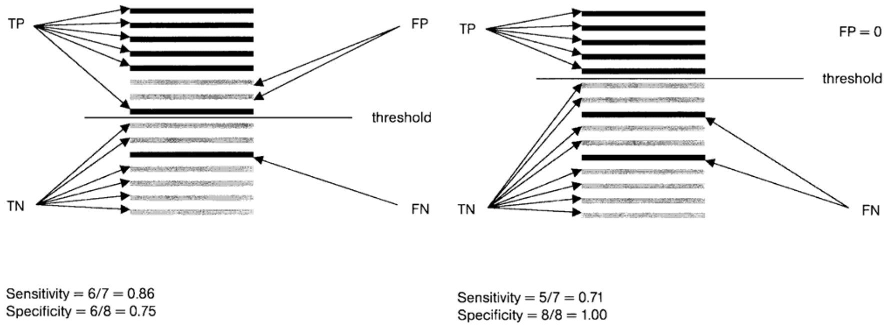
	A	S	D	D	R	E	S
A	1	0	0	0	0	0	0
S	0	1	0	0	0	0	1
S	0	1	0	0	0	0	1
D	0	0	1	1	0	0	0
E	0	0	0	0	0	1	0
D	0	0	1	1	0	0	0
S	0	1	0	0	0	0	1

		A	S	D	D	R	E	S
	0	0	0	0	0	0	0	0
A	0	1	0.4					
S	0							
S	0							
D	0							
E	0							
D	0							
S	0							

		A	S	D	D	R	E	S
	0	0	0	0	0	0	0	0
A	0	1	0.4	0	0	0	0	0
S	0	0.4	2	1.4	0.8	0.2	0	1
S	0	0	1.4	2	1.4	0.8	0.2	1
D	0	0	0.8	2.4	3	2.4	1.8	1.2
E	0	0	0.2	1.8	2.4	3	3.4	2.8
D	0	0	0	1.2	2.8	2.4	3	3.4
S	0	0	1	0.6	2.2	2.8	2.4	4

Figure 4.4 The Needleman and Wunsch alignment algorithm. A path in the matrix corresponds to an alignment. In the example, the thin line in part *a* of the figure corresponds to the first alignment shown in part *b*. The line runs diagonally and therefore corresponds to an alignment where there are no insertions or deletions. The tick line, instead, contains an horizontal line (indicating that the amino acids SDD of the first sequence do not correspond to any amino acid of the second and therefore represent an insertion in the first sequence) and two vertical lines (implying that the amino acid D and the final DS pair of the second sequence do not correspond to any amino acid in the first and is an insertion in the second sequence or, equivalently, a deletion in the first). To compute the optimum alignment we fill the cells of the matrix (part *c*) with a number representing the likelihood that the amino acid in the row is replaced by that in the column. In this example we assign 1 to identical amino acids and 0 to different ones. Part *d* shows the construction of the cumulative matrix as described in the text.

Sensitivity and specificity



Sensitivity = TP / (TP + FN) Specificity = TN / (TN + FP)

Figure 4.8 Examples of sensitivity and specificity values for a database search method. In the figure, dark and light segments, respectively, represent proteins homologous and unrelated to the query sequence. If we select the threshold as shown in the top part of the figure, two unrelated sequences will be labeled as "homologous" and one homologous one as "unrelated". A more stringent threshold (bottom), will eliminate false positives, but will increase the number of false negatives.

True positives vs. false negatives

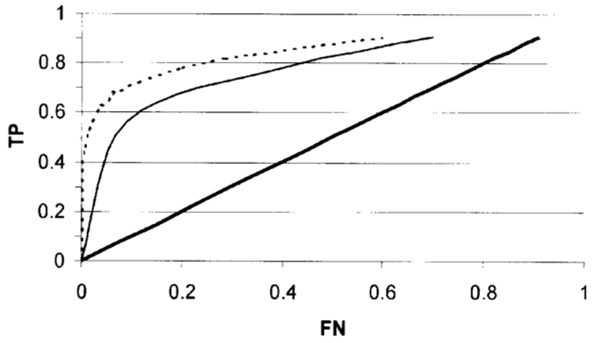


Figure 4.9 Examples of ROC curves. The tick line corresponds to a worthless method, unable to discriminate between positives and negatives. The method represented by the dotted curve is better than that represented by the continuous line: it detects more true positives when finding the same number of false negatives.

Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.

Domains (BLAST)

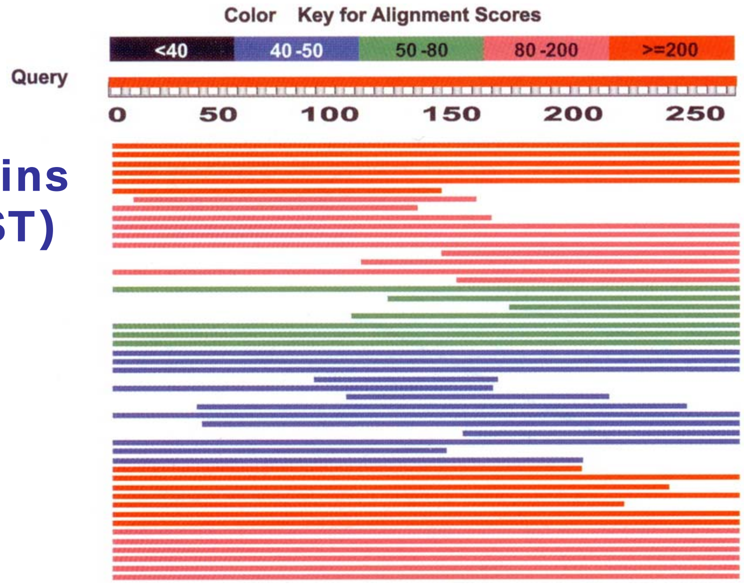
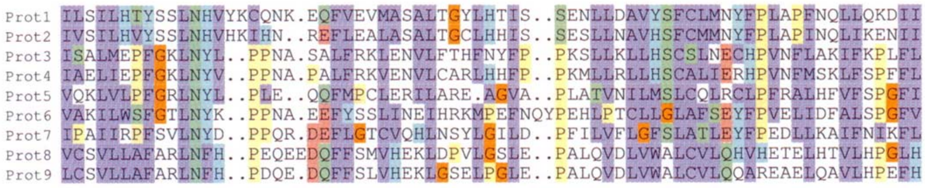


