Strukturelle Bioinformatik
(Masterstudiengang Biochemie)

Strukturmodellierung

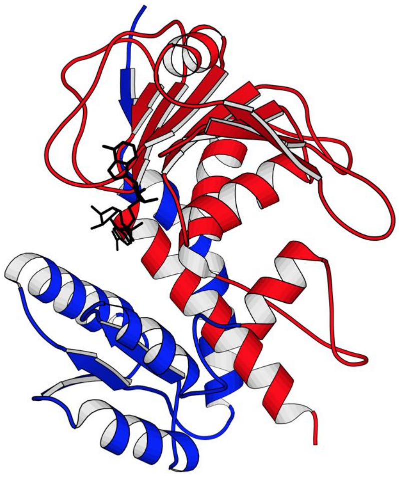
Sommersemester 2019

Peter Güntert & Sina Kazemi

A conceptually simple problem

```
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHL
KTEAEMKASEDLKKAGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKI
PIKYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDIAAKYKEL
GYQG
```

⇩



Methods

1. **Homology modeling/comparative modeling**
   - Similar sequences → similar structures
   - Practically very useful, but requires structural homologues

2. **Fold recognition and threading**
   - Many sequence-wise unrelated proteins share the same structural fold
   - Structures are more conserved than sequences

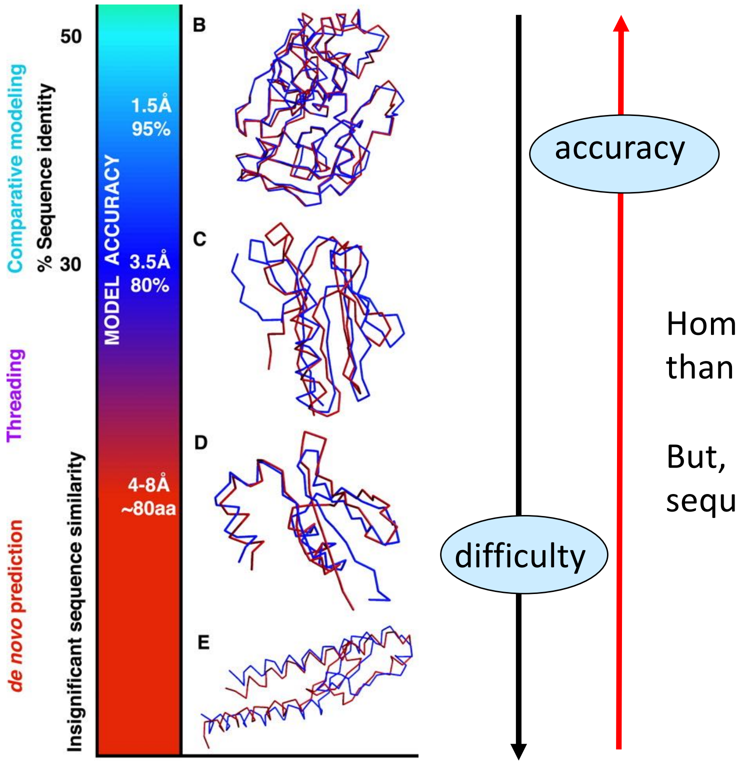
3. **ab initio (or template-free methods)**
   - Can use first principles to fold proteins
   - Do not require templates
   - High computational complexity

Methods for protein structure prediction

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

- **Comparative modeling**: 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.

Structure prediction



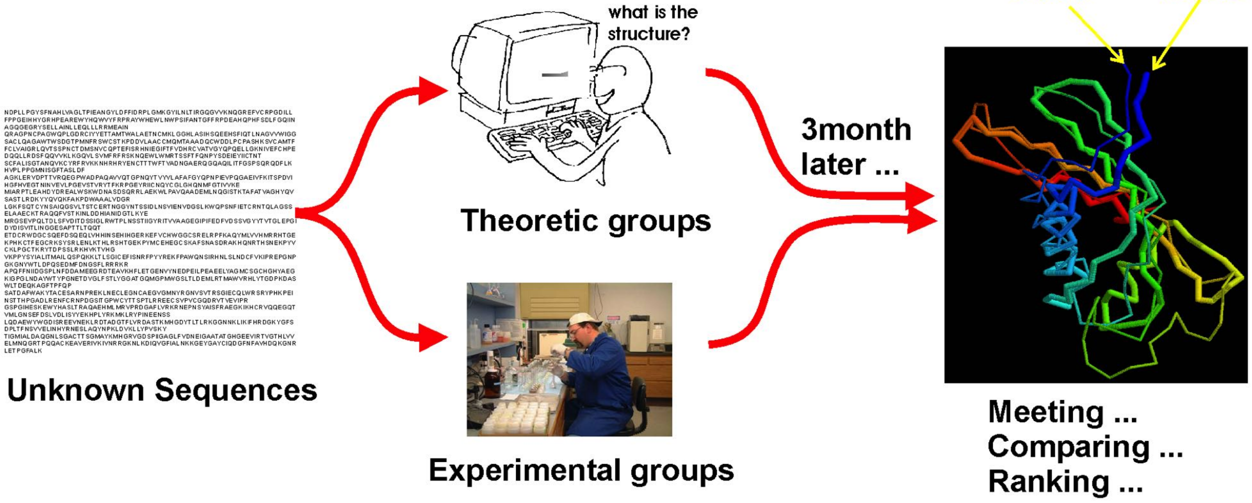
Homology modelling is more reliable than other methods.

But, you can’t always find similar sequences of known structure.

Baker, D, Sali, A. (2001). *Science* 294, 93-96

CASP: Critical Assessment of Techniques for Protein Structure Prediction

CASP (Critical Assessment of Structure Prediction) is a community wide experiment to determine and advance the state of the art in modeling protein structure from amino acid sequence. Every two years since 1994, participants are invited to submit models for a set of proteins for which the experimental structures are not yet public. Independent assessors then compare the models with experiment. Assessments and results are published in a special issue of the journal *Proteins*. In the most recent CASP round, CASP12, nearly 100 groups from around the world submitted more than 50,000 models on 82 modeling targets.

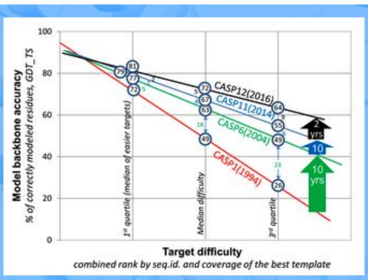


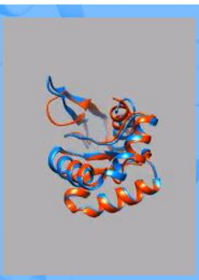
http://predictioncenter.org/

CASP: Template-based modelling

template-based modeling

Models based on templates identified by sequence similarity remain the most accurate. Over the course of the CASP experiments there have been enormous improvements in this area. However, the overall accuracy improvements that we have seen in the first 10 years of CASP remained unmatched until CASP12, when a new burst of progress happened (see the plot). In two years from CASP11 to CASP12 the backbone accuracy of the submitted models improved more than in the preceeding 10 years. Several factors contributed to this, including more accurate alignment of the target sequence to that of available templates, combining multiple templates, improved accuracy of regions not covered by templates, successful refinement of models, and better selection of models from decoy sets due to improved methods for estimation of model accuracy. [Kryshtafovych et al, 2018]





target T0868-D1 (orange)
model 330_2 (blue): GDT_TS=87
best template: 2cw6 (seq.id= 4.2%)

http://predictioncenter.org/

CASP: Ab initio modelling

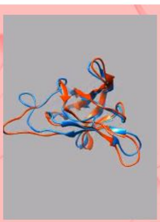
ab initio modeling

Modeling proteins with no or marginal similarity to existing structures (*ab initio, new fold, non-template* or *free* modeling) is the most challenging task in tertiary structure prediction. Probably the first ab initio model of reasonable accuracy was built in CASP4. Since then CASP witnessed sustained progress in ab initio prediction, but mainly for small proteins (120 residues or less, panels 1 and 2). In CASP11 for the first time a larger new fold protein (256 residues, sequence identity to known structures <5%) was built with unpresedented before accuracy for targets of this size (panel 3). The latest two CASPs (2014-2016) also showed a new trend in building better non-template models by successful utilizing predicted contacts (panel 4). [Abriata et al, 2018] Models are shown in blue, targets in orange.

