# CASP4 Abstracts

**Ab-initio category**
**(new fold, secondary structure, residue-residue contact)**

# Jones-Ab , 281

number of submitted models: 88

**FRAGFOLD: Ab initio predictions made by assembling fragments of supersecondary structure**

David T. Jones

*Brunel University*
*email:* David.Jones@brunel.ac.uk

For CASP4 targets which we could not predict using fold recognition methods, our FRAGFOLD (D.T. Jones, PROTEINS. Suppl. 1, 185-191) method was used to generate up to 5 ab initio structures. This approach to protein tertiary structure prediction is based on the assembly of recognized supersecondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm.

For all targets (including CM and FR targets), secondary structures were predicted using PSIPRED (D.T. Jones, J. Mol. Biol. 292, 195-202, 1999). PSIPRED predictions in CASP4 (as opposed to CAFASP2) were generated with a database updated at the CASP deadline rather than the CAFASP2 deadline and with alpha/beta structure weights set to

In detail, our 3-D submissions were calculated using the following procedure:

1. Selection of fragment library. At each sequence position a list of 10 supersecondary structural fragments is generated. These fragments (alpha hairpins, beta hairpins, alpha corners, beta corners, beta-alpha units, beta-alpha-beta units and split-beta-alpha-beta units) are taken from 200 highly resolved protein chains with no chain breaks. The selection process involves ranking the fragments in order of potential energy Z-scores (an ungapped alignment is used for this ranking), though excluding any fragments which had a secondary structure different to the PSIPRED secondary structure prediction (only where the PSIPRED confidence was >= 0.8).

2. Simulation. A classic Metropolis scheme is employed in running the simulation. Random moves are made by selecting either one of the preselected 10 fragments at a randomly chosen sequence position, or a free

choice is made from all 3-5 residue fragments from the entire fold library. These moves are first tested to ensure than the generated structure is physically possible (steric checks) and then accepted if the Metropolis criterion is met. The starting temperature for the simulation is selected by making 500 random moves to the starting conformation and calculating the largest absolute energy change between any two moves. The simulation is started at a temperature corresponding to 10 times this delta-E, and the temperature is halved after either 5000 random moves have been accepted by the Metropolis criterion, or a total of 50000 sterically allowable moves have been tested.

3. Potentials. The THREADER V3.0 potentials were used. These are distance-dependent potentials of mean force compiled from a non-redundant set of protein chains with resolutions < 2.6 Angstroms. For the single cyclic peptide target a simple harmonic bond constraint was added to draw together the N and C-termini. Predicted secondary structure was not incorporated in the objective function. Energies were summed over homologous sequences found by running the target sequence through one iteration of PSI-BLAST. In addition to the mean force terms, simple terms were added to take account of hydrogen bonding in beta-sheets, chain compactness and steric hindrance (tables of minimum CA-CA, CA-CB and CB-CB distances were used).

4. Final model selection. 20 separate simulations were run in parallel with different random seed values on a farm of 12 dual-CPU Linux machines. The 20 final structures were clustered using the NMRCLUST program (L.A. Kelley et al., Protein Engineering, 9, 1063-1065, 1996) and the representatives of the highest ranked clusters were submitted (up to the CASP maximum of 5) as final predictions.

---

# SHESTOPALOV , 027

number of submitted models: 152

**Doublet Code of Protein Secondary Structure and its Application for Secondary Structure Prediction and Fold Recognition**

Shestopalov Boris V

*Institute of Cytology of Russian Academy of Sciences*
*email:* shest@mail.cytspb.rssi.ru

The problem of the protein three-dimensional structure prediction has not yet resolved. We propose to resolve this problem using the Linderstrom-Lang hierarchial model of the protein three-dimensional structure formation [1].

The first step of this process is the secondary structure formation. Then the local folds are formed - the supersecondary structure stage. The final stage is the tertiary structure formation.

We state that all the information on these stages is coded and contained in the previous levels of structure. The protein secondary structure code for water-soluble proteins is now determined (the preliminary versions are described in [2] and [3], some modifications are done for CASP4). The code is doublet one. The alpha-helices are coded by the amino acid residue pairs (i, i+4), the beta-structures - by the pairs (i, i+2), the coil regions - by the pairs (i, i+1). The code is overlapping one and the overlapping is resolved by the selection rule aiming to keep the most number of codons after selection.

During the CASP4 experiment the protein secondary structure code has been used for the secondary structure prediction and the protein fold recognition. For secondary structure prediction the homologous sequences information has been used. The homologous sequences were searched by BLAST2 [4](EMBL service), PSI-BLAST and Conserved Domain Database [5] , NCBI service and PRODOM [6]. The most deverse subset has been used. Then the predicted secondary structure has been confronted with the secondary structures from Protein Data Bank in search of similar sequences of secondary structure elements. The results obtained have been used for the fold recognition. In the case of helix-turn-helix motif our method has been used [7]. In some cases, when it was possible and useful, the expert considerations have been used. After the fold recognition the secondary structure prediction is corrected using the secondary structure alignment of the predicted secondary structure and secondary structures for proteins from PDB, recognized as most similar to the predicted protein. Alignment has been constructed manually, using, when it is possible, Yale structural alignments for PARENT and its homologues [8].

Evidently the result of the fold recognition depends on the quality of the secondary structure prediction and the final secondary structure prediction depends on the quality of the fold recognition. The main restriction of the method used is the application of the doublet code of the secondary structure based on the middle interactions only, excluding long ones. The use of the homologous sequence information, as it is known, does not garantee correct secondary structure prediction as well as the use of the fold recognition results.

We hope to correct this situation after the completion of our theory of the protein three-dimensional structure, now in development. Then all the formal and expert ad hoc schemes developed for CASP4 will became unnecessary. We hope that in future it will be possible to construct the code tables using only pure physical considerations without statistical analysis of PDB data as now.

Five models for secondary structure prediction have been constructed. MODEL A is single sequence prediction (SSP), obtained by DOUBLET CODE METHOD as model 3 in CASP3 [3], with slightly modified code tables, MODEL B is obtained from MODEL A by transforming ambiguous and undetermined regions into COIL. MODEL C is multiple sequence prediction (MSP), obtained by application of DOUBLET CODE and PSI-BLAST, MODEL D is variant of MODEL C, obtained using Prodom, MODEL E is MSP with more expert intervention, including the using of the fold recognition results, described above. MODEL 1 is MODEL E, if absent - MODEL D, if absent - MODEL C, if absent - MODEL B. For the fold recognition it has been used MODEL D, if absent - MODEL C, if absent - MODEL B.

References


1. Linderstrom-Lang K.V. (1952) Proteins and enzymes, Stanford Univ. Press, Stanford, California.
2. Shestopalov B.V. Prediction of protein secondary structure by doublet code method. Mol. Biol., Moscow, Engl. transl.,24/4, p.900-907.
3. Shestopalov B.V.,CASP3, submitted.
4. Yan P. Yuan, Eulenstein, O., Vingron, M. & Bork, P. 1998. Towards detection of orthologues in sequence databases. Bioinformatics, 14, 285-289
5. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z, Miller W & Lipman D.J. /1997/. Nucl. Acid. Res. v. 25, pp 3389-3402.
6. http://www.toulouse.inra.fr/prodom.html.
7. Shestopalov B.V. Amino-acid sequence template useful for alpha-helix-turn-alpha-helix prediction, FEBS Lett. 233: (1) 105-108 JUN 6 1988.
8. Krebs W., Gerstein M. http://bioinfo.mbb.yale.edu/align/server.cgi.

# PSSP , 510

number of submitted models: 32

**Protein Secondary Structure Prediction Using Nearest Neighbor and Neural Network Approach**

G P S Raghava

*Institute of Microbial Technology*
*email:* raghava@imtech.ernet.in

Author developed a method for predicting secondary structure of protein from its amino acid sequence. This method is based on artificial intelligence techniques (AI) called i) Nearest Neighbor approach; and ii) Neural network approach. In our approach four steps are in involved in secondary structure prediction. In first step, secondary structure (SS) is predicted using nearest neighbor method. In second step, SS is predicted using neural network. In third step, SS is predicted using previous two steps based on probability of correct prediction. In fourth and final step, predicted SS is refined using structure to structure prediction approach. This method allows to predict SS from single sequence as well as from multiple alignment. Detail description of these steps is given below.

1. Nearest Neighbor Method
   ----------

The basic idea of nearest neighbor methods is to use the examples closely related to the test instance to determine the secondary structure of the test instance. The success of prediction of this approach is directly depend upon the closely related known examples corresponding to a test instance. The nearest neighbor methods outperform the neural network if there are similar or identical examples corresponding to a test instance but perform poorly in absence of closely related examples. The number of known examples is increasing with time, because the protein structure is continuously increasing.

Author optimizes the parameters used in nearest neighbor method to improve the accuracy of prediction of standard nearest neighbor method. In past the database of known examples was limited because these examples were generated from limited set of proteins (around 126). To overcome this limitation, author generates a database of known examples from all proteins in PDB (PDB, 1998), which maximize the number of examples in database. This improves the accuracy of prediction because performance of nearest neighbor method is directly proportional to number of know examples. Author also estimates the probability of correct prediction of each amino acid as well as distribution over the three states.

One of the major drawback of nearest neighbor method is its speed. This method is very slow and time taken to predict SS of a protein is directly proportional to number of known examples, because it compares query pattern with all known patterns. This method becomes nearly impractical when we consider all known examples in PDB. To overcome this problem, we modified the standard nearest neighbor approach. In this approach closely related patterns were searched by comparing query patterns to only those known patterns, who have three central residues (One central and two adjacent residues) identical to query pattern. This makes our modified approach nearly 800 (20*20*20) times faster than standard approach. It was observed that this decreases the performance of method marginally.

2. Neural Network Method
   ----------

In this study we implement standard feedforward neural network method to predict SS of protein. Our network consists 75 hidden units and an input window of 17 amino acids. We trained network on all protein in PDB. To overcome problem of overfitting, divided our data in two sets, where one was used for training and other for checking overfitting. We also compute the probability of correct prediction based on score.

3. Combined Prediction
---

Performance of two approaches depends upon known examples. If known examples are there than nearest neighbor can outperform neural network but in absence of known example its performance is poor. Predicted SS from these methods were combined to get the best accuracy, which was based on probability of correct prediction.

4. Structure to Structure Prediction
---

Structure to structure prediction approach has been used to refined the predicted secondary structure. In this approach, authors develop a neural network method for predicting secondary structure of a residue from predicted SS. This help in producing more realistic results by, for instance, suppressing helix and strands of length one.

One of the challenge in area of Bioinformatics is to provide wide accessibility to the methods /techniques /information in field of biosciences. Author developed a dynamic web server allow user to predict secondary structure of their proteins from its amino acid sequence, based on the above approach. The common gateway interface (CGI )script of this dynamic web server is written in programming language PERL. This web sever is accessible via Internet:

http://imtech.chd.nic.in/raghava/pssp or
http://mail.imtech.res.in/raghava/pssp/ or
http://www.imtech.res.in/raghava/pssp/

---

# Torda-Andrew , 065

number of submitted models: 93

**Secondary structure prediction from threading results**

Abraham, M, Ayers, D, Dosztanyi, Z, Huber, T,Procter, JB, Russell, AJ, Torda, AE

*Australian National University*
*email:* Andrew.Torda@anu.edu.au

Alignments were calculated and models ranked using the sausage program [1]. Sidechains were fitted using a self-consistent mean-field method [2].

Three force fields were used in three different steps

1. Sequence to structure alignments used a score function which used the identity of only one interaction partner [5]. This allowed us to use the Gotoh method [4] for speed,

while avoiding the frozen approximation or double dynamic programming.

2. Ranking of models used a z-score optimised force field [3]

3. Fed by unbounded optimism or perhaps pure faith, side-chains were placed on the models using a more conventional, physically based, molecular mechanics style force field.

The first two force fields may be knowledge-based, but they were built in complete ignorance of Boltzmann statistics. Instead, the parameters are optimised so as to distinguish native coordinates from a mass of misfolded structures.

A second series of optimisation calculations allowed us to find weights for additional terms for secondary structure predictions [6], sequence similarity and gap penalties.

Finally, the library of templates consisted not of simple protein coordinates, but rather of precalculated fields due to averaging over similar structures.

The alignment code and methodology is undisputably fast. It may occasionally be correct.

For the last few targets, secondary structure predictions were made using a neural net fed on the sausage alignment calculations.

---

[1] Huber T, Russell AJ, Ayers D, Torda AE (1999) Bioinformatics, 15, 1064-1065.
Sausage: protein threading with flexible force fields.

[2] Huber T, Torda AE, van Gunsteren WF (1996), Biopolymers, 39, 103-114.
Optimization methods for conformational sampling using a Boltzmann-weighted mean field approach.

[3] Huber, T and Torda, AE (1999) Protein Sci, 7, 142-149.
Protein fold recognition without Boltzmann statistics or explicit physical basis.

[4] Gotoh, O. (1982) J Mol Biol, 162, 705-708.
An improved algorithm for matching biological sequences.

[5] Huber T, Torda AE (1998) J Comput Chem, 15, 1455-1467.
Protein sequence threading, the alignment problem, and a two-step strategy.

[6] Rost B and Sander C. (1993) J Mol Biol, 232, 584-599.
Prediction of protein secondary structure at better than 70% accuracy.

# Raghava-GPS , 018

number of submitted models: 124

**Pediction of protein tertiary structure based on dihedral angles and secondary structure prediction**

G P S Raghava

*Institute of Microbial Technology*
*email:* raghava@imtech.ernet.in

A method is developed to compute tertiary structure of a protein from its amino acid sequence. This method uses the predicted dihedral angles and secondary structure to build the model. Following steps are involved in this approach i) dihedral angles of protein backbone is predicted using method PDAP developed during this study; i) secondary structure is predicted using standard secondary structure prediction method, to refined the dihedral angles; iii) backbone dependent rotamer library is used to predict dihedral angles of side-chains; iv) tertiary structure is build from dihedral angles of protein backbone and side-chains. v) structure was refined using PROCHECK by removing the unusual bond lengths and angles. Following is the description of procedure

## Prediction of dihedral angles

Author developed a method to predict the dihedral angles of protein backbone from its amino acid sequence. The Protein backbone conformation is divided in seven states where each conformation state represents the allowed conformations of isolated peptide as described by Rooman et al. (1991) (J. Mol. Biol. 221:961-79). Each state can be represent by single value of the dihedral angles (phi,psi,omega). Author developed a seven state prediction method using neural network, which allows to predict, conformation state of residues in protein. The Neural Network used in this study is a 75 hidden units, feedforward network with backpropagation learning. The protein backbone dihedral angles were assigned for each residue from predicted conformation state of residue.

## Refinement of dihedral angles

In second step, secondary structure of residues in protein was predicted using PSSP. The backbone dihedral angles of residues whose secondary structure was predicted as helix and strand by PSSP were substituted by new dihedral angles (--65.30,-39.40) and (-117.00, 142.00) respectively. Residues who were assigned as coils by PSSP were kept unchanged (as assigned in step 1).

## Dihedral angles of side-chain

The dihedral angles of side-chains of each residue were assigned using BBDEP program. This program uses the backbone dependent library to assign dihedral angles of side-chains.

## Building tertiary model from dihedral angle

The dihedral angle information of backbone and side-chain predicted by above procedure was used to build tertiary structure. A computer program was

developed which uses the standard bond length and angles and predicted dihedral angles to generate the tertiary structure of protein in Cartesian coordinate.

Final Refinement by PROCHECK
---

To refined the tertiary structure of protein, we applied PROCHECK validation suite. We clean the structure by taking resolution criteria 2.5 angstrom. New file generated by PROCHECK was used as final structure.

---

# TUDELFT , 155

number of submitted models: 44

**MOLECULAR DYNAMICS SIMULATION OF HYDROPHOBIC COLLAPSING FROM A MODEL IN EXTENDED STATE**

Jaap A. Flohil, Simon W. de Leeuw,

*Delft University of Technology*
*email:* j.a.flohil@tn.tudelft.nl

Protein synthesis is a complex, multistep process with main stages of initiation, elongation and termination (Wimberly et al, 2000; Nature Vol 407 pp. 327-339). This process is mimicked by a protocol in which the early events of folding are simulated by gradually releasing a target polypeptide in extended state from its restraints. An initial simulation model was created by mutation of a fully extended polyglycine into the target sequence. GROMACS (Berendsen et al, 1995; Comp. Phys. Comm. 95 pp. 43-56) with GROMOS96 force field (improved alkane dihedrals) was used for all simulations, and among applied parameters were periodic boundary conditions, temperature coupling and long range electrostatics. For each target, the system was initially built up by a shell of explicit water of 0.6 nm added around the stretched model, and the system was placed in a 12x12x150 nm box. Before each of the following simulations an energy minimization was performed. After adding water or renewing the water, a 10 ps run with position restraints on the protein was done to equillibrate the water. A main collapsing run of 1 ns was performed. This run started with all amino acid positions harmonically restrained, releasing each 10 ps a consecutive residue from its restraints until the complete chain was able to remove free. The first residue restraint was released at the N-terminal side, the last restraint at the C-terminal side. For some targets, e.g. T0088, the residues from the second model were released from their restraints in reversed order. In some cases (T0086) no restraints were applied.
If the residue releasing procedure was completed
before the end of the simulation, then the remaining time was used to continue without restraints. From the 1 ns trajectory, each 1 ps a snapshot was recorded, and for each snapshot the radius of gyration about the principal axes of the protein atoms were computed.
During the collapsing run, most polypeptides folded into a low density globular form. The speed and the efficiency of the collapse depended on the used protocol.
The frame with the conformation in the most compact state
was selected for further refinement. Drifting water molecules having no contact with water shell or protein were removed from the box, and the dimensions of the box were maximally reduced, and empty holes in the box were filled with

water molecules.
For some of the targets, for example T0102, additional information about the structure was used to steer the simulation by external forces. T0102 is a cyclic polypeptide, and the secondary structure was known a priory. The helix structures were formed by applying a distance depended potential on the internal helix hydrogens bonds. The bond formation and ring closure was mimicked by adding a distance dependent potential to the MD hamiltonian. A penalty is added to a linear potential, as long as the distance between the terminal atoms 1 N MET and 70 C TRP exceeds the treshold value of the regular peptide bond distance. During a 3 ns simulation, following after the collapsing run, early formation of secondary structure is visible. The first ns shows a sharp decrease of coil, but an increase in bended structure. Formation of B-bridge and B-sheet was visible in about 10 0f the residues, and remains at constant level during the second ns. Formation of helix like structures occurred very rare. based on the evolution of forming secondary and tertiary segments, as well as backbone-backbone contacts and free energy of the water, the best model conformation was selected for submission. Although a ns simulation is at least 6 orders of magnitude shorter then required to obtain sufficient folding statistics, a rather stable intermediate conformation maybe found and might serve as starting structure for fold recognition or folding acceleration methods.

---

# Prof-server , 385

number of submitted models: 42

**The Prof protein secondary structure prediction server**

Ross D. King Mohammed Ouali

*University of Wales, Aberystwyth*
*email:* rdk@aber.ac.uk

The protein secondary structure prediction program Prof (Ouali & King, 2000) was used to predict all sequences. All predictions were made automatically using the default server (see URL below). Prof predicts structure by cascading multiple neural networks and linear discrimination classifiers. The accuracy of Prof is estimated at 76.7% (using leave-one-out cross-validation) on a non-redundant dataset of 496 non-homologous sequences (obtained from G.J. Barton and J.A. Cuff). This database was especially designed to train and test protein secondary structure prediction methods, and it uses a more stringent definition of homologous sequence than in previous studies. A major strength of Prof is that it can be tuned to discriminate the 3 classes (H, E, C) with an accuracy of up to 78% for beta-strands.