Figure 4.10 Example of the graphical output of BLAST. The example shown suggests that the query protein is formed by two domains, one spanning from the beginning to approximately residue 150, the other from approximately residue 150 to the end of the protein.

http://blast.ncbi.nlm.nih.gov

Multiple sequence alignment



```
Prot1 ILSILHTYSSLNHVYKCQNK.EQFVEVMASALTGYLHTIS..SENLLDAVYSFCLMNYFPLAPFNQLLQKDII
Prot2 IVSILHVYSSLNHVHKIHN..REFLEALASALTGCLHHIS..SESLLNAVHSFCMMNYFPLAPINQLIKENII
Prot3 ISALMEPFGKLNYL..PPNA.SALFRKLENVLFTHFNYFP..PKSLLKLLHSCSLNECHPVNFLAKIFKPLFL
Prot4 IAELIEPFGKLNYV..PPNA.PALFRKVENVLCARLHHFP..PKMLLRLLHSCALIERHPVNFMSKLFSPFFL
Prot5 VQKLVLPFGRLNYL..PLE..QQFMPCLERILARE.AGVA..PLATVNILMSLCQLRCLPFRALHFVFSPGFI
Prot6 VAKILWSFGTLNYK..PPNA.EEFYSSLINEIHRKMPEFNQYPEHLPTCLLGLAFSEYFPVELIDFALSPGFV
Prot7 IPAIIRPFSVLNYD..PPQR.DEFLGTCVQHLNSYLGILD..PFILVFLGFSLATLEYFPEDLLKAIFNIKFL
Prot8 VCSVLLAFARLNFH..PEQEEDQFFSMVHEKLDPVLGSLE..PALQVDLVWALCVLQHVHETELHTVLHPGLH
Prot9 LCSVLLAFARLNFH..PDQE.DQFFSLVHEKLGSELPGLE..PALQVDLVWALCVLQQAREAELQAVLHPEFH
```

Figure 4.11 A multiple sequence alignment. Note that completely conserved amino acids are easier to detect when more sequences are considered.

Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.

Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.

Building structurally divergent regions

- Reinspect alignment, e.g. shift gaps/insertions outside regular secondary structure elements
- Short canonical loops (type I, type II etc.)
- Rely on sequence pattern
- Loops that form compact substructures: internal H-bonds
- Packing inward pointing side-chain between secondary structure elements connected by the loop

Loops with similar conformation

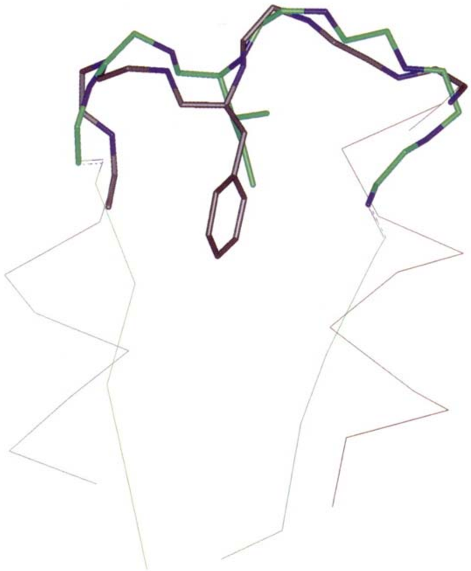


Figure 4.16 The figure shows two loops with similar conformations stabilized by the packing of a central hydrophobic amino acid. Note that one of the loops connects two alpha helices and the other two beta strands.