CASP7: T0283-D1
model 321_1: GDT_TS=75

CASP9: T0581-D1
model 170_1: GDT_TS=71

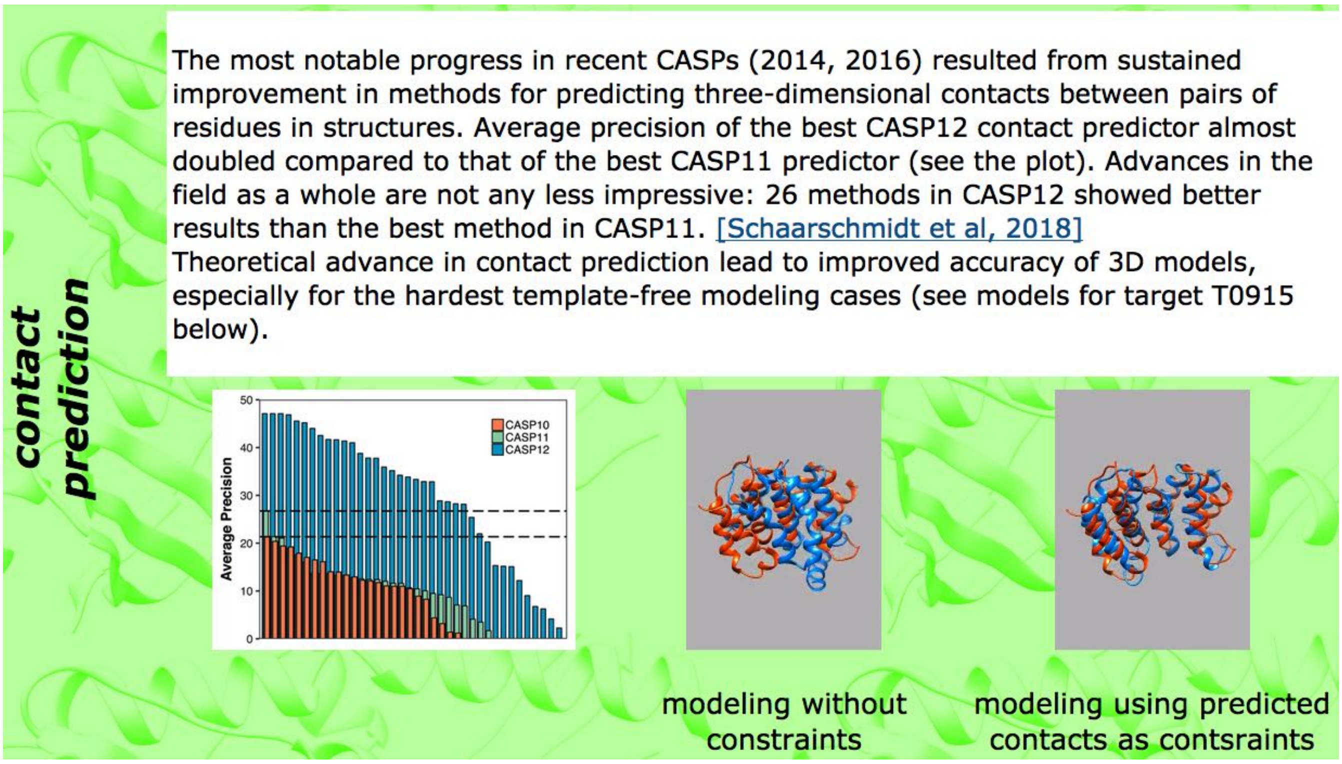
CASP11: T0806-D1
model 064_1: GDT_TS=61



CASP12: T0866-D1
model 325_5: GDT_TS=81

http://predictioncenter.org/

CASP: Contact prediction

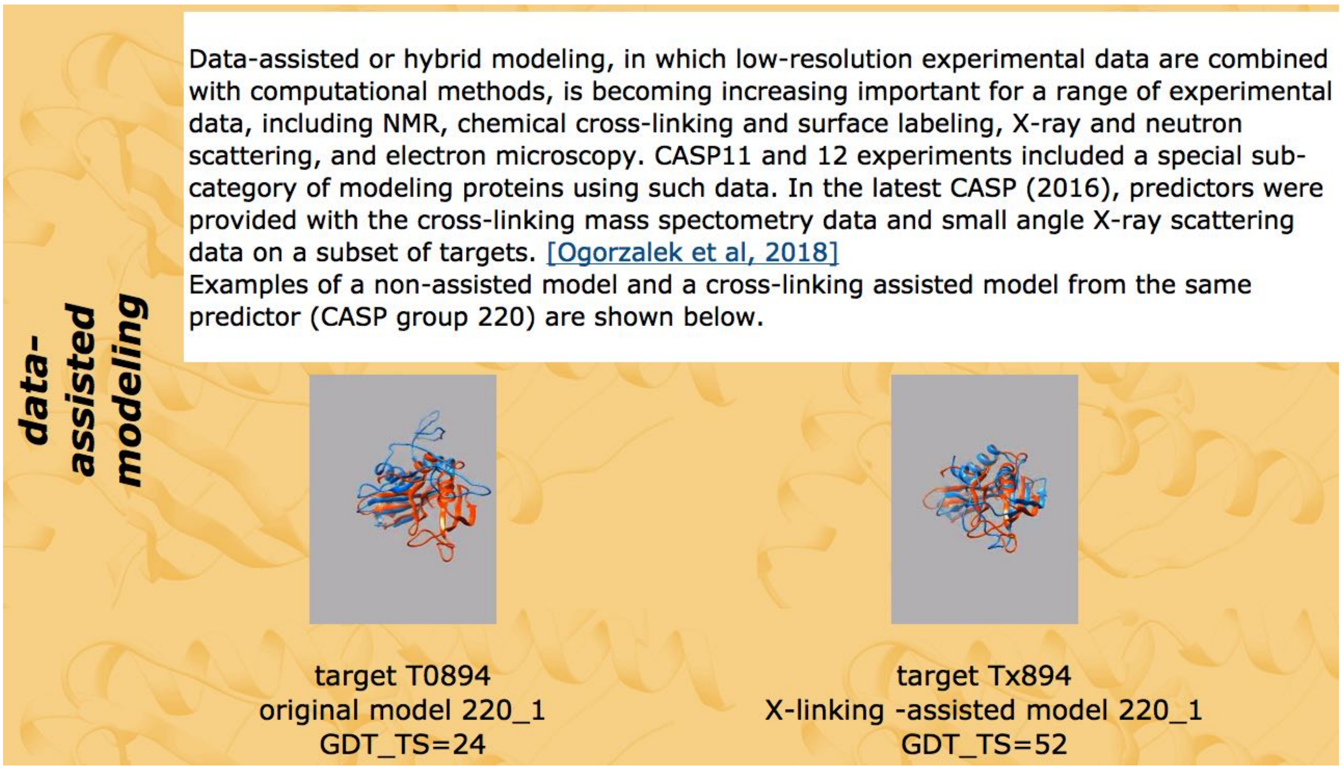


The most notable progress in recent CASPs (2014, 2016) resulted from sustained improvement in methods for predicting three-dimensional contacts between pairs of residues in structures. Average precision of the best CASP12 contact predictor almost doubled compared to that of the best CASP11 predictor (see the plot). Advances in the field as a whole are not any less impressive: 26 methods in CASP12 showed better results than the best method in CASP11. [Schaarschmidt et al, 2018]
Theoretical advance in contact prediction lead to improved accuracy of 3D models, especially for the hardest template-free modeling cases (see models for target T0915 below).

modeling without constraints

modeling using predicted contacts as contsraints

CASP: Experimental data-assisted modelling



Data-assisted or hybrid modeling, in which low-resolution experimental data are combined with computational methods, is becoming increasing important for a range of experimental data, including NMR, chemical cross-linking and surface labeling, X-ray and neutron scattering, and electron microscopy. CASP11 and 12 experiments included a special sub-category of modeling proteins using such data. In the latest CASP (2016), predictors were provided with the cross-linking mass spectometry data and small angle X-ray scattering data on a subset of targets. [Ogorzalek et al, 2018]
Examples of a non-assisted model and a cross-linking assisted model from the same predictor (CASP group 220) are shown below.

target T0894
original model 220_1
GDT_TS=24

target Tx894
X-linking -assisted model 220_1
GDT_TS=52

CASP13 Goals

CASP assesses many aspects of modeling, including the accuracy of protein topologies, atom co-ordinates, and multi-protein assemblies. The experiment also examines the extent to which models can answer questions of biological interest, and how different types of sparse or low resolution experimental data can improve model accuracy.

CASP13 has started in April 2018 and will address the following questions:

- How similar are the models to the corresponding experimental structure?
- Are domain orientations, subunit interactions, and the protein initeractions in complexes modeled correctly?
- How much more accurate are template-based models than those that can be obtained by simply copying the best template?
- How reliable are overall, residue, and atomic level error estimates?
- How much can current refinement methods improve the accuracy of models?
- How effective are approaches to predicting protein three dimensional contacts?
- How well do the models help answering relevant biological questions?
- How helpful is additional information, particularly sparse NMR data, chemical cross-linking, SAXS and FRET?
- In which areas has there been progress since the last CASP?
- Where can future effort be most productively focused?

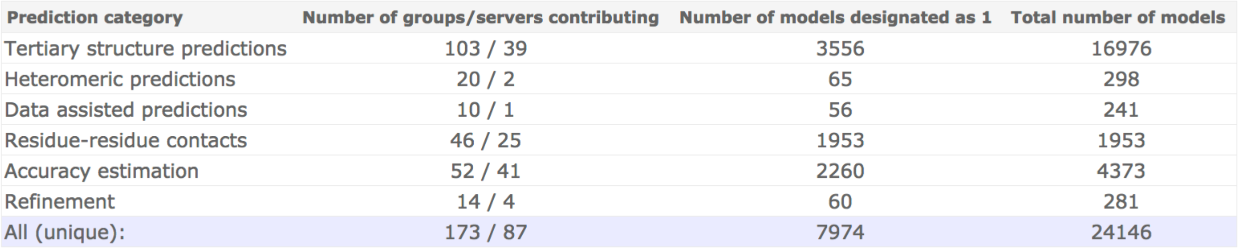
http://predictioncenter.org/casp13

CASP13 Modeling Categories

- The **High Accuracy Modeling** category will include domains where majority of submitted models are of sufficient accuracy for detailed analysis. This category replaces the previous Template Based Modeling category.
- The **Topology** category (formerly Free Modeling) will assess domains where submitted models are of relatively low accuracy.
- The **Contact Prediction** category will assess the ability of methods to predict three dimensional contacts in targets structures.
- The **Refinement** category will analyze success in refining models beyond the accuracy obtained in the initial submissions. For each target, one of the best initial models will be selected, and reissued as the starting structure for refinement.
- The **Assembly** category will assess how well current methods can determine domain-domain, subunit-subunit, and protein-protein interactions. As in CASPs 11 and 12, we hope to work closely with CAPRI in this category.
- The **Accuracy Estimation** category will assess the ability to provide useful accuracy estimates for the overall accuracy of models and at the domain and residue level.
- The **Data Assisted** category will assess how much the accuracy of models is improved by the addition of sparse data. Targets for which such data are available will be re-released after initial data independent models have been collected, together with the available data. Data types are expected to include sparse NMR data, crosslinking data, SAXS data and FRET.
- The **Biological Relevance** category will assess models on the basis of how well they provide answers to biological questions. Target providers will be asked to say what questions prompted the determination of the experimental structure. The usefulness of the models in answering those questions will be compared with the that of the experimental structures.

CASP13 in Numbers

Number of groups registered	**241**
including: *expert groups*	*149*
prediction servers	*92*
Number of tertiary structure prediction targets released	**53**
(including *all-group targets*)	*(45)*
Number of hetero-multimer targets released	**7**
Number of refinement targets released	**7**
Number of assisted prediction targets released	**16**
Targets canceled (all / human)	**(1 / 2)**
Targets available/expired for manual non-QA prediction	**21 / 23**
Targets available/expired for server non-QA prediction	**1 / 51**
Targets available/expired for QA prediction	**6 / 43**
Targets available/expired for assisted prediction	**9 / 7**
Targets available/expired for multimer prediction	**3 / 4**



Prediction category	Number of groups/servers contributing	Number of models designated as 1	Total number of models
Tertiary structure predictions	103 / 39	3556	16976
Heteromeric predictions	20 / 2	65	298
Data assisted predictions	10 / 1	56	241
Residue-residue contacts	46 / 25	1953	1953
Accuracy estimation	52 / 41	2260	4373
Refinement	14 / 4	60	281
All (unique):	173 / 87	7974	24146

http://predictioncenter.org/casp13

Measures of structural similarity

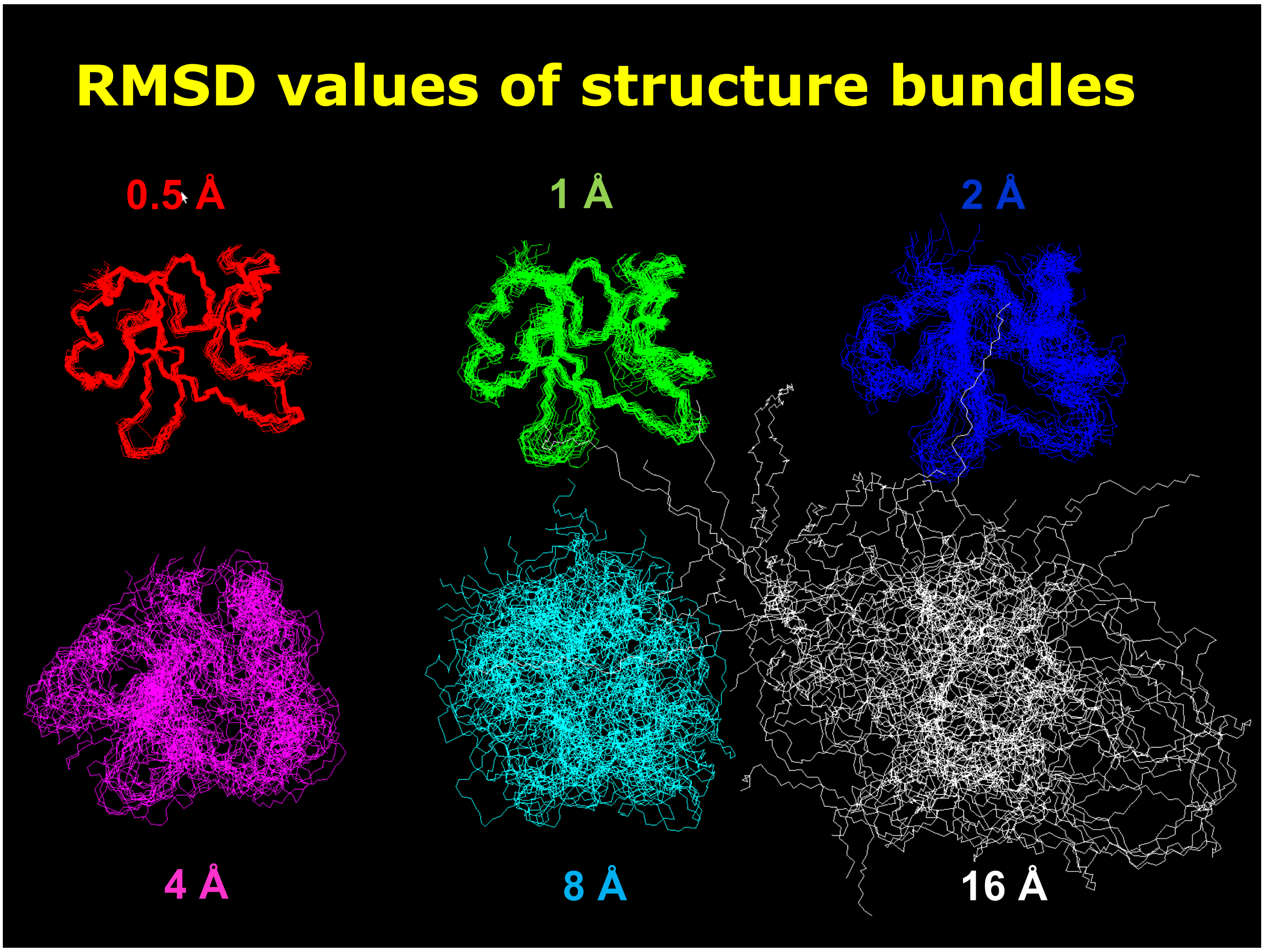
- **RMSD:** Average (root-mean-square) deviation of atom positions
- **GDT-TS:** Percentage of residues that can be superimposed under given distance cutoffs