Prof Server: http://www.aber.ac.uk/~phiwww/prof/
Ouali, M., & King, R.D. (2000) Cascaded multiple classifiers for secondary structure prediction. Prot. Sci 9, 1162-1176

---

# Aberystwyth , 214

number of submitted models: 41

**A Combined Sequence Homology and Secondary Structure Prediction Methodology**

Ross D. King Mohammed OualiAndreas Karwath

*University of Wales, Aberystwyth*
*email:* rdk@aber.ac.uk

```
The following combined methodology was applied to predict the secondary
structure of all sequences.

1) For all sequences the HI sever (URL below) was first run to
identify homologous sequences of known 3D structure.  HI is a sequence
homology program which uses machine learning to bootstrap sequence
similarity searches.  Hi learns rules that are true for sequences
which are clearly homologous to the target sequence, and uses these
rules to discriminate homologous sequences in the twilight zone.

2a) If no homologous sequence was detected of known structure then the
Prof server (Ouali & King, 2000: URL below) was used to predict
secondary structure.

2b) If there was a homologous sequence of known structure, then the
sequence of strongest homology was chosen and aligned to the target
and then used as a template to predict secondary structure.

Occasionally the homology predictions of HI were overruled by hand
and Prof used instead.

HI server: http://www.aber.ac.uk/~phiwww/hi/
Prof server: http://www.aber.ac.uk/~phiwww/prof/
Ouali, M., & King, R.D. (2000) Cascaded multiple classifiers for
secondary structure prediction. Prot. Sci 9, 1162-1176
```

---

# PROF/Rost , 402

number of submitted models: 43

**PROF: profile-based neural network prediction**

Burkhard Rost

*Columbia Univ*
*email:* rost@columbia.edu

```
PROF extends PHD in various ways.
First, the neural network system was retrained on a larger
data set of more than 1000 proteins.
```

Second, the profiles input to the networks were generated
with an iterated PSI-BLAST search against SWISS+TREMBL+PDB.
Third, particular aspects were added to the input units
neural networks.

---

# BioInfo.PL , 031

number of submitted models: 97

**ORFeus, protein structure prediction strategies**

Janusz M Bujnicki, Leszek Rychlewski

*International Institute of Molecular and Cell Biology*
*email:* leszek@bioinfo.pl

Several features of a protein can be inferred based on sequence
similarity or assumed homology with other proteins. These include 3D
structure or general functional description. Sequence alignment is the
most common approach used for the assertion of homology. The
predictive power and utility of homology based prediction methods
increases with the continuously growing database of proteins with
annotated structure or functions. Additional increase in predictive
power can be attributed to the improving accuracy and sensitivity of
sequence comparison methods. Consideration of sequence information
deduced from the family of proteins closely related to the query
protein, as performed by PSI-Blast, enabled a dramatic boost in the
predictive power of sequence comparison methods. The approach of
amplification of sequence information by incorporation of evolutionary
related sequence is being pursued further in a bilateral fashion. The
family of BASIC (Bilaterally Amplified Sequence Information
Comparison) programs represents prediction methods utilizing the
evolutionary information on both sides of the comparison: the query
and the template. ORFeus, the current descendant of this approach was
used during CASP4. The main difference between such prediction
methods and threading approaches is the ability the possibility of a
sensitive exploration of the protein sequence space. The target protein
and its family based profile is compared with more than 100,000 other
profiles. This comparison is expanded in an iterative fashion until a
profile of a family with known structural information is found.
In CASP4, the sequence analysis and the final structural predictions
were supported by the utilization of modern threading approaches,
which were carefully benchmarked before (http://bioinfo.pl/livebench).
The Structure Prediction Meta Server was used for this additional
analysis (http://bioinfo.pl/meta).

---

# Shortle , 001

number of submitted models: 29

**Ab Initio Prediction of Protein Structure based on Modeling the Denatured State as an Ensemble of Overlapping Gapped Fragments**

David Shortle and Michael Ackerman

*The Johns Hopkins University School of Medicine*
*email:* shortle@welchlink.welch.jhu.edu

The method attempts to predict protein structure at low resolution by crudely modeling the denatured state as an ensemble of overlapping fragments. The free energy of fragment ensembles is very roughly approximated by weighting both the energies of individual fragments and an ensemble ‡pseudo-entropy† term, which contains the frequency of occurrence of secondary structure, the handedness of connections and cross-overs, proximity between pairs of segments of secondary structure, and more global features of topology. The central working assumption is that the denatured state ensemble is positioned in the widest free energy well in compact conformation space (Shortle et. al, 1998).

In the first step, the sequence of the target protein is divided into fragments (ungapped 15 - 35 residues, gapped 35 -45 residues) that extend from the end of one turn to the beginning of a second turn. Initially, turns are localized by computing turn propensities (Chou and Fasman, 1980), and subsequently their positions are refined from turn locations in low energy fragment ensembles. These fragments of sequence are threaded through a set of 1600 proteins (the FSSP database of Holm and Sander, 1996), with evaluation of both the empirical pair potential of Bryant and Lawrence (1993) and the secondary structure probability from the helix, strand, and turn propensities of Chou and Fasman. Secondary structure is then predicted from profiles of the frequencies of turns, helices and strands displayed by overlapping sets of 10 to 30 lowest energy fragments that do not exceed a specified level of secondary structure improbability.

Using the predicted central residues of helices, strands, and turns, the PDB is searched again for larger gapped fragments (+/- 3 residues in each predicted turn) with the correct secondary structure and the lowest energy. Sets of overlapping fragments were superposed to identify the most common topological relationship between segments of secondary structure inside and outside the region of overlap. For the predictions submitted to CASP4, visual inspection was the principal method used to identify patterns among helices and strands, and then all of these features were put into a consistent schematic model of the overall topology. In this process, the empirical rules described Jane Richardson and others were used to resolve uncertainties. The initial model, usually built manually by docking sets of fragments from more than one protein, contained gaps of +/- 3 residues in all turns; these were patched using the lego_loop function of the O software (Alwyn Jones). In work since the end of CASP4, superpositioning based on distance matrices and rapid coordinate calculations is being used to build gapped models with little or no human intervention.

If the schematic model resembled a known structural class, protein structures taken from this class were used to build a low energy model. Thus for some targets, the method serves for fold identification. In addition, the secondary structure from fragment ensembles can be used to aid alignment to known structural homologues when there are too few homologues to nail down the secondary structure with confidence using methods such as PHD or PSIPRED. Because fragment assembly into a model was based on human intervention, the method was not immune to additional sources of information and input of the pattern-recognition type.

Shortle, D., Simons, K.T. and Baker, D. Proc. Natl. Acad. Sci. USA 95: 11158 (1998).
Bryant, S.H. and Lawrence, C.E. PROTEINS 16:92 (1993).

Fasman, G.D. Prediction of Protein Structure and the Principles of Protein
  Conformation, Plenum Press, (1989), page 282.
Holm, L. and Sander, C. Science 273:595 (1996).

---

# HeadGordon-Teresa , 383

number of submitted models: 21

**Predicting Protein Tertiary Structure using a Global Optimization Algorithm**

E. Eskow, B. Bader, V. Lamberti, R. H. Byrd, and R.B. Schnabel Computer Science Department, University of Colorado at BoulderCampus Box 430, Boulder, CO 80309, USA.andSilvia Crivelli and T. Head-GordonPhysical Biosciences, NERSC, and Life Sciences DivisionsLawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

*Lawrence Berkeley National Laboratory*
*email:* TLHead-Gordon@lbl.gov

The stochastic/perturbation with soft constraints method (henceforth, the SPSC method) for predicting tertiary structure of proteins combines knowledge about secondary structure with a large-scale global optimization method and a full-scale local minimization algorithm (1-4). It uses secondary structure information both to guide the search and to reduce the size of the search tree through the formulation of soft constraints. The SPSC algorithm is composed of two phases.

Phase I locates some good local minimizers through local minimizations with soft constraints. These minimizations are done using a limited-memory BFGS local minimization algorithm varying all the cartesian coordinates of the atoms in the protein. The soft constraints are derived from predictions of secondary structure obtained from neural networks; we primarily relied on the Psi-pred server for these predictions, although we used J-pred predictions for one protein. Given the primary sequence, the neural networks predict whether the secondary structure of each amino acid should be alpha-helix, beta-sheet, or coil, and provides an indicator of the strength of each prediction. Phase I begins with a starting structure that is the minimum closest to the fully extended chain with all the backbone dihedral angles phi=180° and psi=-180°. The local minimizations encourage the formation of alpha-helices and beta-sheets through the use of penalty (reward) functions; the strength of a penalty (reward) function depends on the strength of a neural network prediction. For the case of beta-sheets, i,j pairs were predicted using the algorithm described in H. Zhu & W. Braun, Protein Science (1999) V.8, 326-342, and we used several most probable pair lists. The local minimizers that result from Phase I therefore contain predicted secondary structure but do not contain any correct tertiary structure.

Phase II improves those minimizers through expensive global minimizations in a sub-space of the torsion angles corresponding to amino acids predicted to be coil. The global optimization for finding the optimal dihedral angles of coil residues uses a method

introduced by Rinnooy-Kan. This global optimization approach is one of the few that provides a (weak) theoretical guarantee of finding global optimum. A brute-force search is avoided by *selectively* doing a local minimization based on whether a new proposed start structure lies within a certain distance metric, and whether its energy is lower than other existing structures; if a new start structure lies within the distance metric of another structure, and is higher in energy, it is assumed to lie within an existing basin of attraction, and is rejected from further computational consideration. Because the probabilistic theoretical guarantee is higher for small dimensional problems, the idea is to select a subset of ~6-10 variables from the space of torsion angles predicted to be coil. A small scale global optimization is performed on the selected torsion angles as variables while keeping the rest temporarily fixed at their current values. This algorithm is general in the sense that subspaces of arbitrary dimension can be explored. However, in practice, the amount of work required to reach the theoretical guarantee is prohibitive for large subspaces. The small-scale optimization produces a number of local minimizers in the subspace of torsion angles chosen. A number of those conformations with low energy values are selected for local minimizations in the full variable space. The new minimizers obtained from the local minimizations are merged into the current list, are clustered and ordered by energy value (see below) and the second phase starts again. The process repeats for a number of iterations, until no further progress is made according to a stopping criteria we describe below.

To guarantee that the search through the space of possible solutions is well balanced in breadth and depth of the tree, a heuristic is used to determine which local minimizer to pick from the list of local minimizers for the next round of phase II. In a balancing stage, the work is balanced over a fixed number of trees. A tree consists of an initial minimizer and all the minimizers generated from it by applying a global optimization on a small subspace of torsion angles followed by a local minimization over the entire space. At each iteration of the balancing stage, the tree with the least amount of work performed on its members so far is selected, and the best configuration in this tree that has not already been used is chosen. After the fixed number of iterations of the balancing stage have been performed, the remaining iterations of Phase II correspond to the non-balancing stage. In this stage, the best configuration is selected, regardless of which tree it comes from. We have found that the combination of breadth and depth in the search of the configuration space contributes to the success of this method.

The SPSC algorithm uses the AMBER molecular mechanics energy function, EAMBER, to represent the protein-protein interactions. We have added an empirical solvation free energy term that models the hydrophobic effect as a two-body interaction between all aliphatic carbon centers. This description is motivated by our recent experimental, theoretical and simulation work to determine the role of hydration forces of model protein systems [5-7]. The Esolv interaction potential has two minima: one for the carbons in contact, and one for the carbons seperated by a distance of one molecular layer of water, with a barrier in between. The benefits of this form are that (1) we introduce a stabilizing force for forming hydrophobic cores, (2) it is a well-defined model of the hydrophobic effect, and (3) it can be described as a continuous potential that is computationally tractable. We tested the effect of the solvation energy function on conformations of the 70 amino acid protein uteroglobin (2utg A) and the 72 amino acid DNA binding protein (1pou), initially using parameter values based on the experimental/theoretical work in [5-7]. We found good agreement between misfolded structures and energies values using this form of potential, but agreement was better using parameter values that exaggerate the stability of the contact

and solvent-separated minima. In addition, it proved beneficial to model the screening of electrostatic interactions by scaling with a dielectric constant of four, which is typical of a dielectric for a protein environment. In extensive testing of Esolv using these parameters with three different alpha-helical proteins, we compared the energies of the crystal structures with approximately 40,000 misfolded conformations generated by our global optimization algorithm. These structures had roughly correct secondary structure but incorrect tertiary structure. For each of the proteins the potential function with Esolv correctly gave the crystal structure as the lowest energy relative to the 40,000 misfolds.

The global optimization algorithm runs on the T3E at NERSC, where for CASP4 we used between 64 and 128 processors. At the end of each Phase II run the algorithm returns between 60-80 of the best (lowest energy) configurations found thus far. These 60-80 structures are clustered into groups in which members of a given cluster are within 5-10?r.m.s.d. of each other (lower bound for small proteins (under 100 residues), upper bound for large proteins) and the members of each cluster are energy ranked. Typically we see on the order of 10-20 clusters in early Phase II. The best configuration from each cluster is used as a starting point for the next round of Phase II (whether balancing or un-balancing). Our experience is that a good sign of convergence is the contraction of the number of distinct clusters to one or two, and an energy value which is no longer changing. We always submitted structures that were our lowest in energy as the first prediction, the second prediction generally as the lowest energy of the member of the second lowest energy cluster, etc.

1. S. Crivelli, T. Head-Gordon, R. H. Byrd, E. Eskow, R. Schnabel (1999). A hierarchical approach for parallelization of a global optimization method for protein structure prediction. Lecture Notes in Computer Science, Euro-Par '99, P. Amestoy, P. Berger, M. Dayde, I. Duff, V. Fraysse, L. Giraud, D. ruiz (eds.), pg. 578-585.
2. S. Crivelli, T. M. Philip, R. Byrd, E. Eskow, R. Schnabel, R. C. Yu, T. Head-Gordon (2000). A global optimization strategy for predicting protein tertiary structure: a-helical proteins. Computers & Chemistry 24, 489-497.
3. A. Azmi, R. H. Byrd, E. Eskow, R. Schnabel, S. Crivelli, T. M. Philip, T. Head-Gordon (2000). Predicting protein tertiary structure using a global optimization algorithm with smoothing. Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches, C. A. Floudas and P. M. Pardalos, editors (Kluwer Academic Publishers, Netherlands), 1-18.
4. S. Crivelli & T. Head-Gordon (2000). A Hierarchical Approach for Parallelization of Large Tree Searches. Submitted to J. Parallel & Distributed Computing.
5. A. Pertsemlidis, A. K. Soper, J. M. Sorenson & T. Head-Gordon (1999). Evidence for microscopic, long-range hydration forces for a hydrophobic amino acid. Proc. Natl. Acad. Sci. 96, 481-486.
6. J. M. Sorenson, G. Hura, A, K, Soper, A. Pertsemlidis & T. Head-Gordon (1999). Determining the role of hydration forces in protein folding. Invited Feature Article for J. Phys. Chem. B, 103 5413-5426.
7. G. Hura, J. M. Sorenson, R. M. Glaeser & T. Head-Gordon (1999). Solution x-ray scattering as a probe of hydration-dependent structuring of aqueous solutions. Perspectives in Drug Discovery and Design 17, 97-118.

# Xia , 446

number of submitted models: 35

Yu Xia, Michael Levitt, Ram Samudrala

*Stanford University*
*email:* yuxia@csb.stanford.edu

We present a combined approach to construct low resolution models of protein structure from sequence information alone [1,2]. Starting from protein sequence, we uniformly sample protein conformational space by complete enumeration on a simple lattice and have a pool of up to 10 billion structures. We then use a variety of selection and reconstruction techniques to both reduce the size of candidates and push the distribution of candidate structures to more native-like, till a single structure model of the protein sequence is generated. A detailed description follows:

First, we exhaustively enumerate all possible compact bounded lattice walks on a tetrahedral lattice to capture the overall protein topology [3,4]. The maximum walk length in our approach is 50; hence each vertex can represent more than one residue for a protein of size up to 200 residues. To obtain a model, a lattice walk is threaded with the target protein sequence and the score is evaluated and minimized using a residue-residue contact function. The 20,000 best scoring structures and their mirror images are selected using a simple lattice-based scoring function. At this stage only protein tertiary topology is captured; there is no secondary structure or side chain information in these structures. To build all-atom models, we fit predicted secondary structures to the selected lattice conformations using the consensus of a combination of methods based on the output produced by the CAFASP server. The consensus predictions are fit to the lattice structures using an off-lattice four-state phi/psi model and a sequential build-up algorithm [5]. Side chains were predicted by simply using the most frequently observed rotamer in a database of protein structures [6]. The all-atom conformations are minimised using ENCAD and scored using a combination of scoring functions that hierarchically reduces the total number of conformations output produced to five which are used for the final submissions.

The scoring functions used for the final filtering include an all-atom distance-dependent conditional probability discriminatory function (rapdf) [7], a hydrophobic compactness function (hcf) [2], a simple residue-residue contact function (Shell) [2], a density-scoring function that is based on the distance of a conformation to all its relatives in the conformation pool, a secondary structure based scoring function that evaluates the match between the predicted structure and the secondary structure of a final energy-minimised conformation, and standard physics-based electrostatics and Van der Waals terms.

We have tested our prediction procedure on numerous proteins, including making predictions at CASP3, with reasonably positive results [1,2,3]. Unlike Monte Carlo and molecular dynamics, our approach is largely deterministic and provides a uniform sampling of structure space. The weakest part in our approach is the low-resolution nature of the tetrahedral lattice model: structural

features such as sharp turns cannot be accurately represented. In general, our procedure is more effective for alpha and mixed proteins than beta proteins. Our main change in this effort since CASP3 has been the use of improved scoring functions to achieve the best discrimination. We are thus able to select conformations that are similar even if the secondary structures are predicted completely inaccurately (as in the case of T102, where the consensus secondary structure prediction predicts two sheets instead of helices which we used "as is" since there is no manual intervention in this method).

[1] Samudrala R, Xia Y, Levitt M. Huang ES. Proteins, S3: 194-198, 1999.

[2] Xia Y, Huang ES, Levitt M, Samudrala R. J. Mol. Biol., 300: 171-185, 2000.

[3] Hinds DA, Levitt M. Proc. Natl. Acad. Sci. USA, 89:2536-2540, 1992.

[4] Hinds DA, Levitt M. J. Mol. Biol., 243:668-682, 1994.

[5] Park B, Levitt M. J. Mol. Biol., 249:493-507, 1995.

[6] Samudrala R, Huang ES, Koehl P, Levitt M. Prot. Eng., 7: 453-457, 2000.

[7] Samudrala R, Moult J. J. Mol. Biol., 275:895-916, 1997.

---

# Ram-Samudrala , 028

number of submitted models: 207

**A segment matching and folding algorithm for ab initio structure prediction**

Ram Samudrala and Michael Levitt

*Stanford University*
*email:* ram@csb.stanford.edu

Our general paradigm for predicting structure involves sampling the conformational space (or generating "decoys") such that native-like conformations are observed, and then selecting them using a hierarchical filtering technique using many different scoring functions. Our goal was to devise a method that would combine the best aspects of the more successful ab initio methods at CASP3. There are three stages to our approach, which is completely automated:

1. Generation of secondary structure prediction

The consensus of the secondary structure predictions from the various

servers at the CAFASP meta-server was used as the secondary structure prediction.

## 2. Searching of conformational space

We initially start with an all-atom conformation where residues predicted to be in helix/sheet by the consensus secondary structure prediction are set to idealised helix and sheet values. The remaining phi/psis are set in an extended conformation. Side chain conformations are predicted by simply using the most frequently observed rotamer in a database of protein structures [1]. New conformations are generated by replacing three phi/psi values for three residues with identical sequence which are obtained from a database of known structures. The scoring function used is primarily a combination of an all-atom distance-dependent conditional probability discriminatory function (rapdf) and a hydrophobic compactness function (hcf) [2]. The conformations are modified based on the fragment insertion approach and and the energies are minimised by using two different protocols: a straight-forward monte carlo/simulated annealing approach similar in spirit to that of Baker and colleagues [3], and a conformational space annealing approach developed by Scheraga and colleagues [4]. A combination of minimisation parametres and scoring functions were used to generate a large pool of conformations.

## 3. Selection of conformations

The conformations generated were minimised using ENCAD and scored using a combination of scoring functions that hierarchically reduces the total number of conformations produced to five which are used for the final submissions. The scoring functions used for the final filtering include a simple residue-residue contact function (Shell), a density-scoring function that is based on the distance of a conformation to all its relatives in the conformation pool, a secondary structure based scoring function that evaluates the match between the predicted structure and the secondary structure of a final energy-minimised conformation, and standard physics-based electrostatics and Van der Waals terms.