Similar loops, different environment

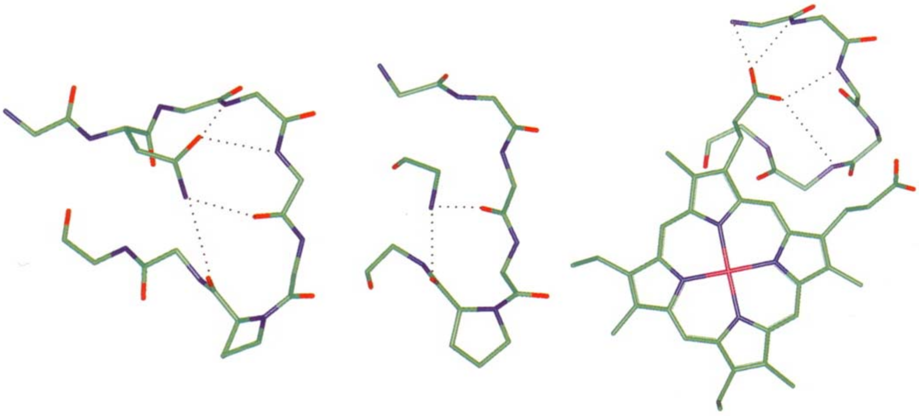
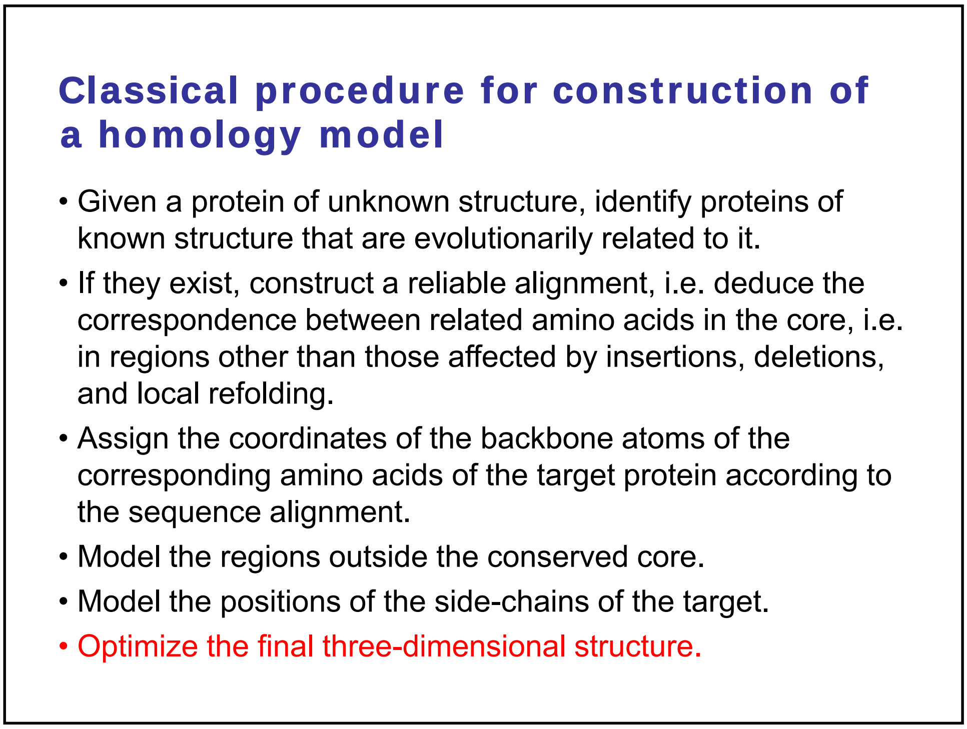


Figure 4.17 The three loops shown in the figure are very similar and stabilized by hydrogen-bonds, however the partners of these interactions are different in the three different proteins (an immunoglobulin, a viral protein, and a cytochrome).

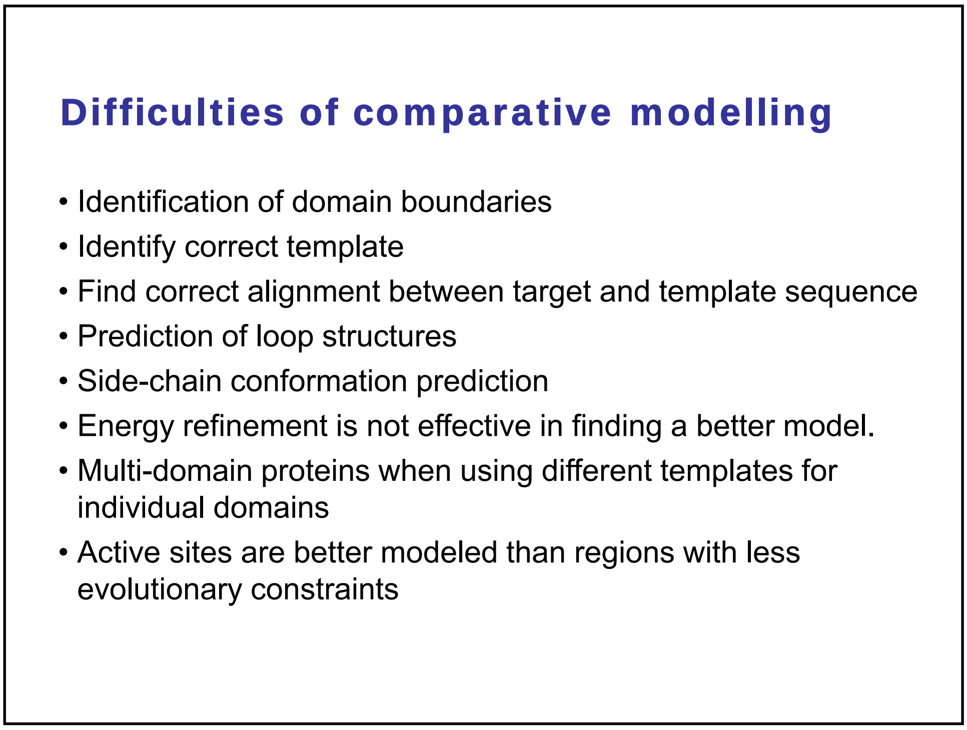
Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.



Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.