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 $\boldsymbol{t}$ → optimale Überlagerung



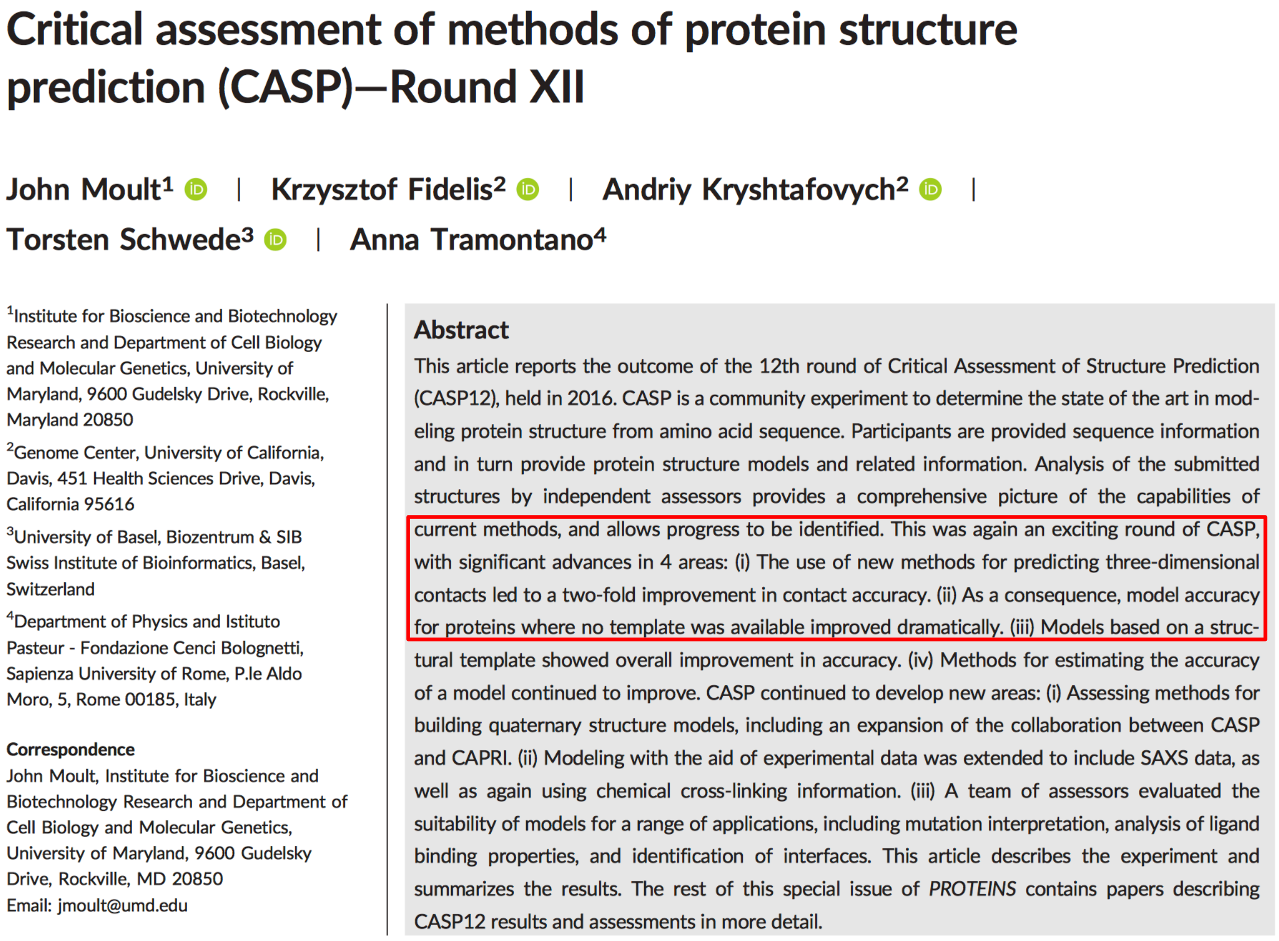
RMSD values of structure bundles

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.



Critical assessment of methods of protein structure prediction (CASP)—Round XII

John Moult[1] | Krzysztof Fidelis[2] | Andriy Kryshtafovych[2] |
Torsten Schwede[3] | Anna Tramontano[4]

[1] Institute for Bioscience and Biotechnology Research and Department of Cell Biology and Molecular Genetics, University of Maryland, 9600 Gudelsky Drive, Rockville, Maryland 20850

[2] Genome Center, University of California, Davis, 451 Health Sciences Drive, Davis, California 95616

[3] University of Basel, Biozentrum & SIB Swiss Institute of Bioinformatics, Basel, Switzerland

[4] Department of Physics and Istituto Pasteur - Fondazione Cenci Bolognetti, Sapienza University of Rome, P.le Aldo Moro, 5, Rome 00185, Italy

Correspondence
John Moult, Institute for Bioscience and Biotechnology Research and Department of Cell Biology and Molecular Genetics, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850
Email: jmoult@umd.edu

Abstract

This article reports the outcome of the 12th round of Critical Assessment of Structure Prediction (CASP12), held in 2016. CASP is a community experiment to determine the state of the art in modeling protein structure from amino acid sequence. Participants are provided sequence information and in turn provide protein structure models and related information. Analysis of the submitted structures by independent assessors provides a comprehensive picture of the capabilities of current methods, and allows progress to be identified. This was again an exciting round of CASP, with significant advances in 4 areas: (i) The use of new methods for predicting three-dimensional contacts led to a two-fold improvement in contact accuracy. (ii) As a consequence, model accuracy for proteins where no template was available improved dramatically. (iii) Models based on a structural template showed overall improvement in accuracy. (iv) Methods for estimating the accuracy of a model continued to improve. CASP continued to develop new areas: (i) Assessing methods for building quaternary structure models, including an expansion of the collaboration between CASP and CAPRI. (ii) Modeling with the aid of experimental data was extended to include SAXS data, as well as again using chemical cross-linking information. (iii) A team of assessors evaluated the suitability of models for a range of applications, including mutation interpretation, analysis of ligand binding properties, and identification of interfaces. This article describes the experiment and summarizes the results. The rest of this special issue of *PROTEINS* contains papers describing CASP12 results and assessments in more detail.

Moult et al. *Proteins* 86, 7–15 (2018).

CASP12 Prediction accuracy

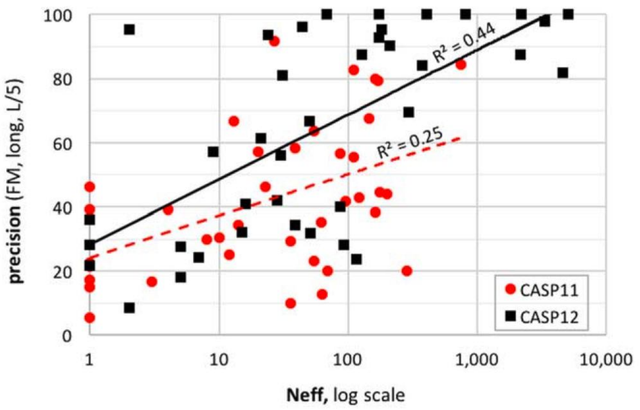


FIGURE 1 Contact prediction accuracy in CASPs 11 and 12 against effective alignment depth. As expected, accuracy increases with alignment depth, and for a number of CASP12 targets with deep alignments, precision is 100%. Best results on the set of free modeling targets are shown. Precision is for the most confidently predicted L/5 contacts separated by >23 residues in the sequence, where L is the target length. Neff is the number of diverse (<90% ID) homologous sequences covering at least 60% of the target with an E-score of 10^{-3} or better, retrieved by HHblits from the uniprot20 database

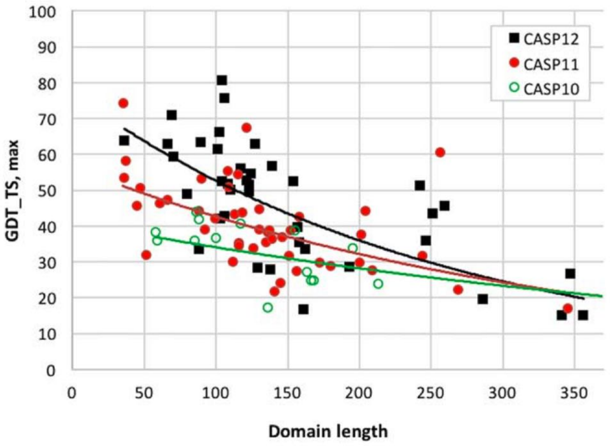


FIGURE 2 Backbone accuracy (GDT_TS) of the best submitted models in the free modeling category for the 3 most recent CASPs, as a function of target length. Good performance for targets smaller than 100 residues mostly reflects earlier improvements in this category. In CASP10, no models longer than 100 residues had GDT_TS >50. In CASP11, 4 crossed this threshold. In CASP12, half of the targets longer than 100 residues do so. (On the GDT_TS scale, 100 is perfect agreement with experiment, 20–30 is typically random, and structures with scores above 50 are largely topologically correct.)

Moult et al. *Proteins* 86, 7–15 (2018).

CASP12 Prediction accuracy

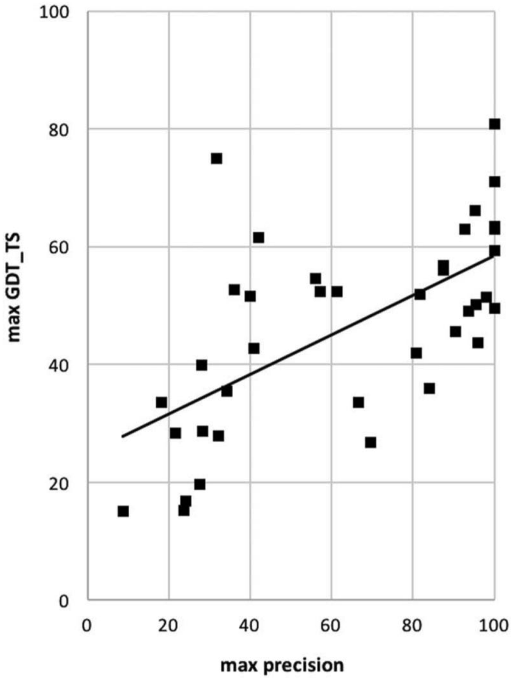


FIGURE 3 Relationship between highest backbone accuracy (GDT_TS) and highest contact prediction accuracy for free modeling targets in CASP12. Average structure accuracy doubles as contact accuracy increases, demonstrating that high accuracy is a consequence of the availability of largely correct contacts. (Precision is for the *L*/5 most confidently predicted contacts separated by at least 23 residues in the sequence, *L* is target length)

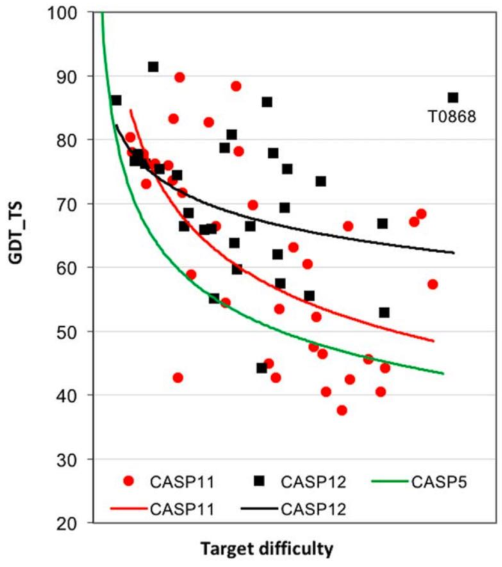


FIGURE 5 Trend lines for best model backbone accuracy (by GDT_TS) in CASP5 (2002), CASP11, and the most recent CASP12, for the template-based modeling targets (TBM and TBM/FM). By this measure, there was only modest improvement in 12 years between CASP5 and 11, but a substantial jump in the last 2 years. Points show the CASP11 and CASP12 best models for each target. The case of T0868 is discussed in the text and shown in Figure 6. The "Target Difficulty" rank of each target is based on its sequence and structure similarity to the closest template[14]

Moult et al. *Proteins* 86, 7–15 (2018).