This work is an attempt at combining three prediction methods from CASP3, the conformational space annealing method by Scheraga [3], the use of fragments to sample conformational space by Baker [4], and using scoring functions of Samudrala et al [1] to not only drive the search method, but also to make the final selections. In addition, there are components that are unique to this approach primarily in the form of the heirarchical filtering methodology employed and in subtle variations of each of the search methods.

[1] Samudrala R, Huang ES, Koehl P, Levitt M. Prot. Eng., 7: 453-457, 2000.

[2] Xia Y, Huang ES, Levitt M, Samudrala R. J. Mol. Biol., 300: 171-185, 2000.

[3] Simons KT, Kooperberg C, Huang E, Baker D. J. Mol. Biol., 268: 209-225, 1997.

[4] Lee J, Liwo A, Scheraga HA. Proc. Natl. Acad. Sci. USA, 96: 2025-2030, 1999

# Avbelj-Franc , 046

number of submitted models: 25

**Prediction of the Three-Dimensional Structure of Proteins**
**Using the Electrostatic Screening Model and Hierarchic Condensation**

F. Avbelj

*National Institute of Chemistry*
*email:* francl@sg3.ki.si

The method for predicting three-dimensional (3D) structure of proteins is based on the electrostatic screening model of amino acid residue backbone conformational preferences (ESM) (1-4). According to the ESM, the stability of backbone conformations is primarily determined by the balance of strengths between local (E-local) and nonlocal (E-nonlocal) main-chain electrostatic interactions. These strengths depend on the amino acid side-chains because the local and nonlocal electrostatic interactions are screened to a different degree by the solvent. The hypothesis is based on an analysis of potentials of mean force obtained from high resolution experimental protein structures (5,1,6-7). The strongest support for the ESM is provided by recent experimental studies, which demonstrated that an enthalpic factor is involved in determining the preferences for alpha-helices and beta-strands (8-9). Further support for the ESM is provided by the NMR studies of denatured proteins (10). The ESM has been used for predicting secondary (2; ESM implemented in the Lifson-Roig theory, Q3 is 69%) and 3D structure of peptides and proteins (6-7). The 3D structure of proteins is predicted from sequence alone. The free energy of a protein as a function of its conformation is obtained from the potentials of mean force analysis of high resolution x-ray protein structures (6,7). The free energy function is simple and contains a small number of fitted coefficients (~50). Various different models for the most critical hydrophobic interactions are used. The minimization of the free energy is performed by the torsion space Monte Carlo simulations (6,7). In the first phase of the minimization procedure only short-range interactions are activated. Short-range interactions are interactions between amino acids less than four residues apart in the sequence. The secondary structures are formed during this initial phase. The long-range interactions are gradually activated in later phases which is causing the hierarchic condensation of alpha-helices and beta-strands into super-secondary and the larger compact structures. Long-range interactions are interactions between the residues distant in the sequence. All heavy atoms including polar hydrogens are included in simulations. The procedure is described in details in references 6 and 7.

References:

1. F. Avbelj and J. Moult, Biochemistry, 34, 755-764 (1995).
2. F. Avbelj and L. Fele, J. Mol. Biol., 279, 665-684 (1998).
3. F. Avbelj, J. Mol. Biol., 300, 1337-1361 (2000).
4. F. Avbelj, P. Luo, R. L. Baldwin, Proc. Natl. Acad. Sci. USA, 97, 10786-10791 (2000).
5. F. Avbelj, Biochemistry, 31, 6290-6297 (1992).
6. F. Avbelj and J. Moult, Proteins: Struc., Funct., Genet., 23, 129-141 (1995).
7. F. Avbelj and L. Fele, Proteins: Struc., Funct., Genet., 31, 74-96 (1998).
8. P. Luo and R. L. Baldwin, Proc. Natl. Acad. Sci. U.S.A., 96, 4930-4935 (1999).
9. M. Lorch, J. M. Mason, R. B. Sessions, A. R. Clarke, Biochemistry, 39, 3480-3485 (2000).
10. M. Hennig, W. Bermel, A. Spencer, C. M. Dobson, L. J. Smith, H. Schwalbe,

J. Mol. Biol., 288, 705-723 (1999).

# bme-mathatiitb , 439

number of submitted models: 5

**Prediction and Refinement Protein Structures by Nonparametric Regression and Heuristic Constraints**

Jyothi S. and Rajani R. Joshi

*Indian Institute of Technology, Bombay.*
*email:* jyothi@math.iitb.ernet.in

We estimate the short and medium range correlations between the primary and tertiary C-alpha distances (dij , |j-i| < 5) using nonparametric discriminant analysis of some features of the primary chain. These features include the global as well as the local measures of the sequence in terms of certain physico-chemical properties of the amino acid residues in the selected segments of the primary chain.

The heuristics derived from these correlations are then used in a nonparametric regression model to estimate the short and medium range C-alpha distance intervals [1]. A large sample of nonhomologous, high resolution proteins of sizes between 75-150 residues, obtained from the PDB, was used as the training sample in model estimation and validation. The model was further validated on a different set of proteins with known structures. The average accuracy of estimation was found to be above 90%.

Long range distance intervals were then estimated from the sequence by imposing certain compactness and globular constraints for a well defined hydrophobic core. A distance geometry program [2] was used to generate the C-alpha trace of the protein using these distance estimates. About 20 to 30 compatible structures were generated which are then ranked on the basis of the number of globular constraint and bump distance threshold violations.

We have used our method to determine the structures of various classes of proteins and have found that the method is well suited (with the Root Mean Square Error ranging between 1A to 3.0A only) for the prediction of local folds. Moreover these local structures are stable and are in good agreement with those computed by the threading methods of [3]. The validation runs on the set of proteins used by [4] have further shown the superiority of our method as compared to DRAGON and XPLOR in terms of accuracy and computational efficiency. For more details of this study see [1].

We have used the method described above to predict the 3d-structures of the CASP4 targets T0097, T0110 and T0091 --- the main criteria for the selection of targets being the number of residues in the chain. The atomic coordinates for these proteins were then obtained from the best ranked C-alpha structure using Maxsprout [5]. We then refine the backbone for good stereochemistry through the minimization of a distance based shape function derived by us. No homology modelling or standard energy function minimization is used in our method.

References:

[1] Jyothi, S. and Joshi, R. R. "Protein Structure Determination by Non-parametric Regression and Knowledge Based Constraints". To be published in Computers and Chemistry

[2] More, J.J. and Zhijun Wu "Distance Geometry optimization for protein structures." (1999) J. Global Optim. 15(3), 219-234.

[3] Sippl, Manfred "Calculation of conformational ensembles from potentials of mean force." (1990) J. Mol. Biol. 213, 859-883.

[4] Aszodi, A., Gradwell, M.J., and Taylor, W.R. "Global fold determination from a small number of distance restraints." (1995) J. Mol. Biol., 251, 308-326.

[5] Holm, L. and Sander, C. "Database algorithm for generating protein protein backbone and side-chain coordinates from C_alpha trace." (1991) J. Mol. Biol. 218, 183-194.

---

# Skolnick-Kolinski , 080

number of submitted models: 142

**Ab initio protein folding using threading based, tertiary restraints**

Andrzej Kolinski, Daisuke Kihara, Marcos Betancourt, Piotr Rotkiewicz, Michal Boniecki, and Jeffrey Skolnick

*Danforth Plant Science Center*
*email:* skolnick@danforthcenter.org

Structure prediction is done using a hierarchical, ab initio folding method that has potentials constructed using predicted tertiary and secondary restraints extracted from weakly significant fragments generated from PROSPECTOR (1), a new threading algorithm, predicted secondary structure, generic protein like potentials, burial terms and protein specific pair potentials. First, using a side chain, center of mass based lattice model (SICHO) of Kolinski and Skolnick (2), initial structures are generated as follows. Using a prediction of secondary structure, gapless threading of structures of comparable size is performed using the matching fractions of the predicted secondary structure to the actual secondary structure of the templates as a scoring function. Fifty lattice chains were built using the 50 best scoring structures as templates. Then, in the preliminary simulation runs, fifty replicas were used. The second iterations used the top 20 (20 lowest energy replicas) as the input pool. The simulation results from the last iteration of the lattice-folding algorithm were subject to a clustering procedure (3) that was also used to make the final fold selection. All folds are locally relaxed using a more detailed off- lattice model comprised of the alpha carbons and a one or two center description of the side chains that depends on the side chain size; this tends to improve the secondary structure of the models. Atomic detail is then added and the resulting structures are reported.

1. J. Skolnick and D. Kihara, Defrosting the frozen approximation: PROSPECTOR: A new approach to threading ,Proteins in press (2000).
2. A. Kolinski and S. A., Assembly of protein structure from sparse experimental data: An efficient Monte Carlo Model ,Proteins 32 475-494 (1998).

3. M. Betancourt and J. Skolnick, Finding the needle in a haystack: Educing protein native folds from ambiguous ab initio folding predictions ,J. Comp. Chem in press (2000).

---

# HenryS , 251

number of submitted models: 43

**THREADNNSSP - Secondary Structure Prediction Combining Local Threading and Salamov/Solovyev NNSSP.**

Henry Schultz and Michael Bass

*Amgen, Inc.*
*email:* hschultz@amgen.com

The Salamov/Solovyev NNSSP Secondary Structure Prediction algorithm ((1995) J.Mol.Biol, 247,11-15) is perhaps the best of the current crop of single method prediction algorithms (King, et al, (2000) Protein Engineering , 13, 15-19). To help correct the known deficiencies of this algorithm we combine local thread predictions with NNSSP to improve overall accuracy.

A reasonable window size is selected and the window is slid across the target protein one AA at a time. Each window is threaded against a 35 on-redundant collection of 3D structures from PDB. The threading Z-scores are used to select windows of optimal prediction and the secondary structure predictions thus obtained are compared to NNSSP results by rules using both the threading information and the NNSSP alpha and beta probabilities. There are further rules for handling multiple overlaps and conflicts. The algorithm generates an initial prediction automatically. Some manual editing may be done afterward to resolve difficult situations using the multiplicity of window thread data. After CASP4, we will develop rules to do this and make the predictor completely automatic.

The algorithm is intended for predicting the secondary structure of novel proteins so no other information about biology, alignments or structural similarity are used.

---

# Rokko , 225

number of submitted models: 30

**Solvent-induced multi-body force field for ab-initio protein structure prediction**

Yoshimi Fujitsuka, Hiroaki Fukunishi, Wenzhen Jin, and Shoji Takada

*Kobe University*
*email:* stakada@kobe-u.ac.jp

We utilize a coarse-grained model of proteins that take into account solvent effects and perform folding simulation of which final structures after quenching to zero temperature are submitted as predicted structures. The most important characteristics of our approach is to combine basically physico-chemical potential energy function with knowledge-based parameter optimization. The latter is base on the energy landscape theory of protein folding.

The protein chain is represented with 4 united atoms per residue; three for backbone atoms and one for side-chain centroid. The side-chain centroid can move among several positions, each of which corresponds to major classified rotamer. All the bond length and bond-bond angles are fixed to the standard values. The interaction potential includes a usual local potential around dihedral angles, the Ramachandran potential, the hydrophobic (HP) interaction, the electrostatic interaction that includes the hydrogen bond (HB), and the van der Waals (vdW) interactions.

One major feature of our model is the hydrogen bonding that depends on local dielectric constant (Takada, Luthey-Schulten, and Wolynes (1999) J. Chem. Phys. Vol.110 11616). Namely, the HB in the protein core is stronger than that at surface. Such a context-dependence is represented in a multi-body functional form. The same kind of dependence is taken into account for other electrostatic interaction (Takada, (2000) Proteins, in press). Physically, this feature is based on solvent and thus we call this interaction model the "solvent induced multi-body force field" (SIMFOLD). Amino acid specificity mainly comes in via the hydrophobic interactions and the vdW interaction between side-chains. The former depends on the nature of a residue, while the latter depends on the nature of a pair of residues.

The model includes many energetic parameters that sensitively affect the prediction power. Thus, we optimize these parameters based on available protein structural database. Here, we are based on the so-called energy landscape theory of protein folding. Parameters are optimized so that the TF/TG for protein libraries are maximized. It is interesting that optimized energy parameters are significantly correlated to experimentally obtained energetic scale for the hydrophobic interaction. This implies potential advantage of combining physico-chemical energy function with knowledge-based parameter optimization.

Another aspect of the model is to use knowledge-based bias. In particular, we rely on the secondary structure prediction (PHD server is used) and introduce an additional bias to the Ramachandran potential. This bias makes the predicted structure much more secondary structure rich. Especially, the beta sheet protein is difficult and is time consuming without this type of bias.

Folding simulation of the standard simulated annealing method is performed from random coil structures. With one time step 0.03ps, folding up to 0.4 microseconds is taken place, during which the system is cooled from 450K to 350K. Based on eye inspection, sometimes we repeat these annealing runs few times for better sampling. Out of several predicted structures, we choose the final structure mostly by the content of secondary structures, compactness of the overall shape, and the total energy scoring.

References

Shoji Takada, Zaida A. Luthey-Schulten, and Peter G. Wolynes, Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer, Journal of Chemical Physics, 110, 11616-11629 (1999)

Shoji Takada, Protein folding simulation with solvent-induced force field: Folding pathway ensemble of three-helix-bundle proteins, PROTEINS, in press.

# CIRB , 392

number of submitted models: 40

**Prediction of contact maps with neural networks and correlated mutations.**

Piero Fariselli, Osvaldo Olmea, Alfonso Valencia and Rita Casadio

*CIRB and Dept. of Biology University of Bologna*
*email:* casadio@alma.unibo.it

Our predictor is based on neural networks which were trained to learn the association rules between the covalent structure of each protein of a selected data base and its contact map. The input coding includes evolutionary information, sequence conservation, correlated mutations, and predicted secondary structures.

The database

For training the network, we use a large set of non-homologous proteins of known 3D structure. The list includes all proteins in the PDB-select list of non sequence-redundant protein structures whose chain was not interrupted and for which alignments with more than 15 sequences were obtained: in total our set includes 173 proteins.

Computing sequence conservation and correlated mutations

Sequence variability was taken from the HSSP database (Sander and Schneider, 1993). In the HSSP definition, variability is 0 when positions in the multiple sequence alignments are completely conserved and it increases proportionally to the number of amino acid changes occurring at that position. Correlated mutations are calculated as previously described (G?el, et.al., 1994). Briefly, a distance array is used to codify each position in the alignment. This position-specific array contains all the residue-residue distances between all the possible pairs of sequences at that position. The correlation value between each pair of positions in the alignment is computed as the correlation of the two arrays for each possible residue pair. Corresponding elements in the arrays contain the distance between the same two sequences in the two positions under comparison. The scoring matrix of McLachlan (McLachlan, 1971) defines the distances between residues. Positions with a percentage of gaps>10% are set at a correlation value of -1 and completely conserved positions are set at a correlation value of 0. The similarity value of gaps is set to a dummy value of 0.

Neural network architecture and input codings

Neural networks have been proven to be one of the most successful methods for prediction of contact maps of proteins (Fariselli and Casadio 1999). In this work we implement a neural network architecture which is similar to that described before. This topology is the best performing one with the problem at hand. A single output neuron codes for contact (output value close to 1) and non contact (output value close to 0). The hidden layer consists of 8 hidden neurons. A new type of input coding was previously introduced (Fariselli and Casadio 1999) and it is also used here. Each residue pair in the protein sequence is coded as an input vector containing 210 elements (20x (20+1)/2), representing all the possible ordered couples of residues (considering that each residue couple and its symmetric are coded in the same way). This is done in order to reduce the number of weight junctions. When single sequence is used, the input neuron coding for the ordered couple of amino acidic residues at positions i and j is set to 1, while the remaining 209 are set to 0. In

order to take into account the sequence neighbours we use a 3-residue long input window, considering both parallel and anti-parallel pairing of the two segments centred at positions i and j, respectively. This procedure requires 1050 (210x5) input neurons. In order to include evolutionary information the binary input is changed in a frequency-based one. This means that for any two positions i and j the frequency of each type of ordered couple obtained from the same sequence in the alignment is computed (Fariselli and Casadio 1999).

The neural network takes into account for each positions in the sequence alignment information derived from sequence conservation and correlated mutations as previously computed (Olmea and Valencia 1997). Moreover the predicted secondary structures are also added as input to the network. The network is trained using the back-propagation algorithm and a balancing procedure (Fariselli et al., 1993) in order to avoid deterioration of the performance due to the large imbalance between contacts (less abundant) and non contacts (more abundant). The weight junctions are randomly initialised in the range [-0.01, 0.01]; the learning rate and the momentum term are set to 0.1 and 0.9, respectively. A sorting procedure based on the network output values is adopted. Contacts are defined as the highest Lp prediction values for a protein of length equal to Lp and are routinely characterised by output activation values greater than 0.75-0.80.

The filtering procedure

To avoid contact overprediction, the predicted pairs are filtered taking into account the amount of contacts that each residue type can make (Olmea and Valencia, 1997). The filtering procedure is based on the occupancy data (or residue-coordination numbers) of each residue. This value is statistically derived from the set of protein structures of the data base and takes into account the secondary structure type and the solvent exposition of each residue. By this, the number of predicted contacts of a residue becomes a function of its structural environment. The occupancy can be therefore considered an estimate of the maximal number of contacts that each residue can make and is used to limit the number of contacts predicted for each residue.


Fariselli P, Compiani M & Casadio R (1993). Predicting secondary structures of membrane proteins with neural networks. Eur Biophys J 22: 41-51.

Fariselli, P. & Casadio, R. (1999). Neural network based predictor of residue contacts in proteins. Prot. Engng., 12, 15-21.

Goebel,U., Sander,C., Schneider,R. & Valencia,A. (1994). Correlated mutations and residue cony?acts in proteins. Proteins, 18, 309-317.

Mclachlan, A. (1971). Test for comparing related aminoacid sequences. J. Mol. Biol., 61, 409-424.

Sander, C. & Schneider, R. (1993). The HSSP data base of protein structure-sequence alignments. Nucleic Acids Res., 21, 3105-3109.

Olmea,O. & Valencia,A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold. Des., 2, S25-32.


---

# osgdj , 221

number of submitted models: 56

**Physical Model Ab Initio Protein Folding**


D.J. Osguthorpe

*University of Bath*
*email:* djosg@mgu.bath.ac.uk

# The Reduced Representation Model and Force Field

## Simplified Geometry Model

The model involves representing the backbone of each residue by one sphere, or 'atom', and the sidechains by up to 3 'atoms'. The side chains of Ala, Val, Ile, Ser, Thr and Pro are represented by 1 sphere, Leu, His, Asp, Glu, Asn, Gln, Cys and Met by two spheres and Phe, Tyr, Trp, Lys and Arg by three spheres. The different number of spheres reflects the anisotropic nature of the average shape of the corresponding side chains. It also enables assigning different characteristics to parts of the side chain of a residue, for example, the side chain of Arg includes a hydrophobic chain and a polar/charged end. Although in this representation many residues have the same number of atoms, they do not lose their unique identity since they have different parameters.

## Simplified Potentials

The potentials required can be split into three major groups, the virtual internal potentials which stabilise the geometry of the protein, secondary structure stabilisation potentials and the global potentials, which deal with the effects of the environment but do not require the environment to be modeled explicitly.

The potential energy function for the model is defined as:

E total = E Internal + E Secondary Structure + E van der Waals + E Global

## Internal Potentials

The values of the parameters were derived by fitting observed distributions of the corresponding internals in experimental structures and by emulating the energy surface calculated using a full atom model.

The internal energy is defined in terms of virtual bond, angles and torsions (or out of plane). A number of functional forms are used, the standard full-atom model harmonic terms, quadratic functions and Gaussian functions plus combinations of these terms. Additionally an out of plane-virtual valence angle cross-term is defined.

E Internal = E V. bond + E V. angle + E V. torsion + E V. oop + E V. oop X V. angle

Virtual angle - virtual angle - virtual torsion angle (theta-theta-phi) cross-terms are defined for dealing with correlations between the two internal valence angles of a torsion angle in the backbone. These are particularly important for turn conformations.