Difficulties of comparative modelling

- Identification of domain boundaries
- Identify correct template
- Find correct alignment between target and template sequence
- Prediction of loop structures
- Side-chain conformation prediction
- Energy refinement is not effective in finding a better model.
- Multi-domain proteins when using different templates for individual domains
- Active sites are better modeled than regions with less evolutionary constraints

Prediction accuracy

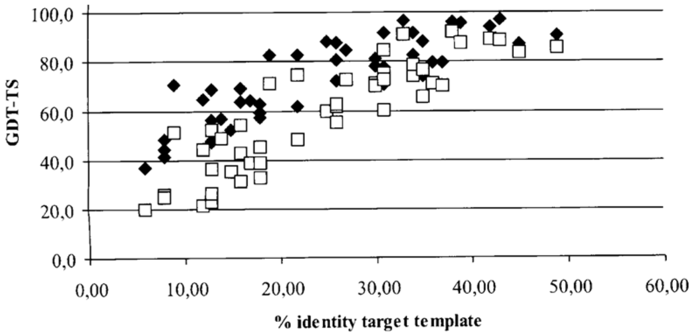


Figure 4.21 The relationship between the GDT-TS of the best (filled symbols) and average (open symbols) models and the sequence identity between the target protein sequence and the sequence of the best structural template. The data are taken from the CASP5 results and indicate that, above 40% sequence identity between target and template sequence, most methods can produce very respectable models. In more difficult examples the best methods can still produce useful results, but the gap between the quality of their results and those that can be obtained on average increases.

Comparative modelling examples

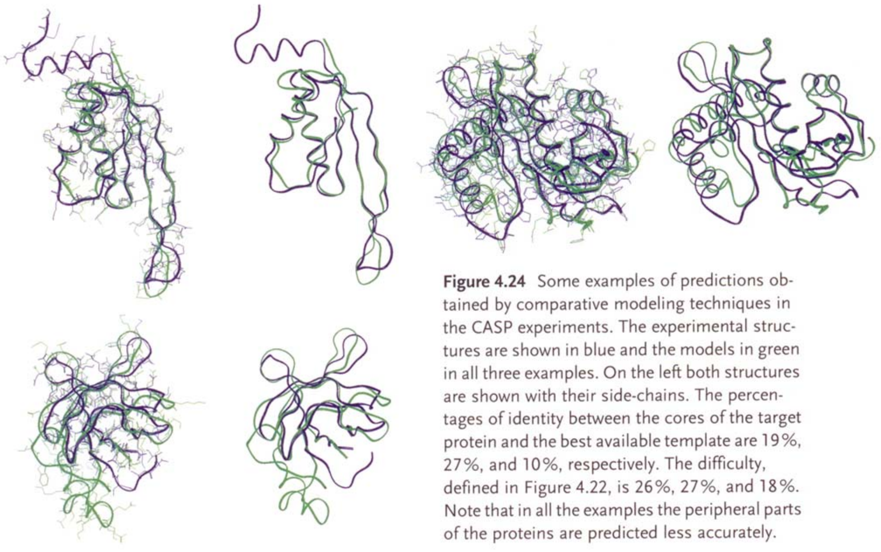


Figure 4.24 Some examples of predictions obtained by comparative modeling techniques in the CASP experiments. The experimental structures are shown in blue and the models in green in all three examples. On the left both structures are shown with their side-chains. The percentages of identity between the cores of the target protein and the best available template are 19%, 27%, and 10%, respectively. The difficulty, defined in Figure 4.22, is 26%, 27%, and 18%. Note that in all the examples the peripheral parts of the proteins are predicted less accurately.

Fold recognition

Methods for protein structure prediction

Methods are distinguished according to the relationship between the target protein(s) and proteins of known structure:

- **Comparative modelling**: A clear evolutionary relationship between the target and a protein of known structure can be easily detected from the sequence.
- **Fold recognition:** The structure of the target turns out to be related to that of a protein of known structure although the relationship is difficult, or impossible, to detect from the sequences.
- **New fold prediction:** Neither the sequence nor the structure of the target protein are similar to that of a known protein.

Structural similarity $\nrightarrow$ Sequence identity

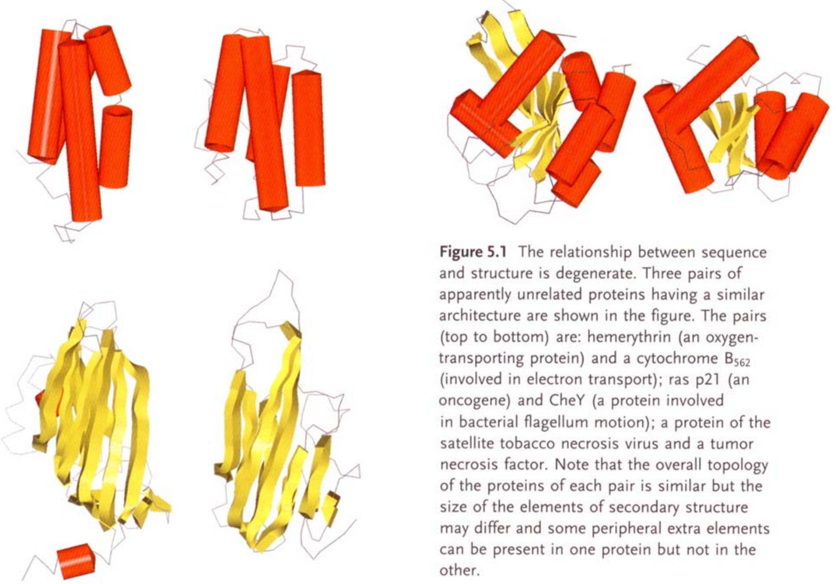


Figure 5.1 The relationship between sequence and structure is degenerate. Three pairs of apparently unrelated proteins having a similar architecture are shown in the figure. The pairs (top to bottom) are: hemerythrin (an oxygen-transporting protein) and a cytochrome B_{562} (involved in electron transport); ras p21 (an oncogene) and CheY (a protein involved in bacterial flagellum motion); a protein of the satellite tobacco necrosis virus and a tumor necrosis factor. Note that the overall topology of the proteins of each pair is similar but the size of the elements of secondary structure may differ and some peripheral extra elements can be present in one protein but not in the other.