CASP12 Prediction accuracy

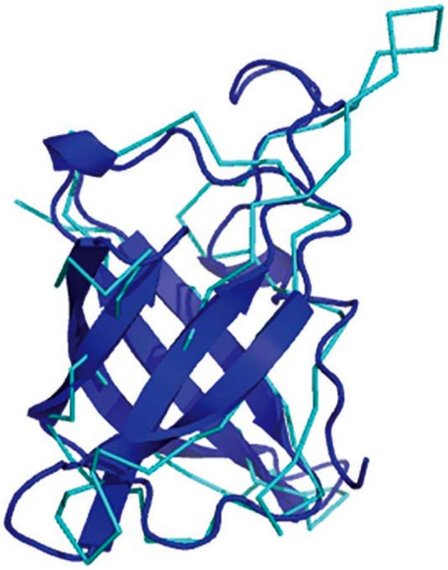


FIGURE 4 Superposition of the best model received for target T0866, the periplasmic domain of MlaD from *E.coli* (blue), with the corresponding experimental structure (turquoise, PDB 4cx8). There were no sequence detectable templates for this protein, and the outstandingly accurate model is largely because of successful prediction of a set of three-dimensional contacts

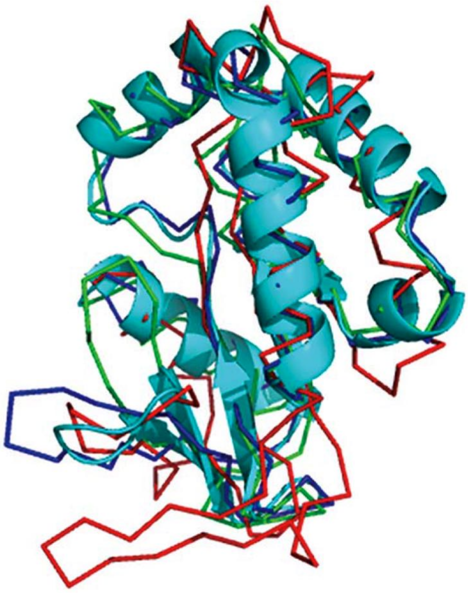
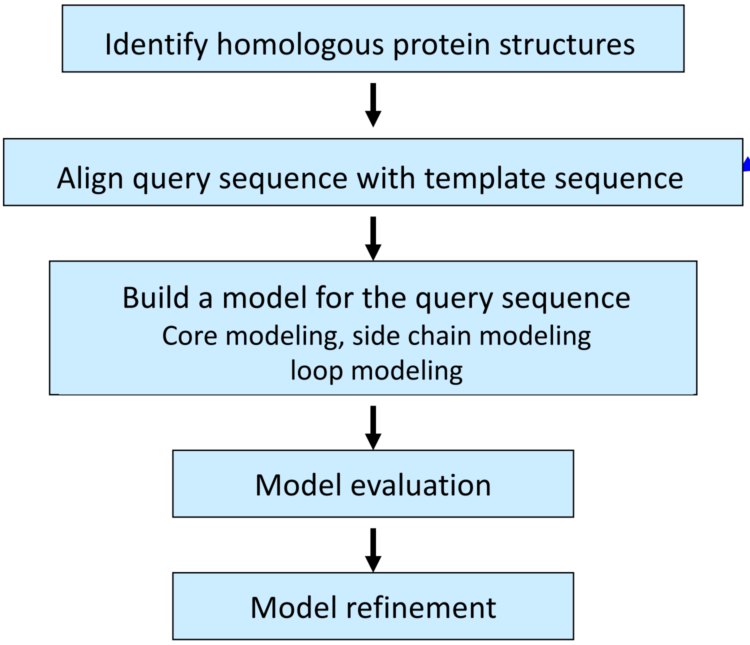


FIGURE 6 Example of accurate template-based modeling for a relatively difficult target, T0868, a bacterial CdiA tRNase toxin. The experimental structure (PDB 5j4a) is shown as a cyan cartoon, with the best homologous template in red, the best server model in green, and the best overall model in blue. There are several obvious areas of improvement over the template, for example modeling of the top left helix, not present in the template, correction of the inter-helical relationship on the top right, and correct replacement of the long template hairpin at the bottom of the structure

Moult et al. *Proteins* 86, 7–15 (2018).

Comparative protein structure modelling (template-based modelling) (homology modelling)

Homology Modeling



Identify homologous protein structures

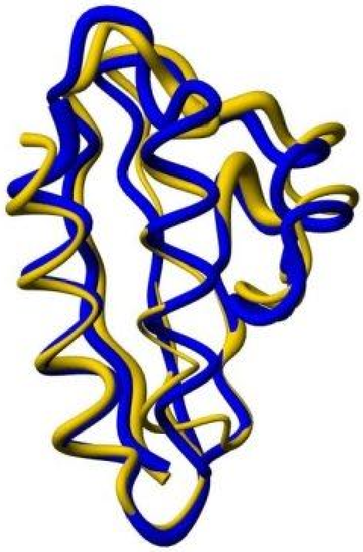
Align query sequence with template sequence

Build a model for the query sequence
Core modeling, side chain modeling
loop modeling

Model evaluation

Model refinement

Very important step



Most of the steps can be automated

HM can give excellent predictions

Threshold for Structural Homology

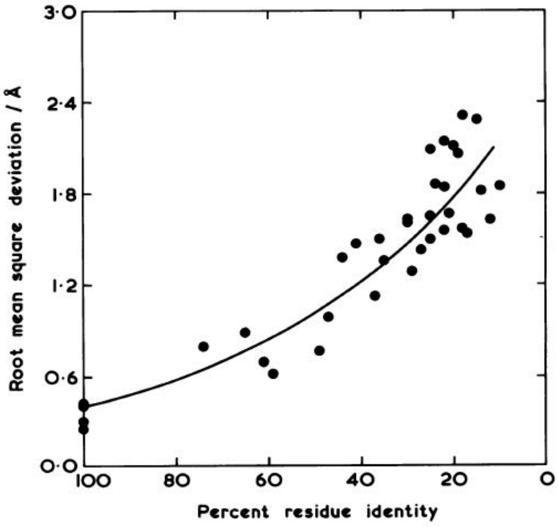
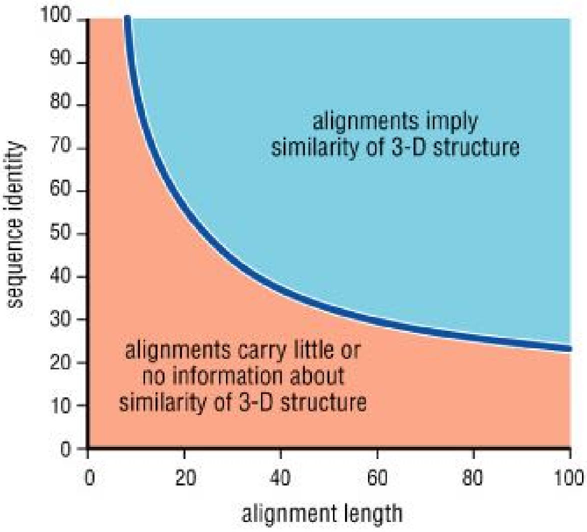


Fig. 2. The relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (see Table II).



From **Protein Structure and Function**
by Gregory A Petsko and Dagmar Ringe

Chameleon Sequences

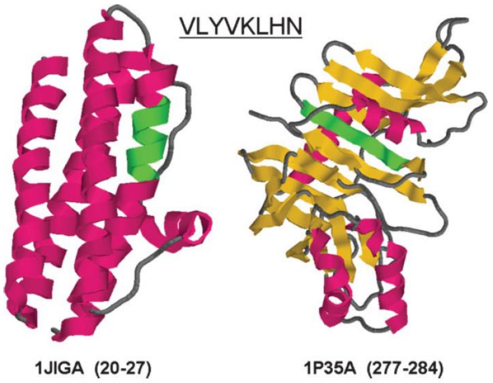


Fig. 1. A chameleon-HS sequence VLYVKLHN (green) in 1JIGA (helix conformation) and in 1P35A (strand conformation). The figure was prepared using Rasmol.[37]

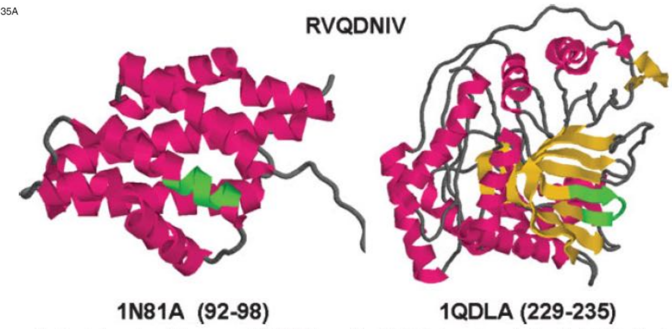
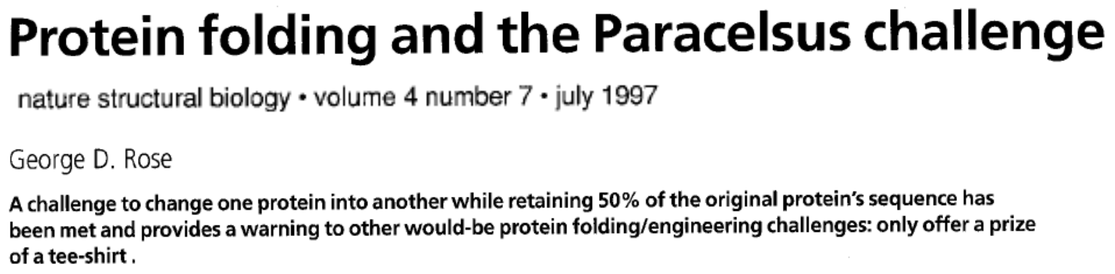


Fig. 2. A chameleon-HE sequence RVQDNIV (green) in 1N81A (helix conformation) and in 1QDLA (sheet conformation). The figure was prepared using Rasmol.[37]

Same short protein sequence adopts different secondary structures

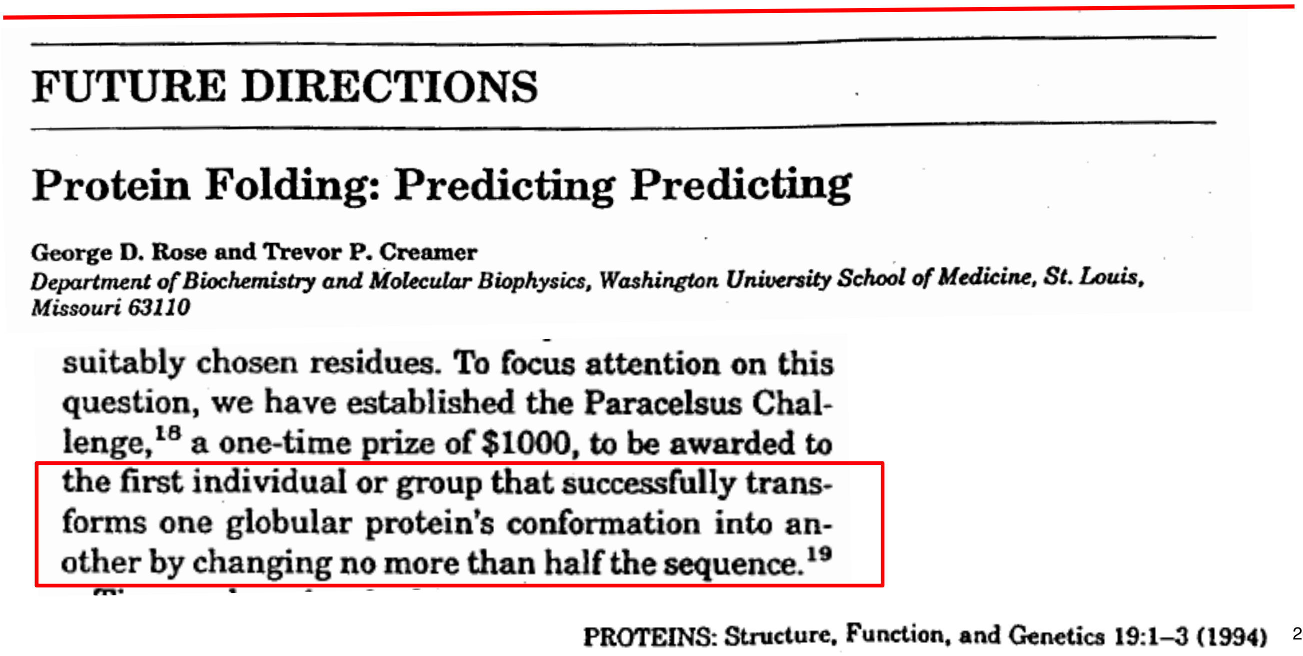


Protein folding and the Paracelsus challenge

nature structural biology • volume 4 number 7 • july 1997

George D. Rose

A challenge to change one protein into another while retaining 50% of the original protein's sequence has been met and provides a warning to other would-be protein folding/engineering challenges: only offer a prize of a tee-shirt.