## Secondary structure energy/Backbone Hydrogen bonding Potentials

With the simplified geometry model only C alpha atoms exist for the backbone and yet backbone hydrogen bonding is very important in the stabilisation of the standard secondary structures. However, the standard secondary structures have a fixed and specific set of

distances between the C alpha atoms. Hence the basic approach was to determine the equilibrium distances between C alpha atoms in 3-10, alpha-helices and parallel and anti-parallel beta-sheets and to use Gaussian functions to stabilise these distances.

E Secondary Structure = E Helix + E Sheet

For the beta-sheets it was also necessary to include some vector terms as well to ensure only when the two strands were aligned was the potential strong. Further improvements were necessary to the sheet potentials as from trial folding runs it became clear that additions were needed to remove conformations created that are never seen in real proteins.

It should be noted that in all cases the secondary structure potentials merely stabilise distances that are found, this is not a pre-imposition of secondary structure. The beta-sheet potentials do a full search of all residue pairs to find any that are close enough to form sheets in each energy calculation.

Secondary structure "prediction" energy

This is a new term added since CASP2 which is required in this model as the basic preference of residues for a secondary structure is due to local interactions between the side chain atoms and the backbone atoms which are missing in this model. Ala, Lys, Arg, Glu, Gln, Leu and Met are assigned a helix preference, while Val, Ile, Thr, His, Phe, Tyr, Trp and Cys are assigned a strand preference. As individual residue conformations only affect the virtual valence angle, the overall preference is specifically increased only for contiguous pairs of residues which both prefer the helical conformation or both prefer the strand conformation.

E Secondary Structure Prediction = E Turn + E Strand

Global/Solvation Potentials

The remaining potentials are used to represent the non-bonded interactions of the residues with each other and the interactions with solvent. The fundamental idea behind the solvation potentials was to use fast approximations to the physical forces involved in real protein structures.

Physical Model Solvation Potentials.

In this potential model the physical forces of solvation were included using simple potential models. The main idea was that most protein atoms should not have an attractive interaction with other protein atoms, reflecting the fact that the real interactions with protein atoms would be replaced by solvent interactions if the atom became exposed, hence its overall energy would not change depending on whether it was buried or exposed. However, the atoms should still have excluded volume so a repulsion potential is required at short distances. An offset Lennard-Jones potential is used, where the well-depth is offset to 0 at the Lennard-Jones radius and the energy is set to 0 for distances between atoms greater than the Lennard-Jones radius. This potential is used for most atoms, in particular the C alpha backbone atoms and any atom which does not have a specific Lennard-Jones potential.

E van der Waals = E Offset Lennard-Jones

Physical Model Solvation Potentials - Hydrophobicity

The next effect to consider is the "hydrophobic" effect. I consider this to be associated with two parts, the Van der Waal's potential between atoms (which is attractive) and effects due to the interactions with water. When side-chains are buried in the hydrophobic core of a protein the only interactions available are the standard Van der Waal's interactions, as there is no water present. Hence side-chain atoms of hydrophobic groups were given a standard Lennard-Jones potential with the energy between the same atom types the enthalpy of vapourisation of the most similar hydrocarbon. This would reproduce the energy of the hydrophobic core when hydrophobic side-chains are buried.

This determined the potential between the same side-chain atom types. For dissimilar side-chain atom types an analysis of the distribution of side-chain atoms around an atom in known protein structures showed to a first approximation little difference in preference between the atoms. This distribution is not that which is created by rules such as the geometric mean rules. A function was created which would give such a distribution and this was used to generate the mixed terms for the Lennard-Jones parameters of hydrophobic side chain atoms.

Having accounted for the potential of hydrophobic side-chains when buried and away from water, a potential for hydrophobic atoms when exposed to water is required. It is only this term which I consider to be truly the "hydrophobic" part, in the sense it reflects the effect of hydrophobic groups on water structure. This was done by introducing a hydrophobic sigmoid potential. (The initial folding work of Levitt had used a sigmoid potential for hydrophobic residues.) For CASP4 this sigmoid potential is only used between the aliphatic side-chain atoms.

A final adjustment to the "hydrophobicity" potential was to give certain groups in residues not normally considered hydrophobic a non-zero Lennard-Jones function so that an interaction existed between them and hydrophobic groups. These groups were not included in the sigmoid potential. Such groups were the Ala C beta, the Thr C beta (because of the methyl group), the C beta of the charged amino-acids Asp, Glu, Lys and Arg and Asn and Gln. It also included the C gamma atom of Lys and Arg. Observations of experimental structures and surface accessibility calculations show that these groups are as buried as any of the atoms in the classic hydrophobic side-chains.

E Global = E van der Waals + E hydrophobic sigmoid

Physical Model Solvation Potentials - Electrostatics
An "inverse" Kirkwood-Tanford model is used for electrostatics, in which the interactions between the charged groups are varied according to their local dielectric environment, defined by counting how many non-polar groups are surrounding them, using a sigmoid function as before. To take into account of ionic strength effects, which are assumed to have an affect at large distances between charges but not at short range (as the Debye-Huckel theory on which this aspect is based assumes an averaged ionic atmosphere around each charge which is certainly not true for charges on the surface of a folded protein), a distance dependent dielectric of the distance squared was used. Electrostatic interactions were computed using a distance cubed term and the same term scaled by the sigmoid function of surrounding non-polar groups. Note that the scaled term accounts for salt-bridges automatically, as interactions between charged pairs not surrounded by non-polars (high dielectric) will be weak but strong when surrounded by non-polars (low dielectric).

E Global = E Electrostatic + E scaled Electrostatic

The other feature of electrostatics that needs to be covered is the difficulty of burying charges. It is actually a much stronger rule of proteins that the charged group of charged residues is exposed than that the sidechains of hydrophobic residues are buried. The simple electrostatic explanation for this is the self-energy of a charge which says it requires a lot of energy to move a charge from a high dielectric region into a low dielectric region.

The same sigmoid function counting the number of non-polar groups surrounding a charge is used as before, scaled by a potential constant which gives a positive energy for burying a charge.

E Global = E charge-non-polar sigmoid

Physical Model Solvation Potentials - Scaling

In the low dielectric environment of the folded protein the stability of the backbone-backbone hydrogen bonds is significantly enhanced as these hydrogen bonds are excluded from solvent and a hydrogen bond is essentially an electrostatic interaction. In the unfolded protein the stability of backbone-backbone hydrogen bonds is likely to be similar to that of backbone-water hydrogen bonds, hence there should be no energy stabilising backbone hydrogen bonds. This effect has been included by scaling the backbone hydrogen bond energy term (E Helix and E Sheet) by a sigmoid function counting the number of surrounding non-polar groups.

Folding Simulations - Simulated Annealing procedure

The starting conformation was an all-extended structure using a rigid geometry procedure based on a standard geometry for the RR model. A random Maxwell-Boltzmann distribution was used to assign initial velocities. The initial temperature was set such the average temperature initially was around 340-350K. 84000x2 steps were run before starting cooling. The annealing protocol was first to reduce the total energy by the energy equivalent to 25 degrees of temperature in 84000 steps followed by 84000 steps at constant total energy. This was repeated three times. 84000 steps at constant total energy followed and then the energy was reduced by 12.5K in 84000 steps. The total energy was then continuously reduced by 6.25K in 84000 steps for 30 runs. The reduction was increased to 12.5K for 5 runs followed by 5 runs at a constant temperature of 100K. Final annealing to close to 0K was done in a single run of 84000 steps.

---

# Lomize-Andrei , 002

number of submitted models: 28

**Assembly of protein cores from regular secondary structures:**
**ab initio and fold recognition techniques.**

Andrei L. Lomize, Irina D. Pogozheva, and Henry I. Mosberg

*College of Pharmacy, University of Michigan*

*email:* almz@umich.edu

3D models of protein cores (complexes of several interacting alpha-helices and beta-sheets, excluding nonregular loops) have been generated for 19 CASP4 targets with no detectable sequence homology to proteins of known structure. The partially automatic procedure described below reproduces main blocks of a large software package that is under development in our group to test the validity of the entire approach and its specific parts. The procedure includes the following three steps.

STEP 1. Ab initio prediction of secondary and supersecondary structure using two different methods:
(a) calculation of alpha-helices, alpha-hairpins, and beta-hairpins in hydrophobically collapsed protein using the program Framework [1];
(b) identification of alpha-helices and beta-strands based on hydrophobicity patterns in multiple sequence alignments [2];
Possible beta-sheet topologies and the structural class of the target (beta-sandwich, beta-barrel, beta-helix, beta-prism, different alpha+beta and alpha/beta structures, alpha-superhelix, or alpha-bundle) were suggested based on a qualitative analysis of results produced by both methods.

STEP 2. Fold recognition. The procedure included the following three parts.
(1) Identification of related PDB structures using a library of "supersecondary nuclei" in proteins [3], and the following criteria:
(a) similar secondary structures of the target and template, including number, order, and lengths of alpha-helices and beta-strands, and identical beta-sheet topologies,
(b) similar biological functions;
Twelve of the nineteen targets considered (T0088,T0094,T0098,T0100, T0101,T0102,T0104,T0107,T0108,T0109,T0118, and T0126) satisfied these criteria, and therefore were designated for fold recognition.
(2) Finding optimal alignment of secondary structures in the target and template that maximizes formation of aliphatic, aromatic, and polar clusters and burial of nonpolar side-chains.
(3) Adjustment of side-chain conformers and the spatial positions of entire alpha-helices to improve close packing, burial of nonpolar groups, and hydrogen bonding.

STEP 3. Ab initio assembly of 3D cores from alpha-helices and beta-sheets - for targets that could not be assigned to any known protein fold in STEP 2 (T0091, T0095, T0097, T0105, T0106, T0110, and T0114). The docking of regular secondary structures (using QUANTA and our unpublished software) sought to optimize burial of nonpolar side-chains, segregation of aliphatic, aromatic, and polar groups into separate clusters, close packing, and hydrogen bonding in simultaneously constructed models of several homologous proteins from the target family. Two different assembly strategies were tested for all-alpha-helical domains: stepwise building of the core from gradually growing structures (T0106 and models 2 of T0095 and T0097), and formation of a nearly complete core (models 1 of T0095 and T0097).

[1] A.L.Lomize and H.I. Mosberg (1997) Thermodynamic model of secondary structure for alpha-helical peptides and proteins. Biopolymers, v.42, pp. 239-269
[2] A.L.Lomize, I.D. Pogozheva, and H.I. Mosberg (1999) Prediction of protein structure: the problem of fold multiplicity. Proteins, Suppl.3, pp.199-203
[3] A.L.Lomize, I.D. Pogozheva, and H.I. Mosberg (1999) Protein structure assembly pathways. Protein Sci., v. 8, Suppl.1, p.86

# DChu11 , 325

number of submitted models: 43

**Fully Automatic Protein Structure Prediction Using Evaluation Functions**

David Chu

*Independent Contributor*
*email:* davidchu@attglobal.net

A wide variety of numerical and statistical methods are utilized to determine the most optimal protein conformation for the prediction of protein secondary structure. At the heart of our technology is the ability to predict protein secondary structure using evaluation functions. Each evaluation function is an algorithm that measures the \223goodness\224 of a given protein conformation. Conformations with positive values are good candidates for protein\222s secondary structure. Conversely conformations with negative values are rejected. Since no single evaluation function alone works the best under all circumstances, we have constructed many evaluation functions to meet all the situations we can think of. The evaluation functions take into account all single and residue pair propensity to be within and at boundaries of loop, strand, and helix. Also taken into account is [protein sequence] to [structure pattern] evaluation function obtained from neural network training and multiple sequence alignment of homologous sequences. Charge distribution, super secondary structure patterns, degree of sequence similarity to sequences of known structures from the PDB are all taken into considerations. Our system is fully automatic and operates totally without human intervention; one simply inputs a protein sequence, and it outputs a secondary structure prediction.

# Baldi , 115

number of submitted models: 43

**SSpro, a web server for protein secondary structure prediction based on recurrent neural networks**

Gianluca Pollastri, Pierre Baldi

*University of California, Irvine*
*email:* gpollast@ics.uci.edu

SSpro is a fully automated system for the prediction of protein secondary structure. The system is based on an ensemble of bidirectional recurrent

neural networks (BRNNs) [1, 2]. BRNNs are graphical models that learn from data the transition between an input and an output sequence of variable length. The model is based on two hidden Markov chains, a forward and a backward chain, that transmit information in both directions along the sequence, between the input and the output sequences. Three neural networks model respectively the forward state update, the backward state update and the input and hidden states to output transition. BRNNs are trained in a supervised fashion using the gradient descent algorithm. The error signal is propagated through the model using the BPTS (backpropagation through structure) algorithm [3], an extension of BPTT (backpropagation through time), used in unidirectional recurrent neural networks.

The system is trained on a set of 1180 structures and tested on a set of 126 structures. The test set is the same on which the first version of the server PHD [4] was trained. The training set is extracted from the Protein Data Bank that was online in April 1999. The structures obtained using NMR or with a resolution worse than 2.0 Angstoms are first removed from the set, then an all-against-all redundancy reduction procedure is run using a rigorous Smith-Waterman algorithm with the Pam120 matrix for pairwise alignments, discarding a sequence if it shows more than 25 0dentity to any sequence in the test set. The same threshold holds for each pair of sequences in the test set. A second all-against-all redundancy reduction procedure is then run on the set thus obtained using a threshold of 50equence identity. The target secondary structure assignments are compiled with the program DSSP [5]. We assign to the class Helix the alpha-helix (H) and 310-helix (G) DSSP classes, to Strand the classes extended strand (E) and beta bridge (B), to Coil the other four classes, consistently with the CASP classification. The system takes as input a profile obtained from a multiple alignment of protein sequences. The multiple alignments are compiled with the program BLAST [6], using default parameters. The database of sequences adopted is the NR that was online in October 1999 (roughly 420,000 sequences). No further check or filter is run on the database. Every sequence in the alignment is assigned a weight proportional to the information the sequence carries with respect to the unweighted profile. A weighted profile is then compiled and used as input for the system.
A set of 11 bidirectional recurrent neural networks is trained on the dataset. For details on the implementation, see [1,2]. The networks contain roughly 70,000 adjustable weights, have normalised exponentials on the outputs and are trained using the relative entropy between the target and output distributions.
The final predictions are obtained averaging the network outputs for each residue. A performance of approximately 76.5orrect residue classification is observed on our independent test set (roughly 800n the training set). SSpro is implemented into a web server that can be found online at the address: http://promoter.ics.uci.edu/BRNN-PRED/

[1] P. Baldi, S. Brunak, P. Frasconi, G. Pollastri, and G. Soda. Bidirectional Dynamics for Protein Secondary Structure Prediction, Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99), Stockholm, Sweden (1999).
[2] P. Baldi, S. Brunak, P. Frasconi, G. Pollastri and G. Soda. Exploiting the Past and the Future in Protein Secondary Structure Prediction. Bioinformatics, 15:937-946, (1999).
[3] P. Frasconi, M. Gori, A. Sperduti. A General Framework for Adaptive Processing of Data Structures. IEEE Trans. on Neural Networks, 9, 5:768-786, (1998).
[4] B. Rost, C. Sander. PHD - An automatic mail server for protein secondary structure prediction. Comput. Appl. Biosci., 10(1):53-60, (1994).
[5] W. Kabsch, C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers, 22:2577-2637, 1983.
[6] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25:3389-3402 (1997).

# Noguchi , 453

number of submitted models: 58

**Prediction of Protein Secondary Structure Using the Threading Algorithm and Local Sequence Homology**

Tamotsu Noguchi

*Electrotechnical Laboratory*
*email:* tnoguchi@etl.go.jp

The SSThread [1]: a method for prediction of protein secondary structure using the threading algorithm relied totally on the global aspect of a protein. We have improved the method by introducing weights of the local sequence homology and the 3D-1D compatibility score (new SSThread). The weight of the local sequence homology gives higher priority to secondary structures in some sequence region whose homology score is high in the structural alignment between the query protein and a protein in the library. And the weight of the 3D-1D compatibility score gives higher priority to all secondary structures in the sequence whose 3D-1D compatibility score is high. A structural library for the threading was reconstructed for predictions of CASP4 targets. The library contains 1692 protein chains, which were selected from PDB at July 11, 2000 by using the system for PDB-REPRDB [2]: a database of representative protein chains from the PDB at the PAPIA (PArallel Protein Information Analysis) WWW server (http://www.rwcp.or.jp/papia/). All selected protein chains are of < 30 identity with one another. On the other hand, a target protein: T0116 of length >500 residues was applied a method that the protein was divided into two sequences to overlap 100 residues each other, and the prediction results of each sequence by the new SSThread were jointed at the overlap region.
We predicted the protein secondary structure of CASP4 targets by the new SSthread and the New Joint Method [3] using the five methods: Qian-Sejnowski, SSThread, Ptitsyn-Finkelstein, Nisikawa-Ooi, Gibrat-Garinier-Robson. The New Joint Method is available from the PAPIA WWW server. Finally, the results of secondary structure prediction for each protein were also jointed as giving a priority to the prediction by each method manually according to the 3D-1D compatibility score. The new SSThread takes a priority of the prediction, in case that the target size is longer than 100 residues and a similar protein with a target, whose total compatibility score is better than -2.8, is found in the structural library by the 3D-1D compatibility search. In these cases, we also submitted the structural alignment between a target protein and the highest score protein.

References

[1] Ito, M., Matsuo, Y. and Nishikawa, K. (1997) "Prediction of protein secondary structure using the 3D-1D compatibility algorithm", CABIOS, 13, 415-423.
[2] Noguchi, T., Onizuka, K., Ando, M., Matsuda, H. and Akiyama, Y. (2000) "Quick Selection of Representative Protein Chain Sets Based on Customizable Requirements" Bioinformatics, 16, 520-526.
[3] Nishikawa, K. and Noguchi, T. (1991) "Predicting Protein Secondary Structure Based on Amino Acid Sequence", Methods in Enzymology, 202, 31-44.

# GNM-AB , 487

number of submitted models: 17

**Ab initio predictions by residue-residue contact matrices**

Galaktionov S., Nikiforovich G.V., Marshall G.R.

*Washington University*
*email:* gregory@ibc.wustl.edu

The cornerstone of our procedure is the prediction of the residue-residue contact matrices, A, based on restoring the elements of their eigensystems, namely, the eigenvectors, Y, associated with the three largest eigenvalues. We have shown that on the basis of the first three terms of eigenvalue decomposition, it is possible to make a reasonable reconstruction of the contact matrix and an excellent delineation of the regions of disallowed contacts. Specific properties of these eigensystems are discussed in the poster presentation (Galaktionov S., Nikiforovich G.V., Marshall G.M., 2000, the CASP-4 Meeting). The set of the eigenvalues of a contact matrix depends only on the size of the protein and can be predicted with good accuracy. The elements of the first eigenvector, Y1, are tightly correlated with the coordination numbers for each residue. Therefore, it is possible to obtain estimations of the elements of Y1 from the vector of the coordination numbers.

Combining the first eigenvector, Y1, with the constant part of A (contacts in positions ai,i+1, ai,i+2; standard contacts within alpha-helices and beta-strands; absence of contacts between the ends of longer elements of alpha-helices and beta-strands), it is possible to obtain a first approximation of the contact matrix, Ao, and of the \223non-contact\224 matrix, Aon, consisting of the largest and the smallest elements, respectively, corresponding to the areas with a very high or low probability of contact.

The most important element of the procedure is the evaluation of the second and third eigenvectors, Y2 and Y3. We have shown that for smaller proteins (< 150 residues) the eigenvectors have a \223block\224 structure, consisting of 3 to 5 alternating positive and negative \223blocks\224 of the elements along the sequence. The absolute values of these elements are correlated with the corresponding elements of Y1. Based on this observation, all possible Y2 and Y3 eigenvectors can be generated by systematic combination of all possible block positions; the most feasible eigenvectors can be selected by checking their orthogonality to Y1 and to each other, conformity with Ao and Aon, etc. For targets we have submitted (< 130 residues), only 1 \226 3 pairs of the Y2 and Y3 met the above requirements and were used for prediction of the new contact and non-contact matrices, A1 and A1n. These were used for the iterative refinement of the elements of Y2 and Y3 until the procedure converged.