Non-uniform distribution of folds

- Few (~10) folds are shared by a large number (~30%) of known proteins
- Large diversity in sequences and functions among members of these “superfolds”

Examples:

- Immunoglobulin fold
- Rossman fold
- TIM barrel fold
- Globin fold

Inverse protein folding problem

Which amino acid sequences fold into a known three-dimensional structure?

Protein folding problem

Which three-dimensional structure is adopted by a given amino acid sequence?

Fold recognition methods

- **3D profile methods**
  Physico-chemical properties of the amino acids of the target protein must “fit” with the environment in which they are placed in the modeled structure.
- **Threading**
  Sequences are fitted directly onto the backbone coordinates of known protein structures.

Profile method for fold recognition

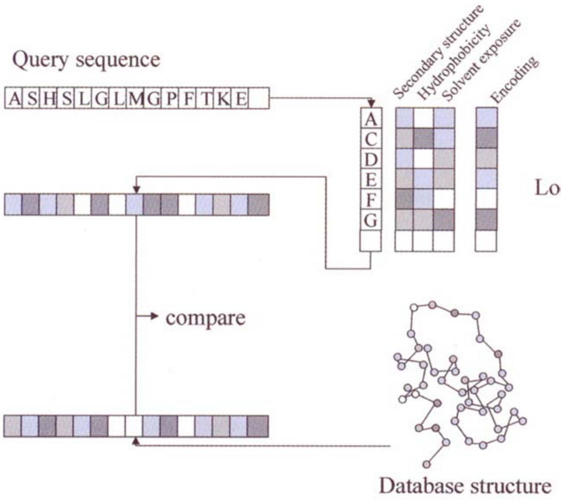
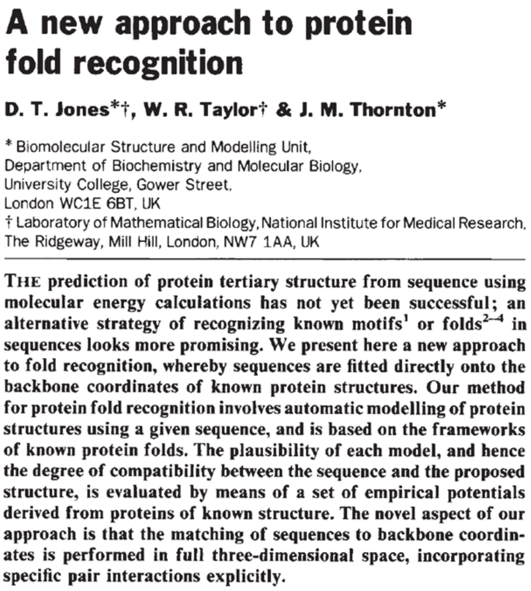


Figure 5.2 Schematic diagram of a possible profile-based method for fold recognition. The amino acids of the query sequence are replaced by a code that summarizes their hydrophobicity and their propensity for secondary structure type and solvent exposure. Each structure in the database is also encoded as a string by assigning a code to each of its amino acid positions. The code reflects their structural environment (secondary structure, solvent accessibility, and hydrophobicity of their environment). This does not depend on the actual amino acid present in the position analyzed. The string encoding the query sequence and each of the strings encoding the database structures are aligned and compared.

Bowie, Lüthy & Eisenberg. *Science* 253, 164-170 (1991)

Threading

- Sequences are fitted directly onto the backbone coordinates of known protein structures.
- Matching of sequences to backbone coordinates is performed in 3D space, incorporating specific pair interactions explicitly.



A new approach to protein fold recognition

D. T. Jones*†, W. R. Taylor† & J. M. Thornton*

* Biomolecular Structure and Modelling Unit,
Department of Biochemistry and Molecular Biology,
University College, Gower Street,
London WC1E 6BT, UK
† Laboratory of Mathematical Biology, National Institute for Medical Research,
The Ridgeway, Mill Hill, London, NW7 1AA, UK

THE prediction of protein tertiary structure from sequence using molecular energy calculations has not yet been successful; an alternative strategy of recognizing known motifs[1] or folds[2-4] in sequences looks more promising. We present here a new approach to fold recognition, whereby sequences are fitted directly onto the backbone coordinates of known protein structures. Our method for protein fold recognition involves automatic modelling of protein structures using a given sequence, and is based on the frameworks of known protein folds. The plausibility of each model, and hence the degree of compatibility between the sequence and the proposed structure, is evaluated by means of a set of empirical potentials derived from proteins of known structure. The novel aspect of our approach is that the matching of sequences to backbone coordinates is performed in full three-dimensional space, incorporating specific pair interactions explicitly.

Nature 358, 86-89 (1992)

Threading

- A library of different protein folds is derived from the database of protein structures.
- Each fold is considered as a chain tracing through space; the original sequence being ignored completely.
- The test sequence is then optimally fitted to each library fold, allowing for relative insertions and deletions in loop regions.
- The ‘energy’ of each possible fit (or threading) is calculated by summing the proposed pairwise interactions and the solvation energy.
- The library of folds is then ranked in ascending order of total energy, with the lowest energy fold being taken as the most probable match.