FUTURE DIRECTIONS

Protein Folding: Predicting Predicting

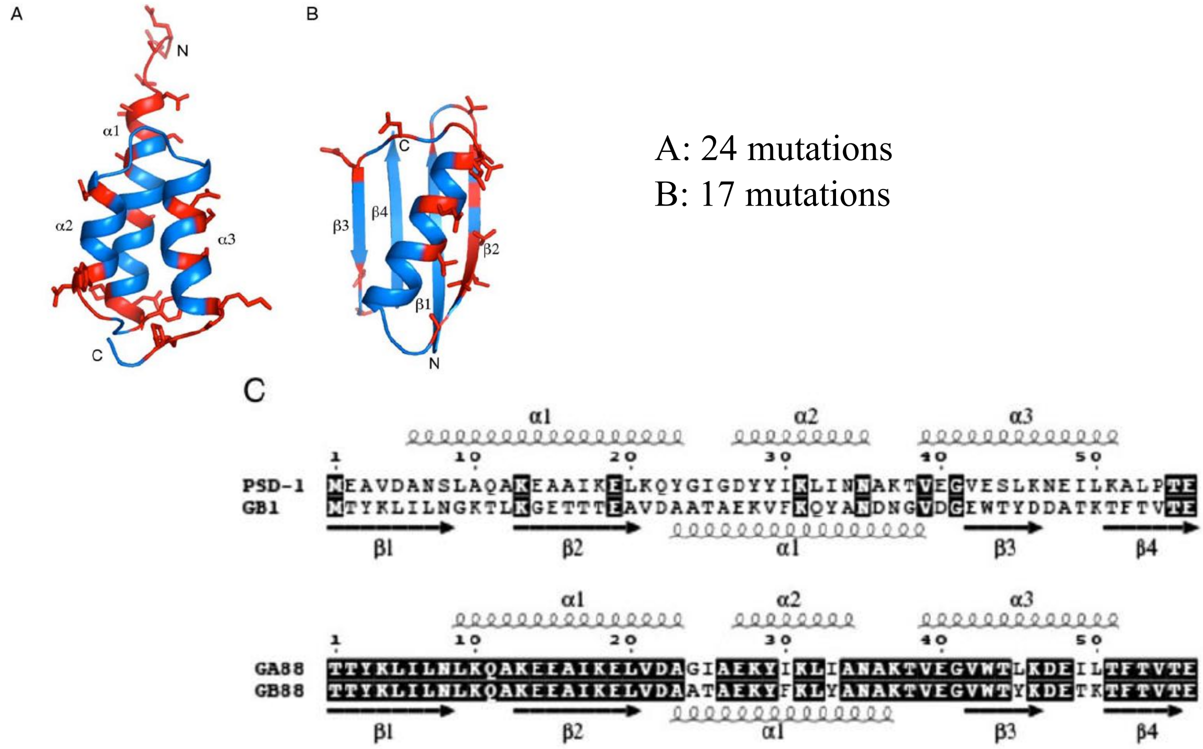
George D. Rose and Trevor P. Creamer
Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110

suitably chosen residues. To focus attention on this question, we have established the Paracelsus Challenge,[18] a one-time prize of $1000, to be awarded to the first individual or group that successfully transforms one globular protein's conformation into another by changing no more than half the sequence.[19]

PROTEINS: Structure, Function, and Genetics 19:1–3 (1994)

NMR structures of two designed proteins with high sequence identity but different fold and function

Yanan He, Yihong Chen, Patrick Alexander, Philip N. Bryan, and John Orban*


A
B
N
α1
α2
α3
C
C
β4
β3
β2
β1
N
C
α1
α2
α3
1
10
20
30
40
50
PSD-1 MEAVDANSLAQAKEAAIKELKQYGIGDYYIKLINNAKTVEGVESLKNEILKALPTE
GB1 MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
β1
β2
α1
β3
β4
α1
α2
α3
1
10
20
30
40
50
GA88 TTYKLILNLKQAKEEAIKELVDAGIAEKYIKLIANAKTVEGVWTLKDEILTFTVTE
GB88 TTYKLILNLKQAKEEAIKELVDAATAEKYFKLYANAKTVEGVWTYKDETKTFTVTE
β1
β2
α1
β3
β4

A: 24 mutations
B: 17 mutations

14412–14417 | PNAS | September 23, 2008 | vol. 105 | no. 38

Target-Template Sequence Alignment

Absolutely Critical:

- Sequence alignment is the bottleneck of the modeling process
- No comparative modeling scheme can recover from an incorrect alignment.

How does one find template(s)?

- The simplest template determination approaches use fairly common database searching methods (i.e., BLAST and FASTA).
- In slightly more difficult cases, multiple sequence alignment and profile-based methods might be used to identify and better align the template to the target sequence.

Target-Template Sequence Alignment

When multiple targets are identified, there are a variety of ways of determining the best — this is a very important step.

Key factors to consider include:

- coverage
- sequence similarity/phylogenetic clustering
- matching of target predicted secondary structure with observed template secondary structure
- structure quality (resolution, R-factor, etc.)
- known functional relationships, etc.

Backbone Model Generation

- For most of the model, creating the backbone structure with a traditional homology modeling protocol is trivial (simply copy the coordinates from one template to the model!). If there is a match within the alignment, the coordinates of the side-chain can be copied as well.

- More recent methods attempt to use multiple structural templates (e.g. if one template has good overlap in one area, while the other has better overlap elsewhere).

Backbone Model Generation

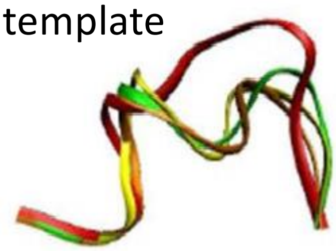
- The program SEGMOD builds the model structure using a hexapeptide fragment library. The model structure is built based on a series of these fragments.

- The widely used program MODELLER generates a series of distance constraints from the template structure, and then builds a model using these restraints in much the same way that is done with NMR structure determination.

One of the advantages of using the satisfaction of spatial restraints method is that it can incorporate various restraints from experiments, such as NMR experiments, site-directed mutagenesis and cross-linking experiments.

Loop Modelling

- Modeling loops that lack coverage within the template is **extremely difficult**, yet **common** due to:
  - Template structure is not well resolved.
  - Sequence divergence
  - Insertions/Deletions



- To make things worse, loop regions vary significantly between model and template even when complete coverage is present.
  - Surface loops tend to be involved in crystal contacts, leading to significant conformational changes dependent upon the unit cell.
  - The exchange of a small to bulky side-chain underneath the loop (within the core) can "push" it aside.
  - Also, remember that loop regions are generally floppy and fluctuate constantly, meaning a fixed conformation may have little biological meaning.

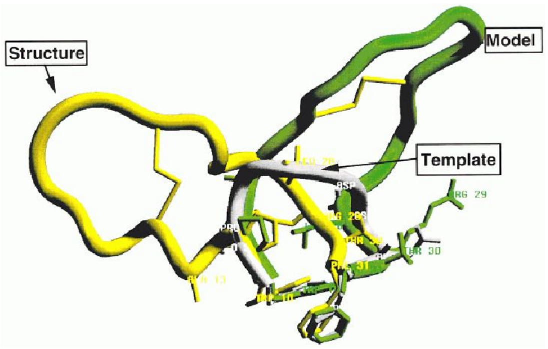
Loop Modeling Methods

Knowledge-based:

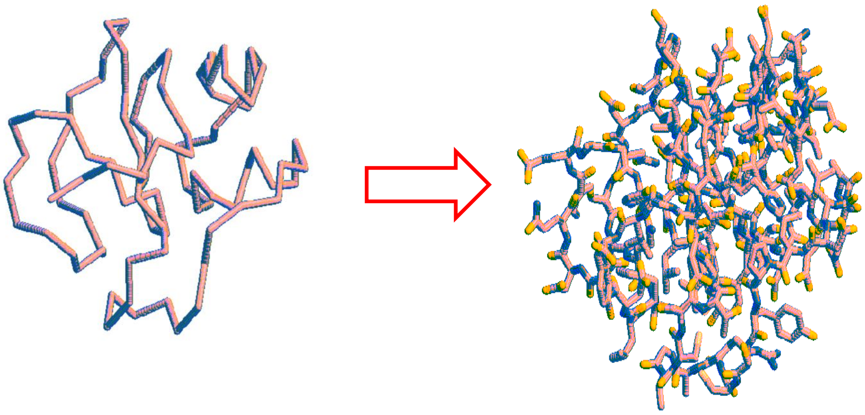
- Find matching loops with the right number of residues and matching endpoints within the PDB.
- In particularly difficult cases (loops longer than ~8 residues), chain fragments together. Based on the premise that irregular substructures are built from combinations of small standard structures.

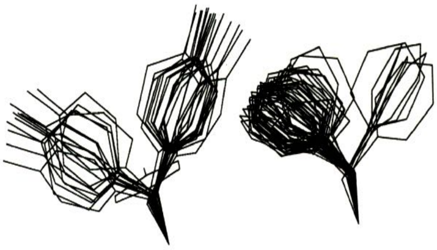
Energy-based:

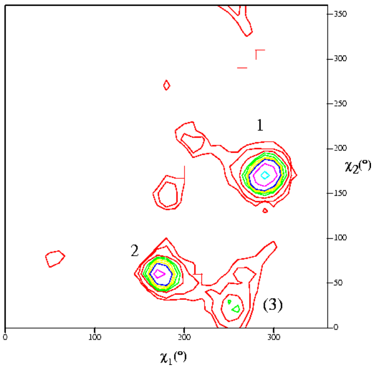
- Generate random loops of right length and endpoints. Evaluate resultant structure with some sort of energy function.



Side-chain Modelling/Packing







Some sort of knowledge-based rotamer library from high-resolution structures is used.

Side-chain Modeling/Packing

Combinatorial explosion:

- Intuitively, it makes sense that the conformation of one residue will affect the conformations of others.
- Fortunately, rotamer space is not limitless.
- Assuming on average 5 rotamers per residue, there are still 5^{100} different combinations to score within a 100 amino acid protein.

Solutions:

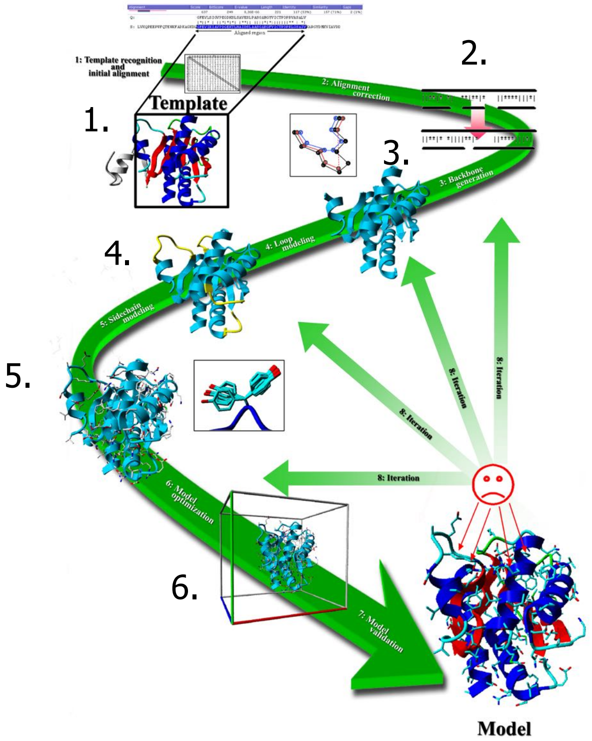
- Certain backbone conformations strongly favor certain rotamers, meaning the others can be ignored.
- More rigid residues can be modeled first, and the more flexible (larger rotamer space) can be modeled subsequently. The advantage of this is that the more rigid residue limits the space that must be explored by the flexible one.
- Nature picks rotamer conformations that maximize packing (minimize voids) and the number of interactions with other groups (i.e. H-bonds, salt bridges, disulfide bonds, etc.).