The computational protocol began with the prediction of elements of regular secondary structure (alpha-helices and beta-strands) by a consensus of statistical methods available at the URL addresses http://insulin.brunel.ac.uk/psiform.html and http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_server.html. (For T0102, the assignment of helical fragments was assumed to be that of Langdon G.M. et al. (1998) J. Biomol. NMR, v.12, pp.173-175.) Coordination numbers for the protein sequence were predicted in general accordance with Rodionov

M.A. & Galaktionov S.G. (1992) Molecular Biology, v.26, pp. 777-83. These were used to predict the contact and non-contact matrices according to the procedure described above. These matrices were further utilized in the build-up procedure, which generated the self-avoiding backbone structures in compliance with the matrices (Nikiforovich G.V., unpublished). For most targets, the resulting structures were too loose, i.e., not satisfying the expected average interresidue C-alpha - C-alpha distance , which has been found to be correlated with the cubic root of number of residues N; the corresponding regression equation being = 4.65N1/3 - 4.42. Therefore, the C-alpha - C-alpha distance matrices calculated for the final backbone structures were corrected by correspondingly \223squeezing\224 their non-constant parts. Routine DG algorithms were used then for reconstructing C-alpha traces from the corrected distance matrices. Finally, the protein backbones were reconstructed via a method involving the optimization of hydrogen bond energies and local geometry penalties, which has been shown to be more robust than existing algorithms when applied to non-ideal C-alpha geometries (Welsh E.A., manuscript in preparation).

---

# Hogue-Feldman , 090

number of submitted models: 22

**Protein Structure Prediction using Trajectory Information**

Howard J Feldman and Christopher W V Hogue

*Samuel Lunenfeld Research Institute, Mount Sinai Hospital*
*email:* feldman@mshri.on.ca

The predictions (T0091, T0097, T0102) were done ab initio, starting with a secondary structure prediction from PHDsec. In the case of T0102 the secondary structure from (1) was used. The prediction of two very long helices in T0091 was suggestive of a coiled coil structure.

Using a map of the conformational space available at each residue, based on the observed conformations of helical, sheet and coil residues in the PDB, we came up with probability distributions for the conformation of each residue based on residue type and biased by 3-state secondary structure prediction, called "trajectory distributions". These were used as input to a modified version of the FOLDTRAJ algorithm (2) which performed a random walk in Ramachandran space rather than alpha carbon space.

Approximately 200,000 structures were generated using the trajectory distributions to build the backbone. Sidechains are placed probabilistically using a backbone dependent rotamer library(3). All residues are chirally and sterically valid, have a minimum of non-hydrogen van der Waal collisions. For T0091 several distance constraints from a known coiled coil structure (1A93) were used to bias the random walk as well.

To account for the cyclic nature of T0102, we observed that a helix was predicted to pass through residues 1 and 70, so rather than generating starting at residue 1, our N-to-C random walk was started at residue 6 (through 70, followed by 1-5) at a loop between two helices. After

generation, those for which the N and C termini were more than a few Angstroms apart were discarded, leaving a few hundred structures. The N and C termini were then "spliced" together in these structures and the residues renumbered properly, the Met being residue 1. No attempt was made to correct the geometry at the splice site, so some minor error is expected in this region.

From the pool of generated structures, various statistics were gathered, and the best models were chosen based on their energy scores(4)(5), radii of gyration, topology, secondary structure content, and whether they actually satisfied the distance constraints in the case of T0091. Energy scores were ignored for T0102 since it is a membrane protein, and the energy scores used were derived for aqueous globular proteins. Only this final step of choosing the "best" models was subjective and non-automated.

REFERENCES


1. Langdon GM, Bruix M, Galvez A, Valdivia E, Maqueda M and Rico M. (1998) Sequence-specific 1H assignment and secondary structure of the bacteriocin AS-48 cyclic peptide. Journal of Biomolecular NMR. 12: 173-175.

2. Feldman HJ and Hogue CWV. (2000) A Fast to Sample Real Protein Conformational Space. Proteins. 39(2): 112-131.

3. Dunbrack RLJ and Karplus M. (1993) Backbone-dependent rotamer library for proteins. Application to sidechain prediction. J. Mol. Biol. 230: 543-574.

4. Zhang C, Vasmatzis G, Cornette JL and DeLisi C. (1997) Determination of Atomic Desolvation Energies From the Structures of Crystallized Proteins. J. Mol. Biol. 267: 707-726.

5. Bryant SH and Lawrence CE. (1993) An Empirical Energy Function for Threading Protein Sequence Through the Folding Motif. Proteins. 16: 92-112.


---

# Harrison-Weber , 058

number of submitted models: 91

**Randomized and Multiple Model Approaches to Homology Modeling and Ab Initio Modeling.**


Ivan Y. Torshin, Irene T. Weber and Robert W. Harrison

*TJU*
*email:* robert.harrison@acm.org


Molecular modeling is a combinatorial, multiple minimum optimization problem. In homology modeling, the known homolog serves as a good starting point for the search, while in ab initio folding there are only limited geometric data. Two complimentary classes of algorithms were explored in our CASP-4 predictions:

randomized algorithms, and multiple modeling algorithms.
Randomized algorithms, either based on the Kohonen self-assembling neural network or an analytic solution for simultaneous circular equations, were used to explore conformational space and delineate regions of allowed molecular geometry. These algorithms are computationally efficient; it was possible to fold most of the CASP-4 ab initio targets several hundred times in a few CPU hours. Multiple models from independent runs of the randomized procedures were used to extract conformations that occurred repeatedly, as this improved the reliability in tests. Hundreds of models were used for ab initio predictions and ten models for homology modeling. AMMP (Harrison, 1999) was used to predict 12 ab initio targets and 30 homology modeling targets.

Randomized Algorithms

Our major focus has been to explore new algorithms for building molecular models and searching conformation space. One general class of algorithms, randomized algorithms, is especially interesting because these algorithms can efficiently find or approximate the solutions to combinatorial and geometric problems (Hertz 1991, de Berg 1997) and can be implemented efficiently on a parallel computer (JaJa 1992). The general idea behind randomized algorithms is to use a set of independent identically distributed random variables to limit the solution to an acceptably small range. Rather than attempt to converge to an ŧexact╡ solution of a mathematical problem, which may not exist or may not be meaningful in the context of protein structure, randomized approaches define a sequence of ever-closer bounds on the ranges of solutions. Two randomized approaches were tested, these were a modified Kohonen neural network with a distance metric and a randomized analytic solution to distance restraints (Harrison 1999). Multiple models were constructed using the distance restraints that were derived from homologous structures or sequences. Averages over the models were then used to develop a single model for submission.

Homology Modeling

Protein folds were recognized using the FFAS server (Rychlewski et al. 2000), the 3D-PSSM server (Kelley et al. 2000), and the screening method we used for ab initio folding. Clustal (Thompson et al. 1994) was used for multiple sequence alignments when possible. The thirty targets 86-90,92,93,99-101,103,104,106,107,109,111-113, 115-123, 125,127, and 128 were modeled. Ten models were generated from each template using either the Kohonen algorithm or the analytic approach (Harrison 1999) coupled with energy minimization and a short run of molecular dynamics. The averaged model was energy minimized to generate the final model. The final models were subjected to 3ps runs of molecular dynamics, which may degrade the accuracy for the high homology examples. The variation among the models was calculated for each atom and used as an estimate of the uncertainty in the positions.

Ab Initio Folding
Ab Initio folding was used for targets 91,94,95,96,97,98,102,105,108,110,114,124 and 126. A simple hydrophobicity-electrostatics potential was supplemented by a sequence-specific empirical potential to improve the stereochemistry of the prediction. Inter-residue distances were estimated by searching the protein database for short stretches of homology from different and unrelated proteins. Simply finding the best local fit for each overlapping window of amino acids does not result in a good self-consistent set of distances. However, when the requirement for chain continuity is enforced, the problem of identifying a self-consistent set of inter-residue distances becomes akin to a convolutional error correcting code which is readily solvable by dynamic programming (Viterbi 1967). This continuity condition is an inherent property of all polymers and provides a significant gain in prediction accuracy. Potential templates were identified for homology modeling by using the proteins that had the most fits for each sequence. The models were generated in three steps. 1) 200 models were generated with the original potential

functions, using C-alpha-only models. Then inter-residue distances (C-alpha-C-alpha) were averaged over all the models. Those distances where the standard deviation was less than 2 angstroms were extracted. 2) A single model was generated that both satisfied the new distance information and minimized the hydrophobicity-electrostatics potential. This model was achiral and can represent either the left or right-handed solution. Secondary structure was identified visually and used to define additional distance restraints (published experimental data on helical locations were used for target 102). 3) All-atom models were built for both the right and left-handed solutions. The best models had right-handed helices.

References

de Berg M., van Kreveld M., Overmars M., and Schwarzkopf, O. (1997) ŧ Computational Geometry ŧ Springer-Verlag

Harrison, R.W (1999), A Self-Assembling Neural Network for Modeling Polymers J. Math. Chem. 26,125-137

Hertz J., Krogh, A., Palmer R.G. (1991) Introduction to the theory of neural computation, Sante Fe Institute studies in complexity lecture notes vol. 1. Addison-Wesley pp244-246,

JaJa J. (1992) An Introduction to Parallel Algorithms, Addison-Wesley pp 433-484,

Kelley LA, MacCallum RM, Sternberg MJ (2000), Enhanced genome annotation using structural profiles in the program 3D-pssm, J Mol Biol 299(2):499-520

Rychlewski L, Jaroszewski L, Li W, Godzik A (2000),Comparison of sequence profiles. Strategies for structural predictions using sequence information Protein Sci 9(2):232-41

Thompson JD, Higgins DG, Gibson TJ (1994), CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res 22(22):4673-80

Viterbi A.J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm IEEE trans inf theory IT-13,260-269.

---

# Sandia-USF , 039

number of submitted models: 41

**Predicting Secondary Structure with Ensembles of Decision Trees**

L.O. Hall and W.P. Kegelmeyer and C. Springer and K.W. Bowyer and N. Chawla and T. Moore, Jr.

*Sandia-USF*
*email:* avatar@ca.sandia.gov

The Sandia-USF approach to the secondary structure prediction of proteins was a purely pattern recognition based approach. The entire PDB as of July 11, 2000 was used as our training set. Structures between 2.5 and 3.0 A resolution deposited before April 1998 were not included due solely to a data handling glitch. This provided us with a training set of over 19,000 protein chains consisting of more than 3.6 million amino acids for secondary structure prediction. The secondary structure was determined by DSSP from the PDB coordinates.

We used the 20 x M matrix output from the -Q option of psi-blast to generate log-probability related values between -17 and 17 that each amino acid would be at a given position. Our approach is similar to that of Jones in CASP3, except that we used the "nr" (non-redundant) database without any filtering. Then, to predict the secondary structure of a given amino acid, a window of size 17 was used with the amino acid being predicted in the center of the window. Hence, there were 160 values to the left of the amino acid being predicted and 160 values to the right of it. These 160 values were made up of the 20 predictions of likelihood that an amino acid was in each of the eight positions to the left or right respectively of the amino acid whose structure was being predicted. Therefore, a feature vector describing a single amino acid in our training set consisted of 340 values ranging between -17 and 17 plus the class of H, C, or E.

For amino acids at the edge of a chain, values of 50 were used as padding. Proteins made up of more than one chain were predicted chain by chain.

Our base prediction engine was a set of decision trees generated by the C4.5 algorithm, a version of which was modified to run efficiently on the ASCI Red supercomputer. Learning algorithms typically require that all data fit in memory, which would require about 3 gigabytes of memory in this case. While it is possible to find machines with the required amount of main memory, such machines are rare. And though there are no difficulties with "convergence" when training with a decision tree, growing a single tree on all of the data requires a prohibitive amount of time.

Hence, we took a distributed approach. We were able to fit 1/8th of the data on one ASCI Red processor which had 256 MB of memory. Eight decision trees could be learned from a disjoint partition of the data. Uniform voting in which each tree gets to contribute a single vote was found to provide the best predictions.

As CASP4 progressed and time permitted we used different disjoint partitions of the data to produce sets of eight trees. In order to track our performance as we made modifications to our decision process, we created a validation set which consisted of all x-ray crystal structures added to the PDB between July 11 2000 and July 28 2000 with resolution better than 3.0A. This set consisted of 146 chains.

We found that higher prediction performance on our validation set could be obtained with 32 trees, that is, our top four sets of eight trees. (Again, each set represents a different disjoint partition of the data and all 32 sets together represent a four times resampling of the PDB). We incrementally added trees as they were obtained during the prediction process. The full 32 trees were used only in the last week.

Finally, we post-processed our predictions, in order to make use of the fact that, most generally, an amino acid's secondary structure will match that of its neighbors.

As a first step, we statistically examined all of the training data. A useful nugget of information, that singleton H's almost never occur, was used in a process we call "H-stomping." A singleton H is removed and replaced by the class which is shared by the majority of its neighbors in window of size 5.

Further, we used a smoothing algorithm to attempt to correct errors in a sequence of predictions. A window size of 5, centered around the amino acid whose secondary structure was being predicted, was used. The prediction strength at each position was the percentage of trees which predicted the class of H, C or E. The class with the highest prediction percentage at the center of the window was compared with a threshold. If the prediction for that class was below the threshold, and if the delta (difference between the prediction percentages of the top two classes) was below a second threshold then it was assigned the majority class of the window. The purpose of the thresholds was to insure that we changed the original prediction only in the cases where our decision tree engine was particularly uncertain about its own accuracy. The smoothing algorithm was updated several times over the course of the contest with the one described here used on the last day.

Thus the process for a target chain is: a) expand the chain into windows of log-probability around each amino acid, b) run each amino acid window through the currently available n trees and obtain the prediction percentages for each class, c) smooth the predicted class data and use the smoothed class as the predicted target class, d) H-stomp the smoothed chain.

The elements of our approach that are likely to be distinctive are (1) the pure statistical pattern recognition orientation, (2) the emphasis on using all available training data, and (3) the incorporation of oversampling methods.

REFERENCES:


@article{qui87,
author  = "J.R. Quinlan",
title   = "Simplifying Decision Trees",
journal = "International Journal of Man Machine Studies, V.27",
year    = "1987",
pages   = "227-248"
}

@book{quib,
author    = "J.R. Quinlan",
title     = "C4.5: Programs for Machine Learning",
publisher = "Morgan Kaufmann",
address   = "San Mateo, CA",
year      = "1992",
}

@book{mitch,
author    = "T.M. Mitchell",
title     = "Machine Learning",
publisher = "McGraw-Hill",
address   = "New York, NY",
year      = "1997",
}

@Article{Jones1,

```
  author =       "David T Jones",
  title =        "Protein Secondary Structure Prediction Based on
                  Position-Specific Scoring Matrices",
  journal =      "Journal of Molecular Biology",
  year =         1999,
  volume =       292,
  pages =        "195--202",
  note =         "www.ideallibrary.com"
}

 The SANDIA-USF AVATAR web site:
http://morden.csee.usf.edu/~chawla
```

---

# zhu , 352

number of submitted models: 54

**Assmebly of protein tertiary structures from Bayesian blocks**

Jun Zhu

*Amgen, INC.*
*email:* junz@amgen.com

There are two main forces that drive protein folding, local and non-local interactions. Local interactions are highly sequence-dependent. Sensitive sequence similarity search method will find favorable local structures. Non-local interactions will stablize local structures in optimal arrangements. This program is to simulate this folding process and devided in two steps: collecting locally similar blocks and assembling those blocks to achieve the minimum engery state.

First, local structure similarities were predicted based on sequence information. Bayesian model (Zhu et al, 1999) was used as sequence similarity search method because it is a natural block-based method, which only aligns regions of high confidence. And also each alignment has a confidence value that can be explored during global energy minimization. Each target sequence was searched against NR database. Similar sequences were collected and used in iteratively training a Bayesian model. Then, the model was searched against PDB45 database. Aligned blocks that met the following conditions were collected: (1)the Bayesian evidence of the pair is less than 0.1, (2) the highest confidence level of residues in the aligned block is higher than 0.5, (3) the length of consecutive high confidence region is larger than 10. Long blocks were broken down to small overlapped blocks of 10~13 AAs.

After obtaining structure blocks, blocks were assembled using simulated annealing method to miniminize energy functions. The energy functions were derived similarly as Simon et al. (1999).

We investigated sequence-independent forces (compactness and secondary structure elements interaction), first order sequence-dependent forces (environment force, non-specific pair distance function), second order sequence-dependent forces ( environment pair interactions, sequence specific pair distance function). The relative weight of each function were estimated using randomly generated structures. We found that compactness expressed in term of radius of gyration, beta-strand interactions and first order environment force have large contribution to the fininal energy function and used in fininal structure predictions.

Monte Carlo simulated annealing process was used to generate structures and minimize energy functions. 5000 iterations were performed in each trial. Temperature was scheduled to decrease linearly. 100 trials were conducted for each target and the resulted predictions were ordered by the same energy functions. The top fives were submitted for evaluation.

References:
(1)J. Zhu, R. Luethy and C.E. Lawrence (1999) ISMB99 297-305.
(2)K.T. Simons, I. Ruczinski, C. Kooperberg, B.A. Fox, C. Bystroff and D. Baker (1999) Proteins 34:82-95.

---

# Kollman-Baker , 498

number of submitted models: 26

**Rosetta/AMBER Hierachy Method, Molecular Dynamics refinement of Rosetta predictions**

Matthew R. Lee David BakerPeter A. Kollman

*UCSF*
*email:* mrlee@cgl.ucsf.edu

For Rosetta, 3 and 9 residue fragments consistent with the query sequence's profile and secondary structure prediction (via SAM, PHD, and PSI-PRED) were assembled using Monte Carlo and a Bayesian scoring function with both sequence dependant and independant terms. The results from several thousand independant runs were filtered and clustered and models selected from cluster centers. Models passing several of the filters prior to clustering have side chain atoms built on by simulated annealing (Dunbrack phi-psi dependant rotamer library) and are then rescored. This score is then used as a last filter prior to clustering. (Simons et al. J Mol Biol 1997;268:209-225). For AMBER, the 20 to 30 most highly populated clusters from Rosetta were subjected to ~150 ps of molecular dynamics with the Cornell et al. (JACS 1995;117:5179-5197) force field. The 1-5 submitted conformations were final snapshots from the trajectories with the lowest average free energies, according to the Molecular Mechanics ∃ Poisson Boltzmann Surface Area free energy function (MM-PBSA) as previously applied towards the study of protein stability of HP-36 (Lee et al. Proteins 2000;39:309-316).

For equilibration in the solvated molecular mechanics representation, we first minimized the Rosetta structures with the steepest descent method for 500 steps, followed by the conjugate gradient method until the RMS of the Cartesian elements of the gradient was less than 0.4 kcal/mol□? Water

molecules alone were then minimized in the same way until the RMS was less than 0.1 kcal/mol?and then slowly heated, while allowing them to move unrestrained for 25 ps (with a 1 fs timestep) in order to fill in any vacuum pockets. The solute atoms alone were then minimized in the presence of ever decreasing positional restraints, thereby allowing them to slowly feel the forces of the now equilibrated waters, until the positional restraints reached zero. Finally, a temperature ramp was used to gradually raise the temperature of the whole system over 20 ps up to 300 K. Equilibration was followed by production-phase molecular dynamics, which involved a 2.0 fps timestep under the isothermal-isobaric ensemble (300 K and 1 atm), the TIP3P model for water, periodic boundary conditions, particle mesh Ewald method (PME) for electrostatics, a 10 ?cutoff for Lennard-Jones interactions, and the use of SHAKE for restricting motion of all covalent bonds involving hydrogen, all within the AMBER 5.0 suite of programs. The MM-PBSA calculation evaluation was performed on single snapshots from the production dynamics trajectory, every 10th ps. The MM-PBSA free energy of each snapshot is approximated as the sum of the internal energy of the protein, estimated from AMBER, and the solvation free energy, estimated from DelPhi for electrostatics and from a linearly dependent molecular surface term for non-polar contributions. The average free energy is then taken and used for comparison among the various conformations.