Knowledge-based (pair) potentials

$$E(r) = -k_B T \ln[f(r)]$$

r distance between two atoms (or some other parameter, like dihedral angles or solvent accessible surface)

$E(r)$ is the energy at r

$f(r)$ is the probability density at r

k_B is the Boltzmann constant

T is the absolute temperature

Statistically derived potentials

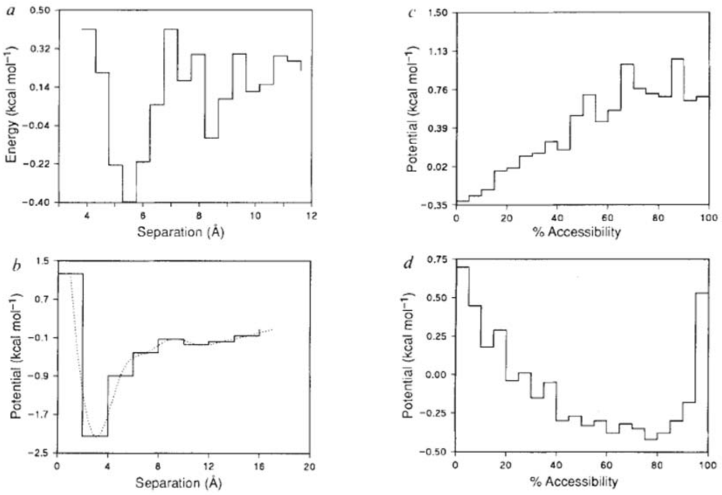


FIG. 1 Samples of the statistically derived potentials are shown. *a*, Short-range ($k = 3$) Ala-Ala $C\beta \rightarrow C\beta$ interaction. Low-energy states are observed for distances around 6 Å, corresponding mainly to α-structure, and 9 Å, corresponding mainly to β-structure. *b*, Long-range ($k > 30$) Cys-Cys $C\beta \rightarrow C\beta$ interaction. The most significant energy minimum around 4 Å corresponds to disulphide bridge formation. *c*, Solvation potential for leucine, and *d*, solvation potential for glutamic acid.

Fold recognition results from CASP

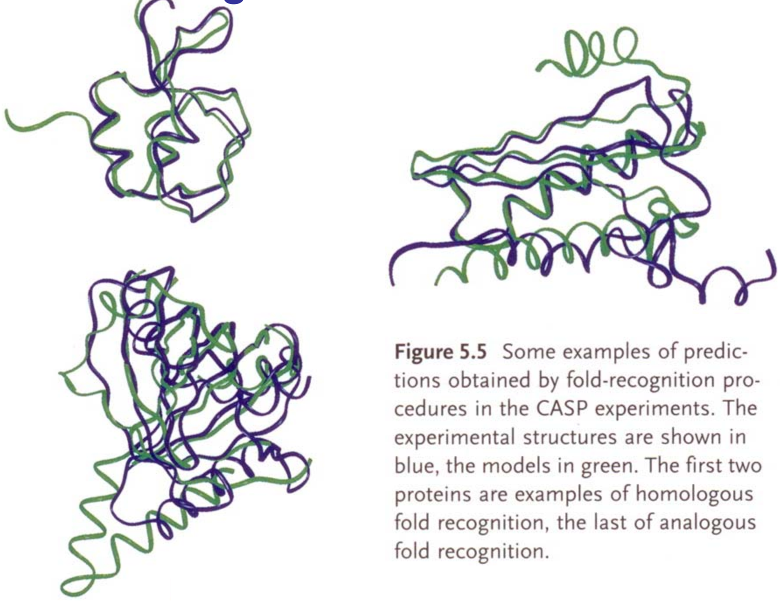


Figure 5.5 Some examples of predictions obtained by fold-recognition procedures in the CASP experiments. The experimental structures are shown in blue, the models in green. The first two proteins are examples of homologous fold recognition, the last of analogous fold recognition.

New fold prediction

Methods for protein structure prediction

Methods are distinguished according to the relationship between the target protein(s) and proteins of known structure:

- **Comparative modelling**: A clear evolutionary relationship between the target and a protein of known structure can be easily detected from the sequence.
- **Fold recognition:** The structure of the target turns out to be related to that of a protein of known structure although the relationship is difficult, or impossible, to detect from the sequences.
- **New fold prediction:** Neither the sequence nor the structure of the target protein are similar to that of a known protein.

CASP: Fragment-based predictions

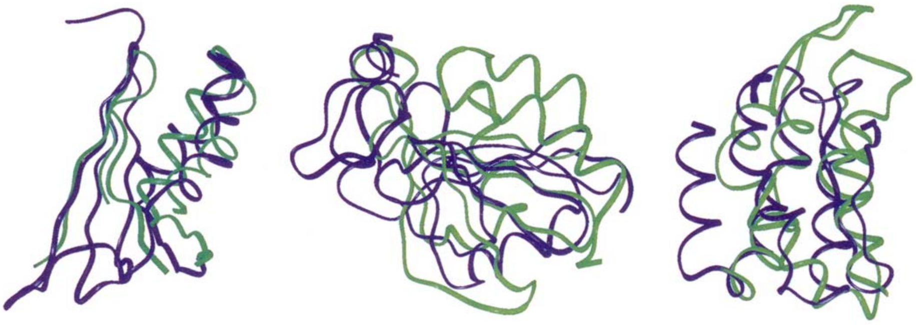
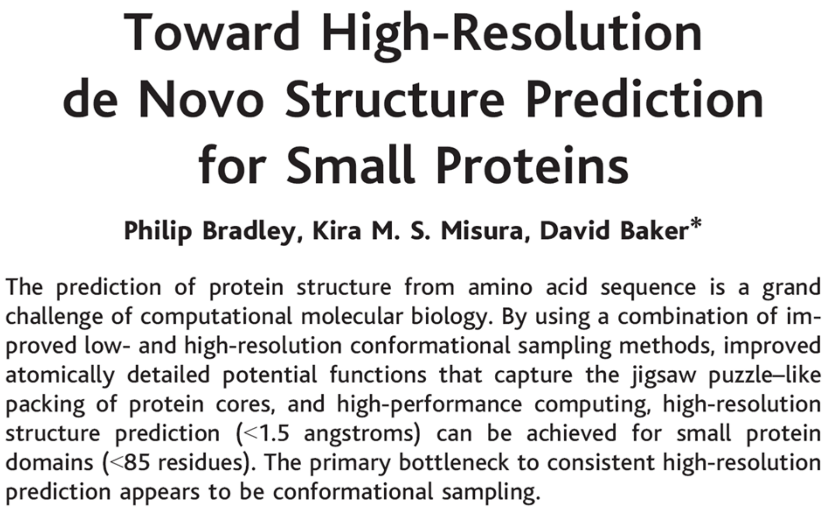