Model optimization

The last step is to optimize the model using some sort of iterative refinement.

- Unfortunately, current force fields are not sufficient.
- While they will remove the few big errors (bumps), they introduce many small errors.

Summary of the steps



1. Pick a template
2. Refine the sequence alignment
3. Build a model of the protein backbone
4. Model loops
5. Add side-chains

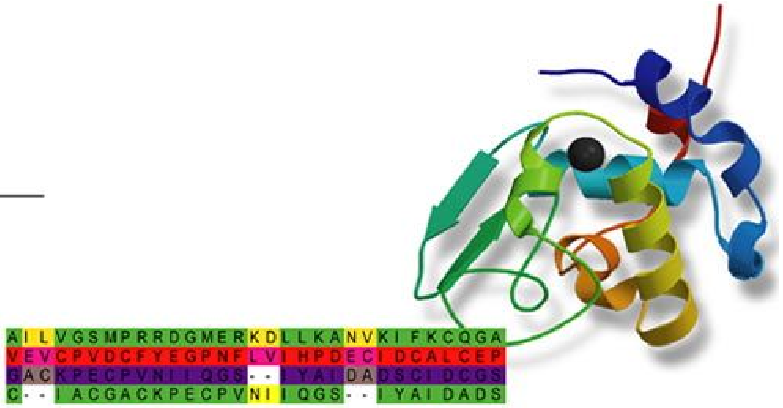
6a. Optimize side-chain configurations

6b. Optimize entire structure

7. Assessment

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints


AILVGSMPRRDGMERKDLLKANVKIFKCQGA
VEVCPVDCFYEGPNFLVIHPDECIDCALCEP
GACKPECPVNIIQGS--IYAIDADSCIDCGS
C--IACGACKPECPVNIIQGS--IYAIDADS

http://salilab.org/modeller/

Given an alignment of a sequence to be modeled with known related structures, MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints 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.

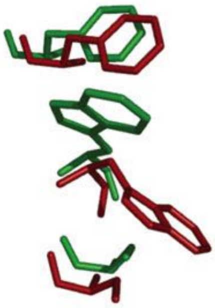
ModWeb: Server for Comparative Protein Structure Modeling using MODELLER
http://modbase.compbio.ucsf.edu/modweb/

SWISS-MODEL

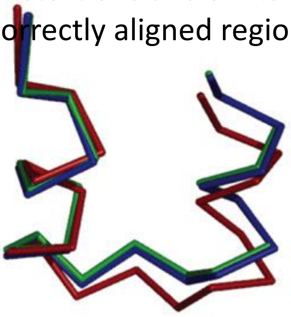
- Swiss-Model - an automated homology modeling server
  http://swissmodel.expasy.org/
- Closely linked to Swiss-PdbViewer, a tool for viewing and manipulating protein structures and models.
- May take hours to get results returned!

Typical errors in comparative modeling

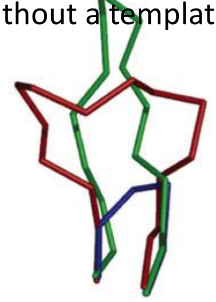
Errors in side-chain packing



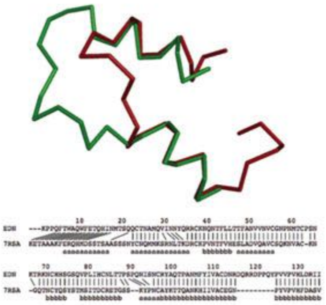
Distortions and shifts in correctly aligned regions



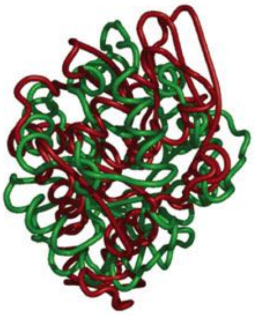
Errors in regions without a template



Errors due to misalignments



Errors due to an incorrect template



Conclusions on homology modeling

- Homology modeling focuses on the use of a structural template derived from known structures to build an all-atom model of the protein.
- Can give **good** overall (fold level) results.
- Yet, the models are **often not good enough** for detailed structure/function analyses.
- In fact, the models tend to look a lot like their templates, meaning a key challenge is picking the right template.
- Detecting meaningful sequence homology in the *Twilight Zone* is very difficult (if not impossible).

Methods for protein structure prediction

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

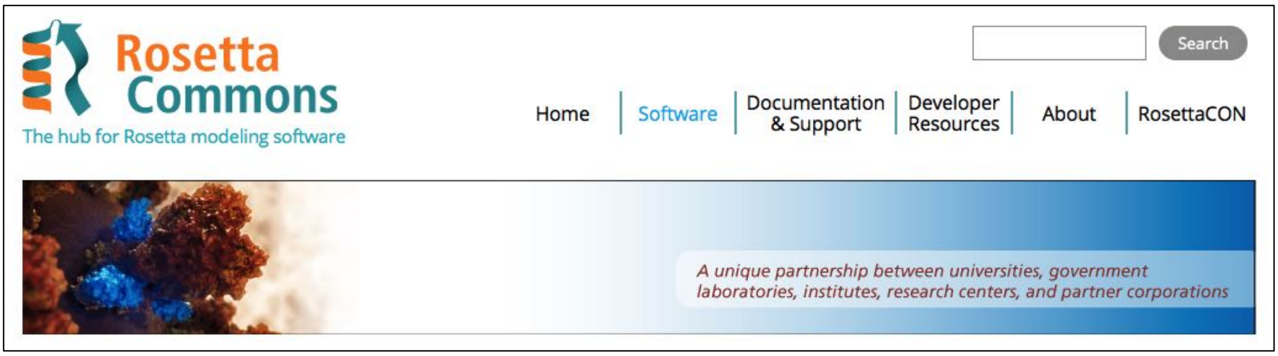
- **Comparative modelling**: A clear evolutionary relationship between the target and a protein of known structure can be easily detected from the sequence.
- **Fold recognition:** The structure of the target turns out to be related to that of a protein of known structure although the relationship is 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.

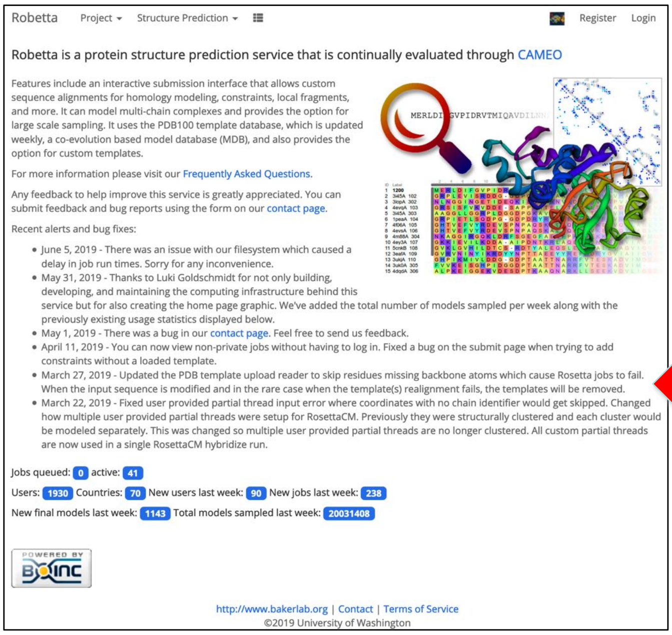
Toward High-Resolution de Novo Structure Prediction for Small Proteins

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)


Rosetta Commons
The hub for Rosetta modeling software
Search
Home
Software
Documentation & Support
Developer Resources
About
RosettaCON
A unique partnership between universities, government laboratories, institutes, research centers, and partner corporations



Robetta is a protein structure prediction service that is continually evaluated through CAMEO

Features include an interactive submission interface that allows custom sequence alignments for homology modeling, constraints, local fragments, and more. It can model multi-chain complexes and provides the option for large scale sampling. It uses the PDB100 template database, which is updated weekly, a co-evolution based model database (MDB), and also provides the option for custom templates.

For more information please visit our Frequently Asked Questions.

Any feedback to help improve this service is greatly appreciated. You can submit feedback and bug reports using the form on our contact page.

Recent alerts and bug fixes:

- June 5, 2019 - There was an issue with our filesystem which caused a delay in job run times. Sorry for any inconvenience.
- May 31, 2019 - Thanks to Luki Goldschmidt for not only building, developing, and maintaining the computing infrastructure behind this service but for also creating the home page graphic. We've added the total number of models sampled per week along with the previously existing usage statistics displayed below.
- May 1, 2019 - There was a bug in our contact page. Feel free to send us feedback.
- April 11, 2019 - You can now view non-private jobs without having to log in. Fixed a bug on the submit page when trying to add constraints without a loaded template.
- March 27, 2019 - Updated the PDB template upload reader to skip residues missing backbone atoms which cause Rosetta jobs to fail. When the input sequence is modified and in the rare case when the template(s) realignment fails, the templates will be removed.
- March 22, 2019 - Fixed user provided partial thread input error where coordinates with no chain identifier would get skipped. Changed how multiple user provided partial threads were setup for RosettaCM. Previously they were structurally clustered and each cluster would be modeled separately. This was changed so multiple user provided partial threads are no longer clustered. All custom partial threads are now used in a single RosettaCM hybridize run.

Jobs queued: 0 active: 41

Users: 1930 Countries: 70 New users last week: 90 New jobs last week: 238

New final models last week: 1143 Total models sampled last week: 20031408

POWERED BY BOINC

http://www.bakerlab.org | Contact | Terms of Service
©2019 University of Washington

Software

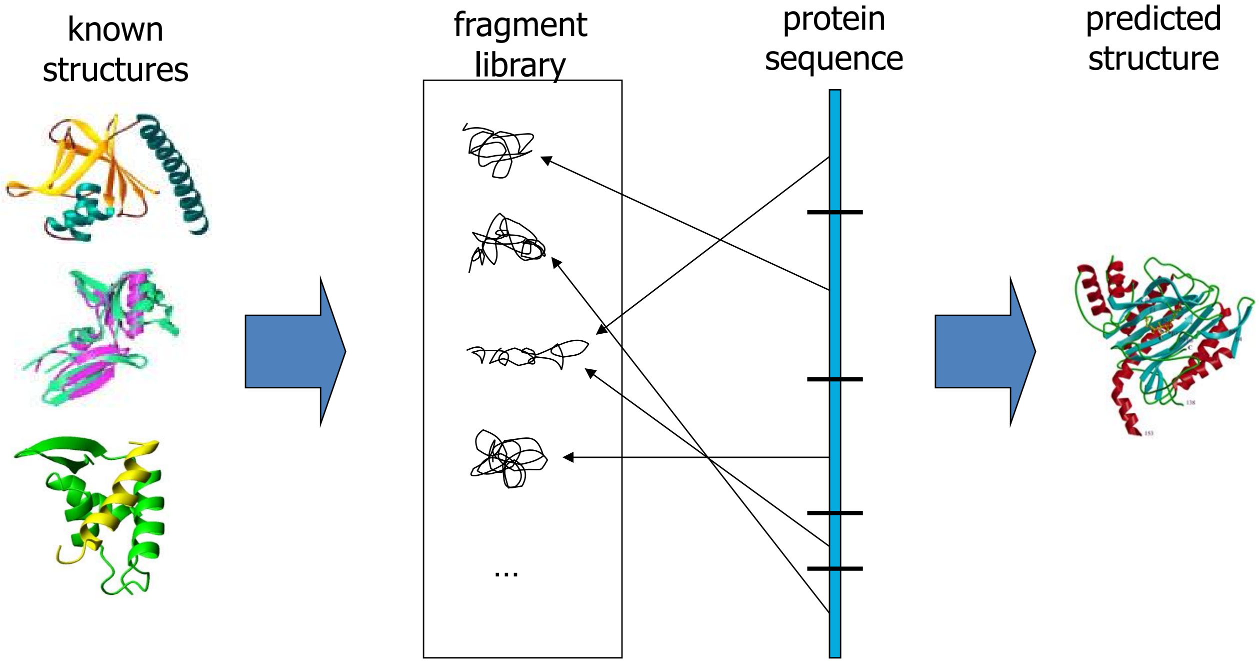
http://www.rosettacommons.org/

Server

http://new.robetta.org/

Rosetta

Assembly of sub-structural units



known structures

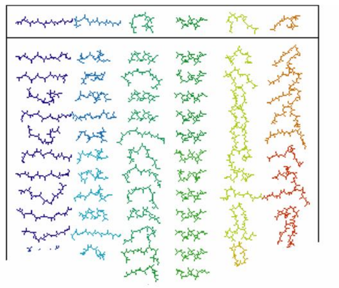
fragment library

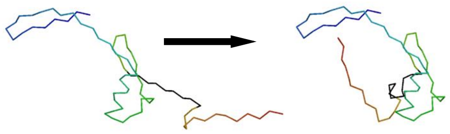
...