---

# Harold-Scheraga , 004

number of submitted models: 80

**Ab initio energy-based protein-structure prediction using the UNRES and ECEPP/3 force fields - test on CASP4 targets**

Jaroslaw Pillardy, Adam Liwo, Cezary Czaplewski, Jooyoung Lee, Daniel R. Ripoll, Rajmund Kazmierkiewicz, Jeffrey A. Saunders, Stanislaw Oldziej, William J. Wedemeyer, Kenneth D. Gibson, Yuan-Jie Ye, Harold A. Scheraga

*Cornell University*
*email:* has5@cornell.edu

The structures of the target proteins were predicted using a hierarchical algorithm consisting of five stages, in which the tertiary structure is predicted at low resolution and then refined. In stage one, the protein is represented by a simplified low-resolution model, in which the atoms of the peptide bonds and side-chains are replaced with two centers of interactions. The virtual bond lengths and the distance between the side-chain centroids and C-alpha atoms are held fixed, but the virtual bond angles and the orientation of the side-chain bond are variable. The interactions of this simplified model are described by a united-residue (UNRES) potential, which was parametrized using primarily a physical potential and a novel Z-score optimization method. Our conformational space annealing (CSA) method was used to search for the lowest-energy families of conformations of the simplified UNRES model. The five families with the lowest UNRES energy were chosen as models 1-5; the structures of these models were then refined in stages 2-5, as described below. In all subsequent stages the C-alpha positions were restrained to resemble the protein topology obtained at the UNRES level.

In stage 2, the low-resolution UNRES models were converted to all-atom backbone representations of the protein. In this stage, all residues were

treated as one of three types of residues (Gly, Pro and Ala), i.e., all non-prolyl and non-glycyl residues were represented initially as alanines.

In stage 3, the polylalanine model of the target protein was refined using our Electrostatically-Driven Monte Carlo (EDMC) method to reduce its all-atom ECEPP/3 energy.

In stage 4, full all-atom representations of the five models are constructed, i.e., the side-chain atoms are added. The ECEPP-energy is then minimized and a search is carried out iteratively residue-by-residue over each consecutive dihedral angle until self-consistency is reached. In stages four and five, hydration energies are included using the SRFOPT algorithm.

In stage 5, the full all-atom structures of models 1-5 are refined using the EDMC algorithm to lower the ECEPP/SRFOPT energy.

REFERENCES

(1) A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackovsky, H.A. Scheraga, Protein Science, Vol. 2, 1715-1731, (1993).
(2) J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, J. A. Saunders, K. D. Gibson, H. A. Scheraga, Int. J. Quant. Chem., Vol. 77, pp. 90-117, (2000).
(3) A. Liwo, J. Pillardy, C. Czaplewski, J. Lee, D. R. Ripoll, M. Groth, S. Rodziewicz-Motowidlo, R. Kazmierkiewicz, R. J. Wawak, S. Oldziej and H. A. Scheraga, RECOMB 2000, Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, pp. 193-200, (2000).

---

# ELAN , 512

number of submitted models: 83

**ELAN-PROT -- a fast elastic net prediction of protein structure**

Keith D. Ball, Burak Erman and Ken A. Dill

*University of California at San Francisco*
*email:* kdb@maxwell.ucsf.edu

ELAN-PROT (ELAstic Net prediction of PROTein structures)
predicts basic tertiary structure (backbone topology)
using the elastic net mean-field annealing method [1]
applied in a novel fashion to approximate the internal free
energy landscape of a protein C_a chain.

The method minimizes a fitness function F analogous in form and behavior
to that of a free energy:

F = E_int + TS,

where E_int is the internal potential energy of the protein chain,
T is a temperature-like annealing parameter, and S behaves as
an entropy.

The internal energy is expressed as

E_int = E_stat + E_bond + E_HP + E_EV + E_2nd.

E_stat is a residue-residue energy function incorporating a knowledge-based statistical contact potential [2]. E_bond is a restraint term for covalent peptide bonds. E_HP is an attractive centrosymmetric term which promotes compactness and hydrophobic burial. The E_2nd term enforces constraints derived from secondary structure predictions obtained from the PHD [3,4] server and the JPRED2 consensus prediction server [5,6]. When run as a fully-automated prediction server, the Predator prediction algorithm [7] was run locally locally to provide secondary structure predictions. An explicit excluded volume potential term E_EV was also added to prevent structural collapse and to promote self-avoidance of the C_a chain.

Terms of the energy contributions E_stat, E_bond and E_2nd take the form of harmonic spring potential:

E = 0.5 * a * (r - r_0)^2

The basis for the use of harmonic, or "Gaussian" springs derives from the Gaussian model of protein self-interactions [8].

The following parameters were used for most predictions:

| | a | r0 |
|---|---|---|
| stat | 0.05 * a_ij | 0.0 |
| bond | 10.0 | 3.8 |
| 2nd | 20.0 | (varies) |

E_HP terms have a relative force constant of 0.1. The excluded volume interaction is modeled by a repulsive harmonic "half-spring", in effect when r < r', with r' = 3.6 A.

In addition to E_stat, there are harmonic virtual "springs" between each residue and a set of virtual points in space. These springs have variable force constants, whose magnitudes vary as a Gaussian of the distance between residues and cities. These points correspond to the cities in the Traveling Salesman Problem (TSP), after which our approach is modeled. The entropy-like quantity S is based on the interactions between "cities" and residues. Our method is not an exact analogy to the application of the elastic net method to the TSP; for instance, the residues to not match up to the virtual "cities" at T=0. Nevertheless, this construct allows us to make a deterministic mean-field approximation to the free energy, thus saving the enormous time cost of a stochastic conformational search.

Since S behaves as an entropy, the function F will be convex for sufficiently high T, and can be minimized simply using steepest-descent. As T is lowered, we track the minimum of F by re-optimizing F at each new temperature. As T->0, we obtain F=E. Hence, by globally optimizing F and annealing down to T->0, we obtain, in principle, the global minimum of E.

Initial chain conformations were generated randomly via a Monte Carlo procedure. The parameter T was lowered after the system was minimized at the current value of T. In practice, we terminated the algorithm at finite T. To create an all-heavy-atom model from this C_a chain, we used PULCHRA version 0.81 [9] to reconstruct the backbone, followed by sidechains were added using the SCWRL algorithm [10] employing a

backbone-dependent rotamer library [11]. Steric clashes were then further resolved by minimizing the structure using the EEF1 energy function [12]. In addition, we used the EEF1 energy as a scoring function to determine the ranking and feasibility of our models. Although the addition of atomic detail is not central to our algorithm, we added it in order to employ the EEF1 energy function, as well as to conform to the desired CASP4 submission format.

References:


[1] Durbin & Willshaw, Nature 1:348-358 (1989).
[2] Thomas & Dill, Proc. Nat. Acad. Sci. USA 93:11628-11633 (1996).
[3] B. Rost & C. Sander, J. Mol. Biol. 232:584-599 (1993); Proteins 19:55 (1994).
[4] B. Rost, Meth. in Enzym. 266:525-539 (1996). http://www.sdsc.edu/predictprotein/
[5] http://jura.ebi.ac.uk:8888/
[6] Cuff J. A. & Barton G. J., PROTEINS 34:508-519 (1999).
[7] D. Frishman and P. Argos, PROTEINS 23:566 (1995).
[8] I. Bahar, A. R. Atilgan & B. Erman, Fold. Des. 2:173 (1997); B. Erman & K. Dill, J. Chem. Phys. 112:1050 (2000).
[9] written by Piotr Rotkiewicz (piotr@pirx.com)
[10] M. J. Bower, F. E. Cohen & R. L. Dunbrack, Jr., J. Mol. Biol. 267:1268-1282 (1997).
[11] R. L. Dunbrack, Jr. & M. Karplus, J. Mol. Biol. 230:543-574 (1993); Nature Struct. Biol. 1:334-340 (1994); Prot. Sci. 6:1661-1681 (1997).
[12] T. Lazaridis & M. Karplus, J. Mol. Biol. 288:477-487 (1998).


---

# Rose-Group , 035

number of submitted models: 19

**LINUS: Protein Folding by Computer Simulation**


Rajgopal Srinivasan and George D. Rose

*Johns Hopkins University*
*email:* raj@grserv.med.jhmi.edu


All predictions are based on the LINUS algorithm (Srinivasan and Rose (1999) Proc.Nat. Acad. Sci, 96: 14258-14263; Srinivasan and Rose, (1995) Proteins, 22:81-99). The algorithm, which operates solely on the amino acid sequence, establishes the protein's intrinsic biases toward each secondary structural state -- helix, strand, turn or coil.

In greater detail, the polypeptide chain is represented in atomic detail by all heavy atoms; no hydrogens are included. Starting with the chain in an extended conformation, five independent Monte Carlo simulations of 10,000 cycles each are performed. A cycle proceeds from the N- to C-terminus, changing the conformation of each three-residue segment at random to one of four possibilities: helix, strand, turn or coil. Accordingly, each cycle generates N-2 structures, where N is the number of residues in the protein. At each three-residue step, the ensuing conformation is screened for the presence of steric clashes. If steric clashes exist, another conformational change is attempted. In the absence of steric clashes, a score for

that conformation is computed, using a simple scoring function which rewards hydrogen bonds and hydrophobic contacts. This trial structure is then either accepted or rejected, using the Metropolis criterion. At the end of every cycle, the secondary structure of the protein is assigned, using a method based exclusively on the backbone dihedral angles. At the end of the 10,000-cycle simulation, the fractional distribution of every residue in each secondary structure state is computed; these distributions are the discovered biases for each residue to populate helix, strand, turn and coil conformations. These distributions are then used to bias the random selection of three-residue conformations in subsequent simulations.

When scoring a conformation, the initial set of simulations only counts interactions between residues that are close in sequence (viz., <7 residues). This interval of allowed interaction is increased incrementally in subsequent simulations. For CASP4, simulations were performed at intervals of 6, 18 and N, the sequence length of the protein. Conformational biases are interpreted as the probability of each residue to be in one of the four secondary structural states, as described above. At each interval, these LINUS-evolved biases were extracted and used to guide the three-residue sampling in subsequent intervals.

Structures submitted for final evaluation were taken from the highest scoring conformers.

---

# Braun-UTMB , 223

number of submitted models: 104

**3-D MODELING OF PROTEIN TARGETS FOR THE CRITICAL ASSESSMENT OF STRUCTURE PREDICTION. COMPETITION (CASP4):**

V.S. Mathura, K.V. Soman, C.H. Schein, Y. Xu, and W. Braun

*UTMB*
*email:* yuan@planck.utmb.edu

The Human Genome Project has revealed many proteins of unknown function. Classification of these sequences can best be done by accurate prediction of their structures, and concurrent assignment to families of known function. We have developed a set of tools for homology modeling of proteins(1,2), based on self-correcting distance geometry (DIAMOD)(3,4,5), multiple sequence alignment (MASIA (6))and energy minimization (FANTOM(7)), that can be used even when the identity to the target is very low(8) (300r less(9)). CASP4 provided us with an opportunity to evaluate our methods impartially and objectively. We submitted a total of 100 models for 27 of the 43 targets, with 15 based on sequence homology. Models for five targets were generated ab initio. The rest used a combination of fold recognition with multiple alignment to improve the sequence register between the target and selected template.

Homology or comparative modeling (CM)

When a suitable template was identified in the Protein Data Bank for a target, our comparative modeling procedure was to: (1) Align the target sequence with one or more template sequences using the program CLUSTALW or alignments suggested by the fold recognition servers(CAFASP) with minimal manual adjustment; (2) extract distance and dihedral constraints with our in-house program EXDIS; (3) build initial models with DIAMOD; and (4) energy minimize using the FANTOM program. For T90, a consensus aligment was prepared manually from the 3D-PSSM, BIOINBGU, FUGUE, GENTHREADER, and Karplus HMM98 and SAM99 results. FANTOM energy

contributions and exposed apolar surface areas calculated with the program GETAREA were used for ranking multiple models for the same target. Where information was available for important residues in the template, such as those within the active site or areas of substrate binding, we compared their location visually in the model structure.

Fold recognition (FR)

When there was not high enough sequence homology with any protein of known structure, threading (fold recognition) was attempted, using the web servers mentioned above and others (PSI-BLAST, 123D and FFAS). Where several methods suggested the same template, a consensus alignment was prepared manually. Manual corrections/adjustments were also used to insure that secondary structures and active site or other critical residues were aligned. For T91, an alignment from 3D-PSSM was manually edited to improve the sequence alignment. We also used multiple sequence alignment of protein families where a fold seemed clear cut. For example, fold recognition identified T88 as a probable Greek key fold and selected yeast killer toxin (1wkt) as a template. Another template structure,1A45, that more closely resembled T88, was selected from a multiple alignment with 57 b/g-crystallins. The indicated gapping pattern from the multiple alignment was used to generate a model.

Ab initio modeling

When a suitable template could not be identified based on homology or threading, but there were clear indications of conserved secondary structure elements based on sequence alignments with related proteins, we prepared ab initio models. The steps for generating ab initio models for T88, T91, T97, T104, and T106 were: (1) Predict secondary structures and exposed/buried residues of the protein from aligned sequences with JPRED and MASIA; (2) convert this information into distance and dihedral angle constraints using the program TRANSLATE; (3) add other constraints derived from any available experimental data for the protein; (4) build models from constraints with DIAMOD; (5)refine initial models by energy minimization FANTOM. We also submitted models based on fold recognition methods for T88 and T91.

Ab initio constraints were used in several other models where appropriate. Di-sulfide bond constraints were added during the modeling of T123 and T125. In another example, for T86, a monomer, a trimeric template of very low identity was identified based on functional similarity and conservation of key active site residues. A multiple alignment with target homologs was used to place probable gaps between the template and target sequences and constraints were extracted from the template according to our usual methods. Ab initio constraints were added at the C-terminal to replace inter-subunit contacts present in the trimer.

Multiple alignments help in FR and CM

We combined these techniques in preparing alignments where the identity between the target and template was very low (such as T86 and T88), when the target had a clear sequence relationship to several templates, or when several sequences related to the target were known. For T101, which had about the same degree of sequence identity/similarity to 6 known protein structures(12-18%), a CLUSTALW multiple alignment of related proteins of the pectate/pectin lyase family was used. This agreed with the DALI alignment of the pectate lyases but not of a structurally related protein, chondroitinase(1DBG). We made models based on the B. subtilis pectate lyase(1BN8) using the multiple alignment to adjust gapping. Other models were based on the fold recognition results for 1DBG (where there was no real consensus for most of the protein).

In keeping with our efforts to use genomic data efficiently in modeling, we used the homologous sequences available for a few of the targets. For T118, PDB- BLAST detected similarity of the C-terminal with 1DDQ-A. The 1DDQ-A sequence was aligned with T118 and related bacterial and fungal polymerase alpha-factors to obtain the gapping used in the submitted alignment. PDB-BLAST also recognized a weak pattern of identity between T126 and 1DMS and 1EG9. Individual multiple aligments of T126 with other olfactory factors and these templates was used to

generate the alignments submitted.

1 Soman, K.V., Midoro-Horiuti, T., Ferreon, J.C., Goldblum, R.M., Brooks, E.G., Kurosky, A., Braun, W. and Schein, C.H. (2000) Biophysical Journal 79:1601-1609
2 Soman, K.V., Schein, C.H., Zhu, H. and Braun, W.A. (2000) Homology Modeling and Simulations of Nuclease Structures. In Methods in Molecular Biology (Humana Press, Totowa, N.J.; editor C.H. Schein) 160(in press for December, 2000).
3 Zhu, H., Schein,C.H. and Braun,W. (1999). J. Mol. Modeling, 5,302-316.

4 Mumenthaler, Ch. and Braun, W. (1995) Protein Science 4, 863-871
5 Zhu, H. and Braun, W. Protein Sci. 1999, 8, 326-342
6 Zhu, H., Schein,C.H. and Braun,W. (2000) MASIA: a program to recognize common patterns and properties in multiple aligned protein sequences. Bioinformatics 16: in press
7 Fraczkiewicz, R. and Braun, W. (1998) J. Comp. Chem. 19, 319-333.
8 Mumenthaler, Ch., Schneider, U., Buchholz, Ch.J., Koller, D., Braun, W. and Cattaneo, R.(1997 ).Protein Sci 6, 588-597.
9 Buchholz, C.J., Koller, D., Devaux, P., Mumenthaler, Ch., Schneider-Shaulis, J., Braun, W., Gerlier, D. and Cattaneo, R. (1997). J. Biol. Chem. 272, 22072-22079

---

# Murzin , 384

number of submitted models: 21

**Distant Homology Recognition and Fold Prediction by a knowledge-based approach using SCOP and Pfam**

Alexey G. Murzin and Alex Bateman

*Centre for Protein Engineering, Cambridge, UK*
*email:* agm@mrc-lmb.cam.ac.uk

As submitted in the Fold Recognition category

Since our team⁄s last performance in CASP2 four years ago, we have been working on the methods that could extend the superfamilies of known structure in SCOP to the sequence families of unknown structure in Pfam and other sequence libraries. We entered CASP4 hoping that this prediction experiment would provide an opportunity to test our new methods. A systematic work on the extension of SCOP superfamilies has already resulted in the structural assignment of many sequence families of unknown structure and, often, unknown function. Indeed, in CASP3, there were at least three targets predictable by this approach. Disappointedly, however, none of the CASP4 targets turned out to be in our list of protein families with already assigned structures.

Therefore, in CASP4 we used essentially the same approach as developed for CASP2 (Murzin A.G. and Bateman A. Distant homology recognition using structural classification of proteins. Proteins, Suppl. 1:105-112, 1997). We searched for probable homologues of the target sequences and available biochemical information on the target protein and/or its sequence family and used the predicted secondary structure to shortlist the SCOP superfamilies, to which each attempted target may belong. Predictions were based on the

discovery of superfamily specific characters. The experience and expertise gained from our working on SCOP and Pfam databases were of a great help in this knowledge-based approach. Also, we tried our knowledge-based approach in the two other prediction categories. We used superfamily specific features to improve the alignments in some of the comparative modelling targets. For several targets, predicted by our approach to be not related to any of the SCOP superfamilies, we attempted the fold prediction using the conservation patterns in the target sequence families, the available biochemical data and/or the empirical folding rules derived from known protein structures.

The choice of prediction format, TS, and the target selection were influenced by the CASP3 Fold Recognition assessment experience (Murzin A.G. Structure Classification-Based Assessment of CASP3 Predictions for the Fold Recognition Targets. Proteins Suppl. 3:88-108, 1999). To ensure the detection of (partly) correct predictions by both sequence-dependent and sequence-independent numerical evaluation procedures, each of our predictions was composed of the regions of confident structure and alignment, the regions of confident structure but tentative alignment, and the regions of tentative structure. The 3D coordinates for the most of the target atoms were the best way to represent this structural mosaic in a single format. As one of us strongly opposed to the NONE prediction, this option was not used. Therefore, in the absence of predicted homologous structure, we either built a 3D model of our prediction "ab initio", or had it dropped. Only one model was submitted for each of the completed predictions. Apart from the two targets whose structures were known to us before they were submitted to CASP4, we did not attempt the large, presumably multi-domain targets without apparent domain boundaries. Because of time limitations, we also ignored late comparative modelling targets including all but one of the predicted members of the P-loop hydrolase superfamily. Due to the presence of characteristic P-loop motifs in their sequences, their homology recognition seemed straightforward, and the actual challenge was the alignment. All other targets were attempted but six or so of them were dropped eventually. In total, we submitted predictions for 21 targets. This include four Comparative Modelling targets, T0090, T0092, T0093(!) and T0103; ten Distant Homology Recognition targets, T0088, T0096_1, T0098, T0100, T0101, T0104, T0108, T0109, T0118 and T0121_2; three targets with predicted known folds (there may or may not be a distant homology), T0095, T0102 and T0114; and four targets with predicted (probably) novel folds, T0086, T0091, T0094 and T0110.