Figure 6.2 Some examples of fragment-based predictions submitted to CASP experiments.

Fragment-based approaches

- Rosetta (David Baker)
- Fragfold (David Jones)



Toward High-Resolution de Novo Structure Prediction for Small Proteins

Philip Bradley, Kira M. S. Misura, David Baker*

The prediction of protein structure from amino acid sequence is a grand challenge of computational molecular biology. By using a combination of improved low- and high-resolution conformational sampling methods, improved atomically detailed potential functions that capture the jigsaw puzzle–like packing of protein cores, and high-performance computing, high-resolution structure prediction (<1.5 angstroms) can be achieved for small protein domains (<85 residues). The primary bottleneck to consistent high-resolution prediction appears to be conformational sampling.

Science 309, 1868–1871 (2005)

Steps of fragment-based structure prediction

- Split sequence into fragments
- For each fragment, search the database of known structures for regions with a similar sequence (“neighbors”)
- Use an optimization technique to find the best combination of fragments

Fragment search

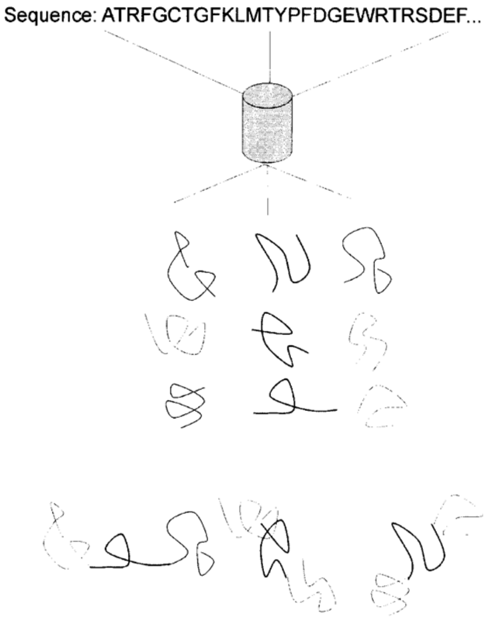
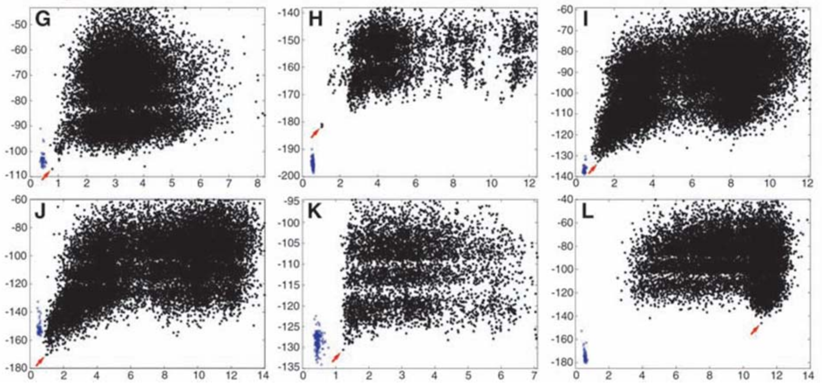


Figure 6.3 Schematic explanation of the first steps of the Rosetta method. The query sequence is split in fragments nine amino acids long. Each fragment sequence is used to search for similar fragments among the sequences of proteins of known structure. Next, the fragments are joined.