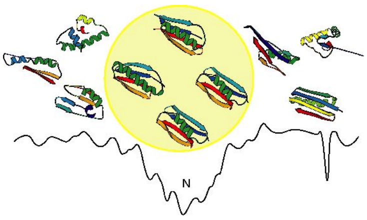
protein sequence

predicted structure

Structure Prediction with Rosetta

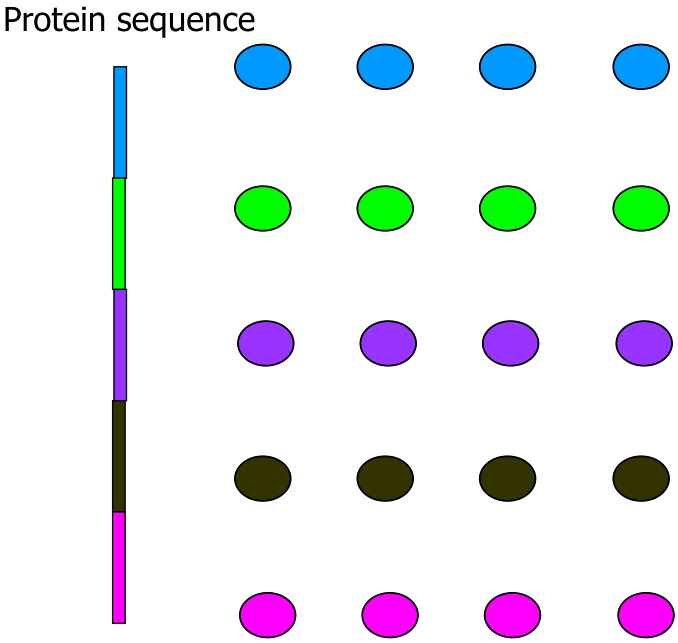






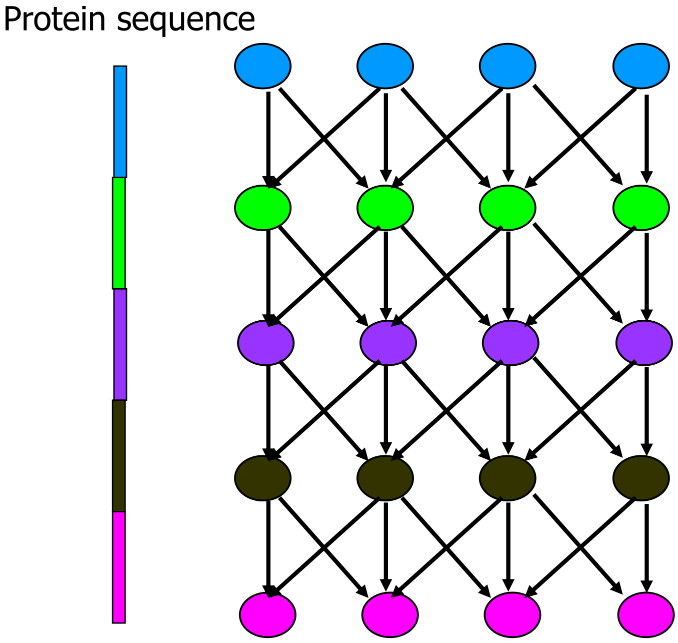
- While not every protein fold is present in the protein databank, all possible conformations of small peptides are.
- Select fragments consistent with local sequence preferences.
- Assemble fragments into models with native-like global properties.
- Identify the best model from the population of decoys.

Modelling



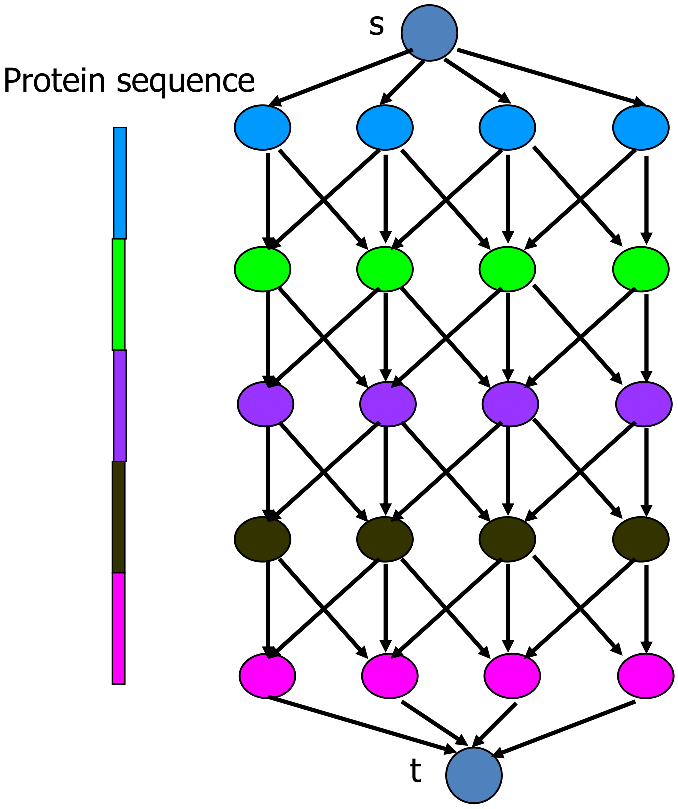
- Model each candidate local structure as a node

Modelling



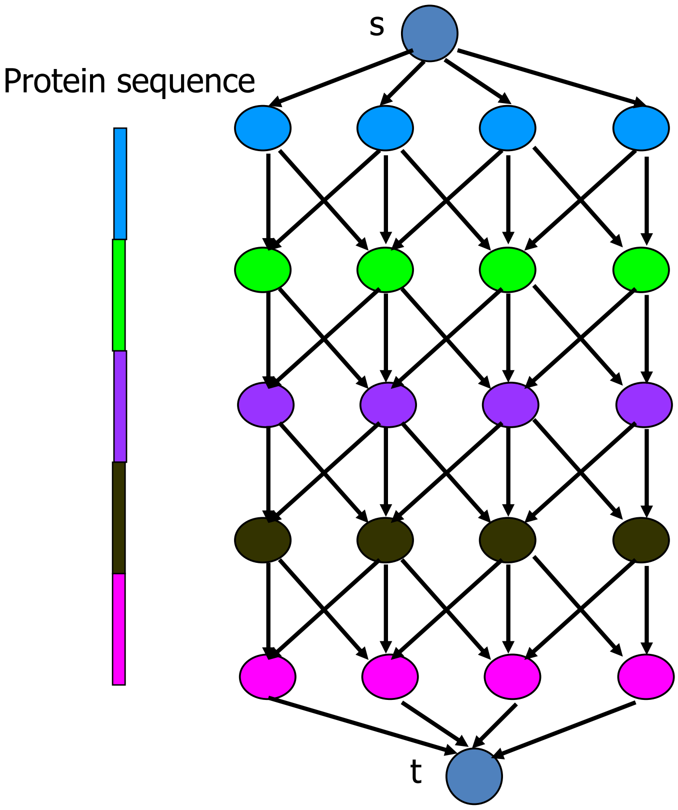
- Model each candidate local structure as a node
- If two consecutive local structure are compatible, an edge joins them

Modelling



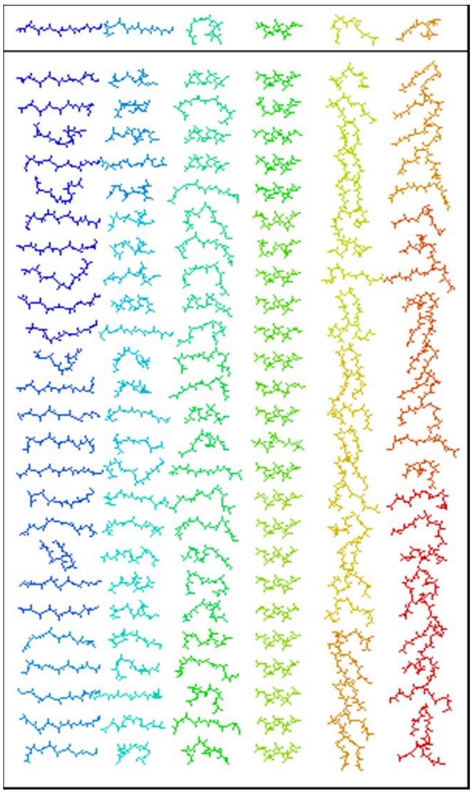
- Model each candidate local structure as a node
- If two consecutive local structure are compatible, an edge joins them
- Add a source s and sink t to the graph

Modelling



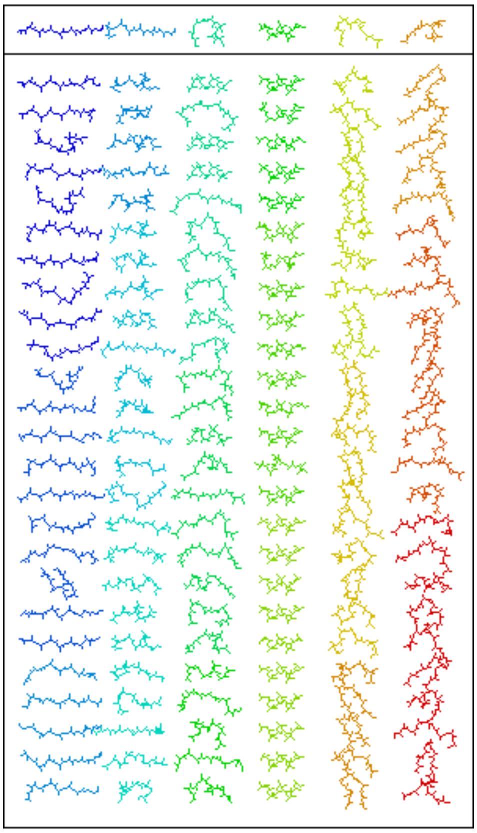
- Model each candidate local structure as a node
- If two consecutive local structure are compatible, an edge joins them
- Add a source s and sink t to the graph
- Each path from s to t forms a candidate structure

Local Sequence Bias – Rapid Approximation of Local Interactions



- While not every protein fold is present in the protein databank, all possible conformations of small peptides are!
- Approximate local interactions using the distribution of conformations seen for similar sequences in known protein structures
- For each sequence window, select fragments that represent the conformations sampled during folding

Rosetta Fragment Libraries



- 25–200 fragments for each 3 and 9 residue sequence window
- Selected from database of known structures
  > 2.5 Å resolution
  < 50 % sequence identity
- Ranked by sequence similarity and similarity of predicted and known secondary structure