Many of the Distant Homology Recognition predictions were based on the result of previous analysis of SCOP superfamilies, for example the pectate lyase beta-helix fold of T0100 and T0101 (Chothia C. and Murzin A.G. New folds for all-beta proteins. Structure 1, 217-222, 1993). There were several cases of d?a vu. T0108 had the same characteristic feature as the CASP4 target T0038 and was modelled on the experimental structure of the latter. In T0121_2, there was the OB-fold signature similar to one we derived for the prediction of T0004. For the fold prediction of T0102, we used the same pseudo "ab initio" approach as we used for the CASP2 target T0042. Incidentally, the predicted fold of T0102 was found to be similar to the experimental fold of T0042. In T0086, there was a probable tandem repeat of two (alpha)-alpha-beta-beta-beta motifs, detected by the analysis of its extended sequence family, analogous to the approach that detected the internal duplication in T0002_2. Similarly, a tandem repeat of two beta-alpha-beta-alpha-beta motifs was detected in the extended T0094 sequence family. Unlike T0002_2, there was no SCOP superfamily assigned for either T0086 or T0094. Both target structures were modelled "ab initio".

One of our CASP2 techniques, not credited properly at the time because it had been used only for the late target T0026, was in great use through most of our CASP4 predictions. For almost every target predicted to belong to a large superfamily with many known structures, a composite template structure was assembled from different fragments of several superfamily structures superimposed onto their common fold. It allowed the selection of the most suitable parts from different structures. In particular, the predicted

structure of the P-loop hydrolase T0104 was assembled from the fragments of several topologically distinct members of this very diverse superfamily to generate a novel topological variant. For a number of our predictions, we also created hybrid templates including fragments of non-homologous structures to model the ‘missing’ parts in the parent structure or even to construct the whole fold. Then we used Modeller to generate the 3D coordinates, automatically sealing the gaps and fixing the stereochemistry of the joints.

---

# baker , 354

number of submitted models: 174

**Ab Initio Structure Prediction Using Rosetta**

Richard Bonneau, Jerry Tsai, Ingo Ruczinski, David Baker

*University of Washington*
*email:* bonneau@u.washington.edu

Ab initio structure predictions for CASP4 were made using ROSETTA. The basic method is described in the CASP3 volume of proteins (Simons et. al, 1999). One of the fundamental assumptions underlying Rosetta is that the distribution of conformations sampled for a given short (3-9 residue) segment of the polypeptide chain is reasonably well approximated by the distribution of structures adopted by the sequence of the segment and closely related sequences in known protein structures. Fragment libraries for each three and nine-residue segment of the chain are extracted from the protein structure database using a sequence profile-profile comparison method as described previously (ref). The conformational space defined by these fragments is then searched using a Monte Carlo procedure with an energy function that favors compact structures with paired beta strands and buried hydrophobic residues. 1000-10000 independent simulations are carried out (starting from different random number seeds) for each query sequence, and the resulting structures are clustered as described previously.21 For CASP4 models were refined using a more physically realistic potential function and only the relatively low free energy structures were subjected to the clustering procedures. The potential function included a hydrogen bonding term, a solvation term based on solvent accessible surface area scaled using atomic solvation parameters, and a packing term using a modified Lennard-Jones potential. The backbone torsion angles were varied using a Monte Carlo procedure with two types of moves: 1) small phi/psi changes at single amino acids, and 2) insertions of 3 residue fragments with subsequent minimization of the relative displacement of the flanking portions of the chain. At each step, sidechains were built onto the backbones using Dunbrack's backbone dependent library and a very rapid simulated annealing protocol. For the largest proteins, these relaxation runs used largely as a post-clustering filter because of resource limitations. For larger targets as many as 200,000 independent simulations were carried out and the population was then trimmed considerably by filters designed to remove systematic errors and incorrect fold. It was found prior to CASP4 that Rosetta produces high contact order structures too infrequently and that many structures had multiple paired strands. Simple filters were thus used to remove low contact order structures and structures with many unpaired strands prior to the refinement and clustering steps. For proteins with clear sequence homologues, independent simulations were also run on 2-5 of the homologues as described above. The resulting populations were then clustered

simultaneously to identify conformations frequently populated by most/all of the homologues. The ideal result of this procedure is that only the correct free energy minima will be present for all of the homologs folded, and that the false minima (incorrect conformations) will be sufficiently different for different aligned sequences so as to be diluted out as the homologous decoy populations are combined (Bonneau, In Press).

Bonneau R, Strauss CEM,Baker D. Improving the Performance of Rosetta Using MSA information and Global Measures of Hydrophobic Core Formation. Proteins, In Press.

Simons KT, Bonneau R, Ruczinski I, Baker D. Ab Initio Protein Structure Prediction of CASP 3 Targets Using Rosetta. Proteins 1999; Sup 3:171-176.

---

# Levitt , 012

number of submitted models: 180

**Ab Initio Folding by Repeated Energy Minimization**

Chen Keasar and Michael Levitt

*Stanford University*
*email:* michael.levitt@stanford.edu

The Levitt group work for ab initio modeling was a collaboration between Michael Levitt and Chen Keasar. It was based on torsion angle energy minimization with predicted secondary structure. Using different random starting points generated many models. These models were then reduced to the five submitted by the same procedure used for all the models submitted by the Levitt group. We considered 8 CASP4 targets to be Ab Initio targets (T0086, T0091, T0097, T0102, T0105, T0106, T0110, T0114) and built models for 7 of these (T086 with 164 residues was larger than the others and was skipped). The targets modeled had between 70 and 128 residues.

The energy minimization, which used the method published some time ago (Levitt, M. Protein Folding by Restrained Energy Minimization and Molecular Dynamics. J. Mol. Biol. 170, 723-764 (1983)), differed from that work by virtue of three added terms: (a) Hydrophobic residues were encouraged to form clusters, (b) main chain hydrogen bonds were cooperative so as to reward the arrangements of hydrogen bonds that are found in protein alpha-helices and beta-strands and (c) charged residues were restricted to the proteins surface by a cooperative hydrophilic term. Because the energy function had to be very smooth with a continuous value and continuous first and second derivatives, the modified program was a tour-de-force on the part of Keasar as he implemented his concepts of allowed and disallowed patterns of hydrogen bonds. The vast conformational space was sampled by repeating the minimization from different random starting conformations. We generally generated between 20,000 and 60,000 such decoy conformations for each target.

The methods used for secondary structure prediction (used as an initial constrain for the minimization) and for choosing among the decoy conformations were taken from the CAFASP servers as explained in the abstracts for Comparative Modeling and Fold-Recognition. What follows is similar to the relevant parts of those abstracts.

This work was greatly aided by the availability of the output of all the 30 or so servers participating in CAFASP on the CAFASP web site at http://cafasp.bioinfo.pl/target. In general these results were available within hours of the target sequence announcement and we never felt the need to consult the original servers in any way.

The consensus secondary structure was taken from Levitt's parsing of the CAFASP server results also used for Fold-recognition and Comparative Modeling. For secondary structure we used the results from the CAFASP files jpred, psipred, target99, pssp and sspro. Because jpred is itself a consensus prediction server, our tabulation depended on the work of a very large number of groups that we cannot acknowledge here. For each target we produced a summary file that listed the secondary structure as follows:

```
T0106_jpred2_sq_
AACEPVRIPLCKSLPWEMTKMPNHLHHSTQANAILAMEQFEGLLGTHCSPDLLFFLCAMYAPICTID
T0106_jpred2_sq_Cys    nCys=10                    C       C
C        C      C
T0106_0.Conservation_ss_                         ..........6.............................
..6....+...5..75+..6.6.+...
T0106_1.psipred_ss_
....EE.HHHH.................HHHHHHHHHHHHHHHH.....HHHHHHHHHH........
T0106_2.target99_ss_
....EEEE.E..................HHHHHHHHHHHHHHHHH..HHHHHHHHHH........
T0106_3.jpred2_ss_jpred
............................HHHHHHHHHHH..........HHHHHHH..........
T0106_3.jpred2_ss_phd predicti
............................HHHHHHHHHHHH........HHHHHHHHHE........
T0106_4.pred2ary_ss_
............................HHHHHHHHHHHH..HH......HHHHHHH.........
T0106_jpred2_ss_JNETALIGN
............................HHHHHHHHHH.............HEEEEE.........
T0106_jpred2_ss_JNETFREQ
......EE....................HHHHHHHHHHH..HH......HHHHHHEE.........
T0106_jpred2_ss_JNETHMM
............................HHHHHHHHHHH.........HHHHHHHHHHH.......
T0106_jpred2_ss_JNETPSSM
....HHHHHHH.......EE.........HHHHHHHHHHH..HHH...HHHHHHHH....E......
T0106_jpred2_ss_dsc prediction
.......................HH.....HHHHHHHHHHHHHHH......HHHHHH..........
T0106_jpred2_ss_jnetpred
.....HHHHH..................HHHHHHHHHHH.........HHHHHHHH..........
T0106_jpred2_ss_nnssp predicti
............................HHHHHHHHHH.............EEEE..........
T0106_jpred2_ss_pblock predict
.......EE.....................HHHHHHHHHH.EEEE......HEEEXE....EE...
T0106_jpred2_ss_predator predi
.........................HHHHHHHHHHHH.............HHHHHHH........
T0106_jpred2_ss_zpred predicti
............................HHHHHHHHHHHHHHHH...EE..HHEEEE...EEEEE..
T0106_pssp_ss_
............HH..............HHHHHHHHHHHHH.........HHHHHH......EEEE
T0106_sspro_ss_
..........H.................HHHHHHHHHHHHHHH.....HHHHHHHHHH........
```

For more complete results see our "private" site at: http://csb.stanford.edu/levitt/casp1234. During the CASP event, information contained in that site was updated regularly by Levitt and shared with the different CASP4 groups in my lab headed by Samudrala, Xia, Fain and Koehl respectively. This is the only information that was shared. Each group then went on to make their own comparative models (Samudrala, Koehl and Levitt) and/or ab initio models (Fain, Levitt, Samudrala, and Xia). There was no comparison of models, as

each individual preferred to use CASP as an opportunity to prefect their methods rather than to "win" CASP.

Finally the best models were selected as follows. Use the rapdf probability score (Samudrala, R & Moult, J. An All-atom Distance-dependent Conditional Probability Discriminatory Function for Protein Structure Prediction. J. Mol. Biol., 275: 893-914, (1998)) to choose the best 1000 models. Cluster all these 1000 models into 10 clusters (using bottom-up hierarchical clustering based on inter-structure CA coordinate RMS deviation). For each model we use the rapdf score, Samudrala's HCF hydrophobic compactness score, Keasar's surface energy, and the number of hydrogen bonds to rank the conformations in each cluster. Finally choose the five lowest energy models never including more than one model from a given cluster. Occasionally manual intervention was used in deciding the rank of the models in the official submission to CASP. For this we viewed the models to judge general protein like shape and also used the coverage (for comparative modeling). For example, a model with a less favorable energy score may be ranked above a model with better score if the first model covered more of the target sequence.

There were some difference for two of the targets. For T102, a cyclic 70-residue protein, we tried two methods: (a) we added Cys residues at each end of the chain and used this to generate an approximately cyclic model. (b) we had no constraint but then closed the ends by Cartesian coordinate energy minimization. This gave us a total of 60,000 decoys that were discriminated as described above. For T0106, a 128-residue protein with 10 Cys residues, we assumed there would be 5 SS bonds. There are 10!/2^5 = 113,400 possible combinations, which is too large for exhaustive examination of them all. Instead we did a quick analysis of common SS patterns and choose three different patterns. Each pattern was then used with the predicted secondary structure in three independent energy minimization runs. Al the decoys were pooled and discriminated as described above.

---

# Friesner , 414

number of submitted models: 150

**Ab Initio Tertiary Structure Prediction using Global Minimization of a Size-dependent Potential Energy Function**

An, Y. Eyrich, V.A.Gunn, J.Pincus, D.L.Standley, D.M.Friesner, R.A.

*Columbia University*
*email:* rich@chem.columbia.edu

We attempted ab initio prediction for a number of small helical proteins (or helical domains), specifically for cases that either were said to be ab initio targets in the CASP4 additional information, or for which we were unable to locate a homologue using our fold recognition technology. An outline of the algorithms that we employed in these calculations is as follows:

(1) Secondary structure prediction: predictions of the target sequence were obtained from four public servers: PSIPRED, JPRED, SSPRO, and PHD.

(2) Tertiary folding simulations: The computational details of the tertiary folding simulations are briefly described as follows:

(a) An off-lattice model containing backbone atoms plus a pseudo atom representation of the side chain for each amino acid was employed.

(b) The geometrical variables in the simulation were the phi and psi angles in the loop regions; angles in alpha helical regions were fixed to ideal values (-57, -47 degrees).

(c) The potential was a function of the distance between the side chain pseudo atoms and the identities of the interacting residues. The functional form was a general cubic spline that allowed great flexibility along with rapid computation of energies and gradients. In general, hydrophobic-hydrophobic interactions were attractive, and hydrophilic-hydrophilic interactions were repulsive, as in the statistical potential of Sippl and coworkers [1]. However, the potential was designed to vary as a function of protein size; we have found this modification to be essential for obtaining reasonable results for test cases. The size dependence was implemented by collecting distance statistics from proteins of a given size-group (the training set). The potential function was optimized iteratively so as to render the training set proteins stable (i.e., after local minimization), while maintaining the smallest energy gap possible between native conformations and their locally minimized counterparts.

(d) Simulations of protein structures were carried out via a Monte Carlo plus minimization algorithm [2] along the lines proposed by Li and Scheraga [3], with a number of modifications to improve efficiency. The Monte Carlo code has been developed to run in parallel using the MPI protocol over a network of inexpensive personal computers.

(e) When the folding simulations were complete, the resulting structures were clustered and ranked according to total energy. A combination of ranking by energy and visual inspection was used to select the models to be submitted as predictions.

References:

[1] Casari, G., Sippl, M.J. (1992). Structure-derived hydrophobic potential. J. Mol. Biol. 224(3), 725-732.

[2] Eyrich, V. A., Standley, D. M. & Friesner, R. A. (1999). Prediction of protein tertiary structure to low resolution: Performance for a large and structurally diverse test set. Journal of Molecular Biology 288(4), 725-742

[3] Li, Z. Q. & Scheraga, H. A. (1987). Monte-Carlo-Minimization Approach to the Multiple Minima Problem in Protein Folding. Proceedings of the National Academy of Sciences of the United States of America 84(19), 6611-6615.

---

# roland-luethy , 309

number of submitted models: 13

**Three-dimensional structure prediction using simplified structure models and Bayesian block fragments.**

Roland Luethy and Jun Zhu

*Amgen Inc.*

*email:* rluethy@amgen.com

The method used for CASP4 has three main components: 1. a simplified representation of protein structures that can be locally modified. 2. A structure modification method based on selecting blocks from know 3D structures. 3. Evaluation of structures and optimization. The simplified models are based on a sequence of internal coordinates: the torsion angles between four consecutive Ca atoms and angles between three Ca atoms. In order to generate different structures blocks were randomly selected from database of structural blocks which was created in the following way: First a Bayesian sequence model of the target was created. Then the database of known structures filtered with an identity cutoff of 45% was searched for high confidence alignments. The ungapped blocks from these alignments were kept. Different structures were generated by randomly selecting blocks from this database and substituting them. To evaluate structures cartesian coordinates for the Ca and Cb atoms were reconstructed using constants for all distances and the angles needed to reconstruct the Cb positions. These structures were then evaluated using knowledge based potentials derived from know structures. The potentials used were a residue specific pair-wise distance potential, a residue specific number of contacts potential, a compactness function and a penalty for too close contacts.

---

# Brooks-Feig , 040

number of submitted models: 10

**Multiscale Modeling Protocol for ab initio Structure Prediction**

Michael Feig and Charles L. Brooks, III

*The Scripps Research Institute*
*email:* meikel@scripps.edu

We have used a novel multiscale modeling approach to predict protein structures ab initio. This approach combines fast lattice-based Monte Carlo sampling of a low-resolution side-chain based model with a more accurate energetic description of all-atom models reconstructed from the simple lattice models based on the CHARMM molecular mechanics force field.

For the lattice-based modeling we have used the MONSSTER program by Kolinski et al.
[A. Kolinski, J. Skolnick: Proteins (1998), 32, 475-494]
that implements a sophisticated empirical energy function for side-chain only (SICHO) protein models where each residue is represented by a single particle located at the respective side chain center of mass. Using this energy function combined with a Monte Carlo simulated annealing protocol as in MONSSTER it is often possible to fold up initial random chain configurations to structures in the vicinity of the native structure, especially if some restraints (from limited experimental data or contact map predictions) can be provided. However, it is usually difficult to correlate the

energetic score in the low-resolution model with the distance of the model from the native structure which is necessary to select structures closest to the native state from an ensemble of predicted conformations.

Better energetic discrimination is expected if a more accurate energy function for a more detailed protein model is used. This can be achieved by reconstructing all-atom models from the side-chain based models and minimizing under the influence of a full molecular mechanics force field. We have used the CHARMM19 force field combined with a generalized Born solvent approximation in the CHARMM program [B. Brooks et al.: J. Comp. Chem. (1983), 4, 187-217] for this purpose. Including a generalized Born solvent approximation improves the energetic description significantly over environments with constant or distance-dependent dielectric constants at moderate computational cost compared to much more expensive explicit solvent simulations.
An efficient reconstruction procedure has been developed by us recently [M. Feig, P. Rotkiewicz, A. Kolinski, J. Skolnick, C. Brooks: Proteins (2000), 41, 86-97] for the purpose of facilitating the transition to low-energy all-atom models from lattice-based models.

Following the approach described above by generating an ensemble of lattice-based structures with MONSSTER followed by all-atom reconstruction and minimization in CHARMM will not improve the resulting structures significantly - some small improvement often occurs during the CHARMM minimization - but allows a better selection of structures closer to the native conformation based on the CHARMM energy estimate. Sampling towards the native conformation can be further improved if the "best" structures, according to the CHARMM energy function, are used as starting conformations for a new cycle of short low-resolution lattice-based simulations that will then sample conformations around these "best" structures from the previous run. This effectively amounts to lattice-based sampling of low-resolution models biased by the all-atom CHARMM energy and compensates for deficiencies in the approximate energetic description of the low-resolution model while maintaining much more extensive conformational sampling than what would be feasible in all-atom simulations.

Finally, even when the CHARMM energy function is used the variance of energy values within a conformational class is relatively large compared to differences in average energies between different classes. Therefore, a more suitable approach to finding the "best" structures within an ensemble than simply sorting all structures according to their energy value would be to apply a clustering technique that separates different conformational classes and allows the comparison of average energies between clusters. The "best" structures can then be selected more reliably as the best structures within the best cluster.

In our CASP4 prediction efforts we used a multi-step protocol as outlined above. Initially, 2000 lattice-based simulated annealing runs were performed starting from random conformations (in some cases with weak restraints from contact map predictions from CAFASP2). Secondary structure predictions were also used to provide some bias accordingly during the lattice simulations. The resulting models were then reconstructed to all-atom structures, minimized in CHARMM and clustered according to the minimized all-atom conformations. The best structures of the best clusters according to the CHARMM energy function were then used as starting conformations for another round of 2000-3000 very short

lattice runs followed by all-atom minimizations and clustering.
3-4 of these refinement cycles were performed for each target
before representatives of the best clusters were inspected visually
and by computational measures using PROCHECK and WHATIF
[http://biotech.embl-ebi.ac.uk:8400] to select the most "protein-like"
structures as the final predictions.

---

# Dill-Ken-A , 139

number of submitted models: 9

**Using Local Homology Information in a Systematic Conformational Search for three CASP4 targets"**

Kaizhi Yue and Ken A. Dill

*University of California, San Francisco*
*email:* yue@zimm.ucsf.edu

We have submitted structure predictions for targets T0110, T0118 and T0091. We use a systematic conformational search program, Geocore, in which phi/psi angles and side chain dihedrals assume discrete values. We use a simplified potential that considers hydrophobic interactions, hydrogen bonding and vdW interactions. (For details, see " Folding proteins with a simple energy function and extensive conformational searching", K. Yue and K. A. Dill, Protein Sciences, volume 5, page 254-261) The conformations are represented at both united atom level and at the level of secondary structures. Secondary structure elements are treated as rigid bodies and packing between secondary structures are handled as rigid body docking (for details, see "Exploiting Constraints on Secondary Structure Packing in Protein Conformational Search", K. Yue and K. A. Dill, Protein Sciences, in press).