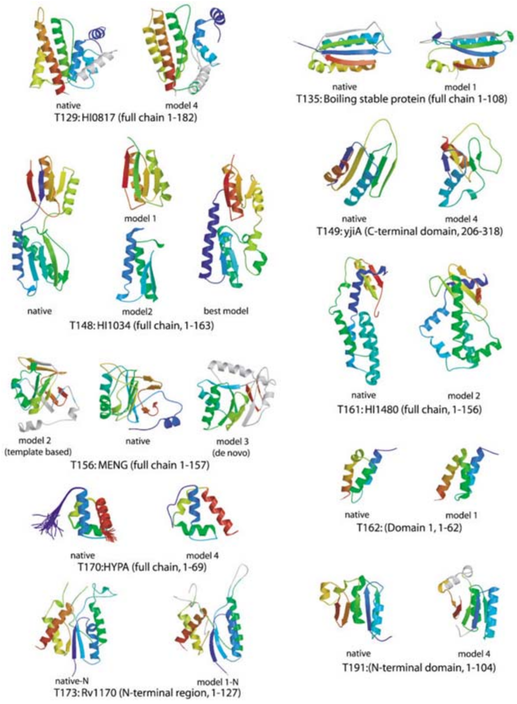
Energy vs. accuracy



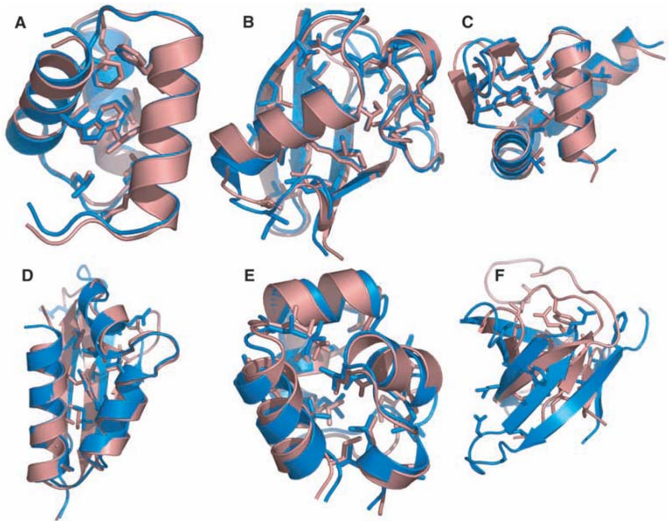
Plots of C^{α}-RMSD (x axis) against all atom energy (y axis) for refined natives (blue points) and the de novo models (black points). Red arrows indicate the lowest energy de novo models.

ROSETTA results in CASP5

Ribbon diagrams of predictions made by using the fragment insertion approach. The native structure and best submitted model are shown colored from the N-terminus (blue) to C-terminus (red). For T148, the best generated model is also shown, and for T156, both template-based and fragment insertion based models are shown. For targets T173, T135, T156, and T191, colored regions deviate from the native structure by <4 Å, and gray regions deviate by >4 Å. For targets T129 and T156, colored regions deviate from the native structure by <6 Å C$^\alpha$ RMSD, whereas the gray regions deviate by >6 Å.


native
model 4
T129: HI0817 (full chain 1-182)
native
model 1
T135: Boiling stable protein (full chain 1-108)
model 1
native
model2
best model
T148: HI1034 (full chain, 1-163)
native
model 4
T149: yjiA (C-terminal domain, 206-318)
native
model 2
T161: HI1480 (full chain, 1-156)
model 2
(template based)
native
model 3
(de novo)
T156: MENG (full chain 1-157)
native
model 1
T162: (Domain 1, 1-62)
native
model 4
T170:HYPA (full chain, 1-69)
native
model 4
T191:(N-terminal domain, 1-104)
native-N
model 1-N
T173: Rv1170 (N-terminal region, 1-127)

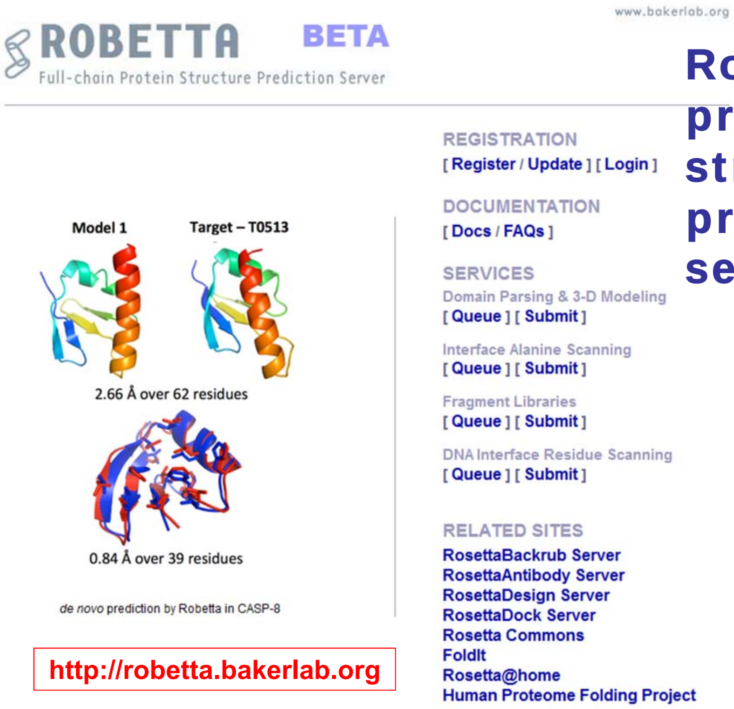
High-resolution de novo structure predictions



Superposition of low-energy models (blue) with experimental structures (red) showing core side chains.

A: Hox-B1
B: Ubiquitin
C: RecA
D: KH domain of Nova-2
E: 434 repressor
F: Fyn tyrosine kinase

Robetta protein structure prediction server



www.bakerlab.org

ROBETTA BETA

Full-chain Protein Structure Prediction Server

Model 1 Target – T0513

2.66 Å over 62 residues

0.84 Å over 39 residues

de novo prediction by Robetta in CASP-8

REGISTRATION

[Register / Update] [Login]

DOCUMENTATION

[Docs / FAQs]

SERVICES

Domain Parsing & 3-D Modeling

[Queue] [Submit]

Interface Alanine Scanning

[Queue] [Submit]

Fragment Libraries

[Queue] [Submit]

DNA Interface Residue Scanning

[Queue] [Submit]

RELATED SITES

RosettaBackrub Server

RosettaAntibody Server

RosettaDesign Server

RosettaDock Server

Rosetta Commons

FoldIt

Rosetta@home

Human Proteome Folding Project

http://robetta.bakerlab.org

Literatur

- Anna Tramontano: *Protein Structure Prediction, Wiley-VCH*, 2006.