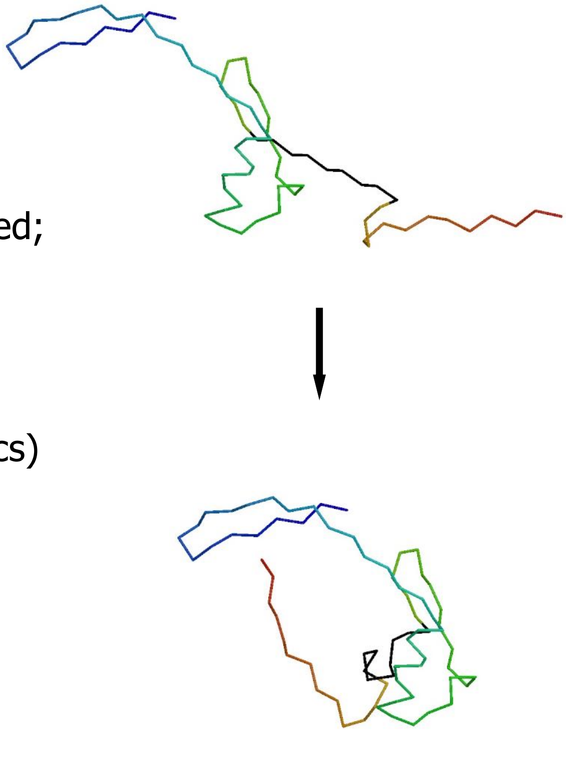
Scoring Function

The ideal energy function

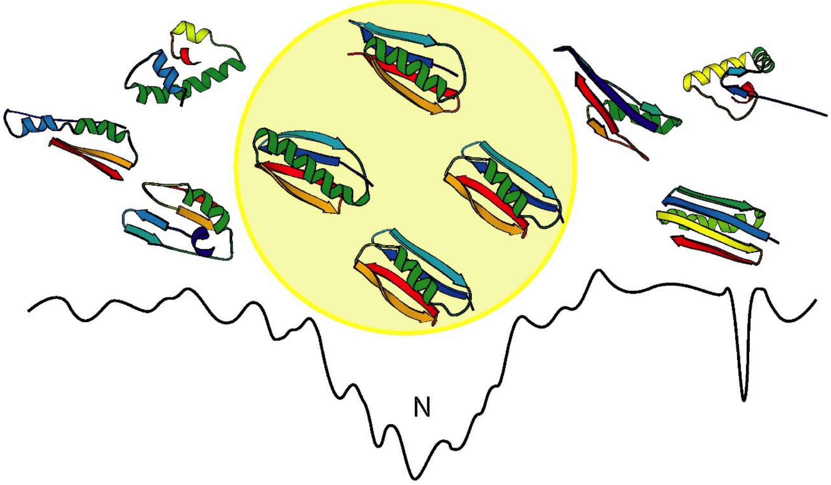
- has a clear minimum in the native structure
- has a clear path towards the minimum
- Global optimization algorithm should find the native structure.

Rosetta Potential Function

- Derived from Bayesian treatment of residue distributions in known protein structures
- Reduced representation of protein used; one centroid per sidechain
- Potential Terms:
  - environment (solvation)
  - pairwise interactions (electrostatics)
  - strand pairing
  - radius of gyration
  - Cβ density
  - steric overlap

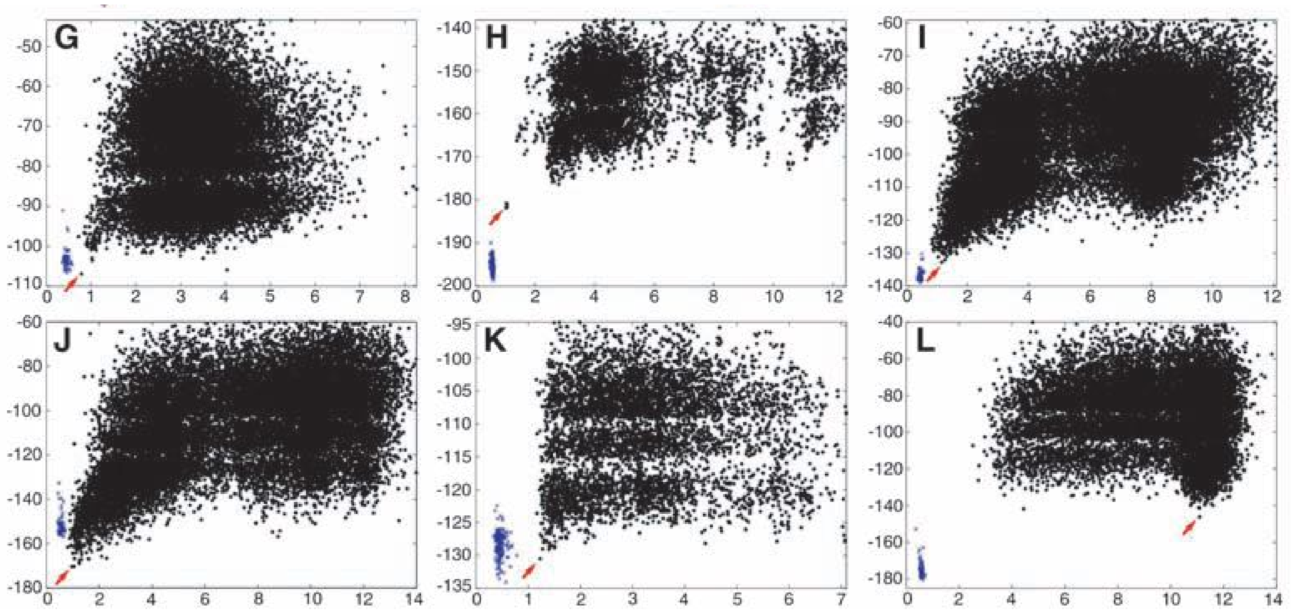


Decoy Discrimination: Identifying the Best Structure



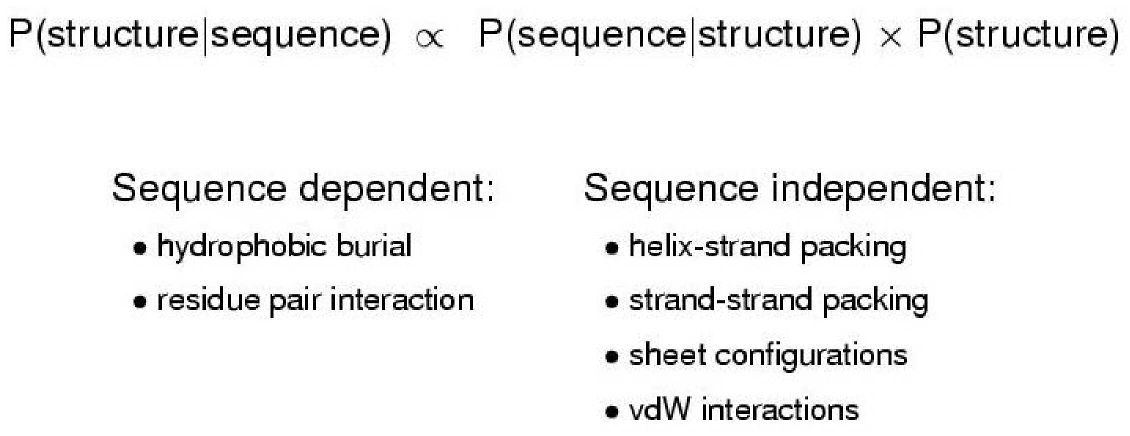
- 1000–100,000 short simulations to generate a population of 'decoys'
- Filter population to correct systematic biases
- Full atom potential functions to select the deepest energy minimum
- Cluster analysis to select the broadest minimum
- Structure-structure matches to database of known structures

Rosetta: Energy vs. Accuracy



Plots of C^a-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.

The Rosetta Scoring Function



$$P(\text{structure}|\text{sequence}) \propto P(\text{sequence}|\text{structure}) \times P(\text{structure})$$

Sequence dependent:

- hydrophobic burial
- residue pair interaction

Sequence independent:

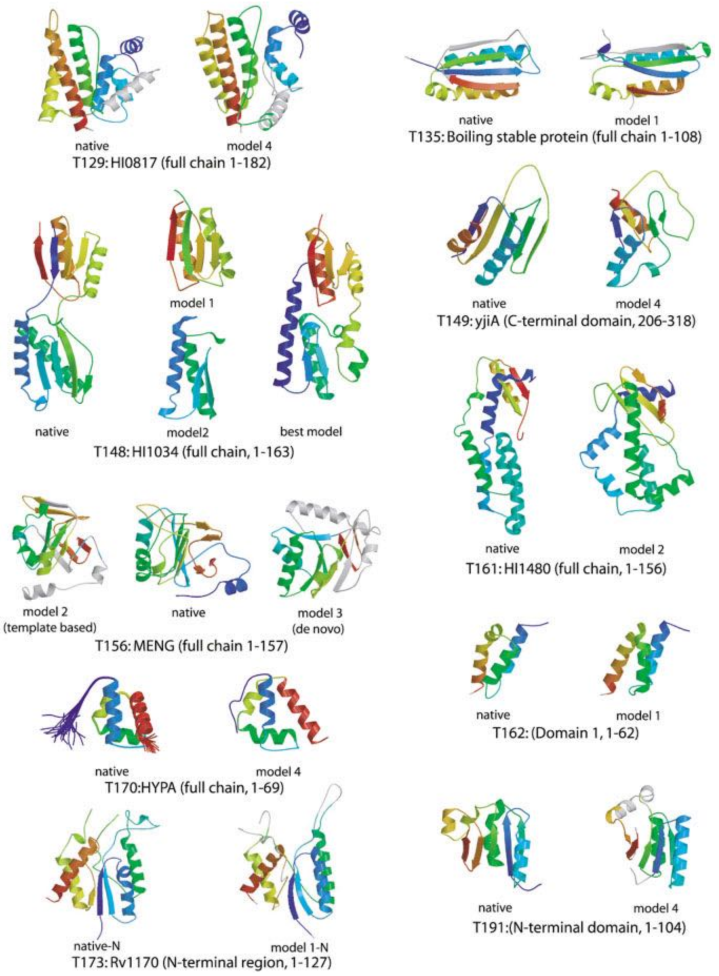
- helix-strand packing
- strand-strand packing
- sheet configurations
- vdW interactions

ROSETTA search algorithm Monte Carlo/Simulated Annealing

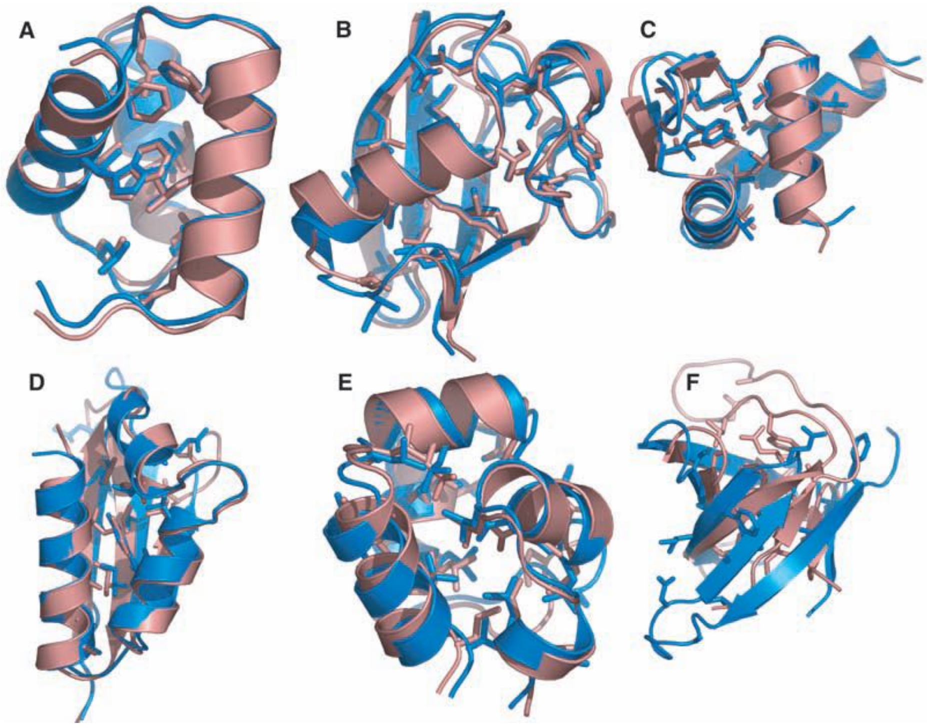
- Structures are assembled from fragments:
  - – Begin with a fully extended chain
  - – Randomly replace the conformation of one 9 residue segment with the conformation of one of its neighbors in the library
  - – Evaluate the move: Accept or reject based on an energy function
  - – Make another random move, taboo list is built to forbid some local minimums
  - – After a prescribed number of cycles, switch to 3-residue fragment moves

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

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