In the conformational search for CASP4, we have included an experimental module in which homologous conformational templates similar to I-sites libraries ( "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions", Simons, K. T. and Kooperberg, C. and Huang, E. and Baker, D.,J Mol Biol, 268:209-225,1997) are used. The templates are extracted from Protein Data Bank for chain segments of lengths 4 to 11 residues. This information is combined with secondary structure prediction result ("Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool". J. Levin, J. Garnier. Biochim. Biophys. Acta, (1988) 955, 283-295.) to determine the discrete choices for each residue position in the search.

Because of the limitation of time, only a small portion of the conformational space is sampled. The program collects around 400 high scoring (low energy) conformations as its output. Some of these high scoring conformations are subjected to energy minimisation in cartesian coordinates using the EEF1 force field ("Effective Energy Function for Proteins in Solution", T. Lazaridis and M. Karplus, Proteins, 35:133-152, 1999), and the lowest energy structures were submitted. The minimizer is developed by John Chodera.

---

# Wolynes , 032

number of submitted models: 68

**Ab initio Structure Prediction with Associative Memory Hamiltonians**

Corey Hardin,Michael Prentiss,Zadia Luthey-Schulten and Peter G. Wolynes.

*University of Illinois*
*email:* c-hardin@uiuc.edu

Our entries in the ab initio category of CASP4 were produced with the use of the associative memory energy function described in (1) and (2). At short to moderate sequence separations the energy function is an associative memory Hamiltoian (3)(4) constructed from a database of folding patterns contained in a standard set of known structures, and at long sequence separations it has the form of a simple pairwise contact potential. The structure of the CASP target sequence is obtained via simmulated annealing in a molecular dynamics simmulation. A second energy function is then used to search the low energy states from the molecular dynamics trajectory and select a final prediction.

The associative memory (AM) potential is based on correlations between a target's sequence and the sequence-structure patterns in a set of database proteins - the memories. Residue pairs in the target (ij) are associated with pairs in the memory (i'j') using a sequence-structure threading algorithm (5). In the case of ab initio predictions it is expected that the memory set will not contain any protein homologous to the target. Thus only fragmentary local in sequence patterns are expected to be found by the threading procedure. CASP4 targets were threaded against a subset of the Protein Database consisting of structuraly unique, well resolved x-ray structures less than 200 residues long and of the secondary structure class suggested by standard secondary structure prediction algorithms. In cases were searches of sequence databases or other bioinformatic techniques suggested a distant homolog, it was included in the database. The AM potential was then constructred from the highest scoring memories for each target. The paramters for all of our energy functions have been optimized using energy landscape theory (6).
In order to minmize the AM potential via molecular dynamics and keep the problem computationly tractable, we have developed a minimal model of the protein backbone (7). An initial molecular dynamics simulation was performed for each target, and then a set of low energy structures was assembled by performing a series of rapid quenches from around the expected folding temperature. For longer targets ( N > ~100), the initial simulation was often started from the structure of the highest scoring memory, in order to avoid topological problems which can arise from our limited backbone model. The target sequence was then threaded onto these structures using the same threading procedure used in the construction of the AM potential. The lowest energy structure was chosen as model 1, and structures with similar energies, if any, were submitted as additional models.

References

1. Hardin, C., Eastwood, M., Luthey-Schulten, Z., and Wolynes P. Associative Memory Hamiltonians for Structure Prediction without Homology: Alpha-Helical proteins. Proceedings of the National Academy of Sciences 2000 , in press.

2. Koretke, K., Luthey-Schulten, Z., and Wolynes P.,
Self-consistently Optimized Energy Functions for
Protein Structure Prediction by Molecular Dynamics.
Proceedings of the National Academy of Sciences
1998 , 95:2932-2937.

3. Friedrichs, M., and Wolynes, P. Toward Protein Tertiary Structure
Recognition by Means of Associative Memory Hamiltonians.
Science 1989. 246:371-373.

4. Friedrichs, M., Goldstein, R., and Wolynes, P.
Generalized Protein Tertiary Structure Recognition Using
Associative Memory Hamiltonians. Journal of MOlecular Biology
1991. 222:1013-1034.

5. Koretke, K., Luthey-Schulten, Z., and Wolynes P.,
Self-consistently Optimized Statistical Mechanical
Energy Functions for Sequence Structure Alignment.
Protein Science 1996. 5:1043-1059.

6. Onuchic, J., Luthey-Schulten, Z., and Wolynes P.,
Theory of Protein Folding: the Energy Landscape
Perspective. Annual Review of Physical Chemistry 1997.
48:539-594.

7. Hardin, C., Luthey-Schulten, Z., and Wolynes P.
Backbone Dynamics, Fast Folding, and Secondary Structure Formation
in Helical Proteins and Peptides. Proteins:Structure, Function,
Geneetics 1999. 34:281-294

---

# Yoon , 152

number of submitted models: 107

**Simulation of the Protein Folding Structures**

Chang N Yoon, Taesung Moon, Jin K Lee, Hyun J Kim

*Korea Institute of Science and Technology*
*email:* cody@kist.re.kr

To simulate the folding structures of a protein, we used a simple off-lattice model with the unified-residue point, which represents the alpha carbon of each amino acid in the protein model. This model has two angle variables, one for the angle between two consecutive virtual bonds, residues i to j and j to k, the other for the rotational angle of the virtual bonds consisting of residues i, j, k, and l. In order to generate the protein conformations the Monte Carlo method was used with the starting point of random-coil conformations. During this procedure the range of the i-j-k angle was limited between 60 to 150 degrees. Among the trajectory data obtained from the navigation through the

potential surface, about half of them were accepted and stored. The knowledge-based potential was used to obtain the potential energy surface. It was derived from the known protein structures. The total number of the accepted conformations was about 10E3 and the total steps for one run were about 10E8. Finally, all the conformations were clustered using the energy and cRMS between the alpha carbon traces. Then the obtained representative conformations were minimized with the potential energy.

---

# Bystroff , 055

number of submitted models: 251

**Fully automated ab initio protein tertiary structure prediction using HMMSTR and Rosetta**

Chris Bystroff, Yu Shao, Vestienn Thorsson, Kim Simons and David Baker

*Rensselaer Polytechnic Institute*
*email:* bystrc@rpi.edu

A publically available web-based server has been set up which combines the strengths of three previously-described methods. For CASP4, the inputs to the server were single sequences of the targets, although the server also allows several input formats of multiply aligned sequences. The prediction of tertiary structure is carried out in five steps:

(1) Psi-blast is used to obtain a multiple alignment, from which a sequence profile is calculated. e-value cutoff = 0.001

(2) A hidden Markov model, HMMSTR (Bystroff C., Thorsson V., Baker D. J. Mol Biol 301(1):173-90 2000) was used to predict Markov states for each position in the multiple alignment. HMMSTR is a hidden Markov model based on the I-sites Library of sequence-structure motifs. (Bystroff & Baker, J. Mol. Biol. 281:565-577(1998)) HMMSTR Markov states represent positions within local structure motifs, expressed as backbone angles (phi,psi) with confidences. A profile of Markov state probabilities is output. If the confidence of a single position backbone angle prediction exceeded 0.70, then that residue was constrained (fixed) to those angles for the remainder of the process. However, if more than one-third of the residues had a confidence in excess of 0.70, then the confidence cutoff was raised until at most one-third of the residues were fixed.

(3) A set of 25 nearest-neighbor fragments was found for each 3 and 9-residue segment in the target by comparing the predicted Markov state profiles to the known Markov states in the database of all known proteins. This is the moveset for the Monte Carlo conformational search.

(4) The fragments are assembled by Monte Carlo Fragment Insertion as performed by Rosetta (Simons et al, PNAS 1998). The starting conformation is a randomly-chosen set of fragments that span the protein. At each step a fragment is chosen at random from the moveset and is inserted, then the backbone angles are converted to 3D coordinates. The move is accepted or rejected based on teh Rosetta energy function. Rosetta minimizes a knowledge-based energy function derived from the database of known structures. The energy is inversely related to the conditional probabilities of pairwise

secondary structure packing geometries and other structural characteristics, described in the above reference. Long target sequences were divided into overlapping segments having at most 36 un-fixed residues each and having 18 un-fixed residues of overlap with neighboring fragments. Each fragment was simulated separately and the 15 lowest-energy conformers were kept for the next step.

(5) Predictions of increasing length were generated by combining the 15 predictions from each of two overlapping segments. A fusion of two predictions was tested for every overlapping position and every fragment pair. The five fused structures with the lowest energy, measured the same way as in Rosetta, were kept. Overlapping fragment preditions were extended in this way until the fragments were full length. The server returns the coordinates for the five lowest energy predictions. They are generally very similar because of this fusion process. (honduras.bio.rpi.edu/~isites/hmmstr/server.html)

---

# Ho-Kai-Ming , 375

number of submitted models: 63

**Ab initio prediction of protein structures using a coarse-grained model**

James R. Morris, Kai-Ming Ho, Cai-Zhuang Wang, Drena Dobbs, Tzu-Liang Chan, Young-Ok Im,Josh Koch, Neil Voss, Dion Harmon

*Iowa State University and Ames Laboratory*
*email:* jrmorris@ameslab.gov

As part of the CASP4 experiment, we have made a number of ab initio predictions for protein structures. The key to our work is a coarse grained mechanical model of proteins, similar in spirit (though very different in details) to other simple models (Scheraga, Domany). Our model is a "two bead" model: each amino acid is characterized by a backbone bead and a side-chain bead. These two beads and adjacent backbone beads are connected harmonically. The model also includes residue-specific second-nearest backbone bead interactions (allowing for secondary-structure tendencies of the residues); hydrogen bonding interactions (saturable weak interactions that allow for alpha helix and beta sheet formation); residue-backbone repulsion; and residue-residue interactions. These latter interactions are currently very simple: all interactions include a repulsive portion at short distances, while two hydrophobic residues or two oppositely charged residues will include an attractive tail. Late in the contest, the potential was modified to give preference to right-handed helices.

We are currently working on more accurate and detailed residue-residue interactions. Our model successfully stabilizes the native structure of a number of all alpha and all beta proteins, ranging in length from 80-250 residues. These test were performed by beginning in the native structure, then relaxing and annealing the structure. Subsequent examination showed little major change in the backbone topology; RMSD of the C-alpha positions was typically on the order of 6 angstroms. For the all beta proteins, however, blind searches (using techniques described below) revealed that there existed all alpha structures with lower free energy. This indicates that further work on the model must

focus on correctly balancing between the free energies of helices vs. sheets. Similarly, mixed alpha/beta native structures were found to be unstable, and would transform to all alpha structures.

For the prediction, we used two approaches. The first of these is a variation of simulated annealing. We would perform this in two stages. In the first stage, repulsive interactions and interactions associated with secondary structure were initially turned off, allowing for a fast collapse and formation of a hydrophobic core. These interactions were then "ramped up" to full strength during this initial stage. The second stage was then a conventional simulated annealing approach - conventional molecular dynamics with a slow cooling schedule. The second of the approaches was a parallel, genetic algorithm version of this same approach, with the initial interactions during the first stage guided by a guess at the contact list. After a number of geometries had undergone this process (typically 16-32 geometries run in parallel), the contact matrix of the lowest energy final structure was chosen, and a series of geometries were again generated. This process was repeated 5-10 times. We found that the latter process did not significantly improve the results, compared with a more direct approach of performing very long simulated annealing runs on several geometries. These approaches have performed reasonably in a number of test cases, including "blind" predictions using sequences from known structures.

Given the results on test cases, our work on the CASP4 experiment focussed on sequences whose secondary structure was predicted to be all alpha, with some work on those which had beta sheet regions predicted with high confidence (using PHD in all cases). Those with predicted beta sheet regions had hydrogen bond assignments made to bias the system towards that secondary structure. We also examined cases where a portion of the structure had high sequence homology to a known protein structure, but where a significant segment of the sequence (60 residues or more) had no such homology. We then attempted to predict the non-homologous region, while keeping the remaining region fixed to that of the known homologous structure.

---

# Gerloff , 003

number of submitted models: 13

**Incorporation of Human-Derived Constraints from Active/Functional Site Models in Protein Tertiary Structure Assembly**

Zeti A. M. Hussein, Melanie H. McCarthy-Troke, Bernard J. Mitchell, Cairan R. E. Duffy, Siu-Wai Leung, G. M. Cannarozzi and Dietlind L. Gerloff**: corresponding author

*University of Edinburgh*
*email:* D.Gerloff@ed.ac.uk

We have submitted tertiary structure predictions for eight CASP4 target proteins (see below) in order to investigate the potential of knowledge and/or predictions about functional sites on these proteins for being used in combination with established methods. The prediction categories in which our predictions will be considered, as well as their assigned degrees of difficulty, are likely to vary.

The applicable categories could range from ab initio modelling (tertiary assembly of predicted secondary structure elements) over fold prediction to threading alignments. Two targets for which the fold prediction was very "obvious" (T0100 and T0101) might be designated comparative modeling targets by some assessors. Accordingly, we are submitting this method abstract to all three categories.

Besides our emphasis on formulating distance and/or geometrical constraints for our models based on functional site knowledge, or prediction, the only unifying link between our submissions is the use of predicted Surface/Interior/ActiveSite/Parse positions (termed SIAP-predictions in the following) according to the approach by Benner and Gerloff, which is most conveniently described in:

S. A. Benner, G. M. Cannarozzi, D. Gerloff, M. Turcotte and G. Chelvanayagam (1997). Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. Chem. Reviews 97, 2725-2843.

Below, we provide more explicit information regarding the methods used in each of the predictions. We wish to emphasize that this information is by no means to be considered redundant with the information given in the header text of individual prediction submissions made to CASP4. Specifically in the header text, we attempted to highlight the most relevant clues that led us to each prediction (or, rather, those that we thought to be relevant without knowledge regarding the experimental structures), and from which underlying functional assumptions these clues were derived. In this way, we hope to provide a transparent account of our prediction strategy from which we will be able to learn which of our assumptions were valid, and could, potentially, be used more generally in tertiary structure prediction after further validation on known structures.

T0086 - Chorismate Lyase - UBIC
---

This prediction rests primarily on the assembly of secondary structures around a putative active site. FUNCTIONAL SITE PREDICTION included speculations regarding a plausible mechanism for the catalyzed reaction, and the roles of conserved functional residues. Incidentally, using one monomer of the trimeric Bacillus chorismate mutase structure as scaffold, and assuming an approximately equivalent location of substrate binding led to a moderately satisfactory model.

FOLD PREDICTION: assembly of predicted secondary structures + pathway neighbors were examined preferentially. FOLD AGREEMENT WITH CAFASP2: no. INCREASED THE NUMBER OF HOMOLOGOUS SEQS: yes (genome projects via PEDANT). SECONDARY STRUCTURE PREDICTION: mainly from CAFASP2 servers. SIAP-PREDICTION: yes. CONFIDENCE IN TOPOLOGY PREDICTION: medium.

T0087 - PPase - PPX1
---

This submission consists of manual fold recognition prediction(s) and threading alignments for two predicted domains in target T0087. FUNCTIONAL SITE PREDICTION was possible only for the first domain. It centers in the predicted location, and ligands, for a Mn2+ ion which is relevant for catalytic function, and the compatibility of the predicted fold and threading alignment with our expectations with respect to composition and geometry of a polyphosphate hydrolase site. The predictions use the fold of the Thiamine(-PP) Binding Domain as the parent fold. Interestingly our prediction rules out a Rossmann fold for the first domain.

FOLD PREDICTION: combinatorial assembly of predicted secondary

structures + functional hypothesis (Mn2+-site; pyrophosphate-binding).
FOLD AGREEMENT WITH CAFASP2: no. INCREASED THE NUMBER OF
HOMOLOGOUS SEQS: yes (genome projects via SAMT99 (CAFASP2 #15)).
SECONDARY STRUCTURE PREDICTION: mainly from CAFASP2 servers.
SIAP-PREDICTION: yes. CONFIDENCE IN TOPOLOGY PREDICTION: high (first
domain) + medium (second domain).

T0094 - Cyclic Phosphodiesterase - CPDase
---
This submission is a tentative manual threading alignment,
assuming that the weak sequence similarity to histone acetyl-
transferase (1ygh), as detected by pdb-blast (CAFASP server #1),
bears any significance. FUNCTIONAL SITE PREDICTION included our
emphasis to group together most of the functional residues that
were conserved in two putative homologs, and to fulfill the spatial
requirements for two disulfide bridges. The difficulties with
fitting the predicted secondary structure, as well as the different,
inferred, active site location of target vs. template, could indicate
that the assumptions are in fact invalid, and our prediction false.

FOLD PREDICTION: BLAST against PDB (CAFASP2 server #1).
FOLD AGREEMENT WITH CAFASP2: yes. INCREASED THE NUMBER OF
HOMOLOGOUS SEQS: yes? (Geobacter genome, distant homolog?, via NCBI).
SECONDARY STRUCTURE PREDICTION: mainly from CAFASP2 servers (esp. #7).
SIAP-PREDICTION: yes. CONFIDENCE IN TOPOLOGY PREDICTION: medium-low.

T0098 - Spo0A C-terminus - SP0A [solved: 1FC3]
---
Our prediction represents one possible arrangement of alpha-helices that
would seem compatible with our secondary and SIAP-predictions, as well as
limited FUNCTIONAL INFORMATION & HYPOTHESES. The latter included knowledge
of DNA-binding specificity at a bi-partite regulatory site, albeit with
uncertainty regarding the number of DNA-contact sites per protein domain.
We [falsely!] chose to presume two sites, and proposed a fold containing
two HTH-motifs seen in an earlier CASP-target (T0079). A possible Zn2+-
binding cluster of residues seemed to support our prediction, while
protein-to-DNA sequence-sequence correlation (1) was pointing towards
the correct location for the HTH motif, albeit at sub-significant scores.

FOLD PREDICTION: functional hypothesis (DNA-binding; 2x) + assembly of
predicted secondary structures + deja vu. FOLD AGREEMENT WITH CAFASP2:
yes (HTH-structural motif) + no (2 HTH-motifs and relative orientation).
INCREASED THE NUMBER OF HOMOLOGOUS SEQS: yes? (M.leprae genome, distant
homolog?, via NCBI; variation between sequences very limited).
SECONDARY STRUCTURE PREDICTION: CAFASP2-servers + Benner&Gerloff.
SIAP-PREDICTION: yes. CONFIDENCE IN TOPOLOGY PREDICTION: medium.

T0100 - Pectin Methylesterase - PMEA [solved: 1QJV]
---
This submission is a manual threading alignment using a single-stranded,
right-handed beta helix as the parent fold, as this was suggested by the
majority of CAFASP2-submissions for this target. The most important anchors
for the threading alignment came imposed by PUTATIVE DISULFIDE BRIDGES in
other members of the T0100-family, and FUNCTIONAL SITE PREDICTIONS
(of putative active site residues, their relative locations and their
locations with regard to the putative pectin binding site). These clues
were used in combination with known as well as predicted, structural
constraints that apply specifically to repetitive folding topologies
of this kind (reviewed e.g. in (2)). [In retrospect, all of our
assumptions seem to have proven valid; however, we overpredicted an
insertion after approx. 290aa, which corresponded, in reality, to a full
turn of the beta-helix].

FOLD PREDICTION: obvious. FOLD AGREEMENT WITH CAFASP2: yes.
INCREASED THE NUMBER OF HOMOLOGOUS SEQS: no.
SECONDARY STRUCTURE PREDICTION: not used. SIAP-PREDICTION: yes.
CONFIDENCE IN TOPOLOGY PREDICTION: high.

T0101 - Pectate Lyase - PELL
---
This submission is a manual threading alignment using a single-stranded, right-handed beta helix as the parent fold, as this was suggested by the majority of CAFASP2-submissions for this target. In contrast to T0100, only little help could be obtained through speculative disulfide bridges. Instead the alignment is based on FUNCTIONAL SITE PREDICTION of candidate residues forming a calcium-binding site which is relevant for enzymatic activity in most members of this family of pectate lyases. Incidentally, our hypothesis would place the calcium ion at a similar location as it is found in co-crystal structures of other pectate lyases (e.g. 1bn8). Further help came from fold-specific structural anchors, e.g. N-ladders (2).

FOLD PREDICTION: obvious. FOLD AGREEMENT WITH CAFASP2: yes.
INCREASED THE NUMBER OF HOMOLOGOUS SEQS:yes (genome projects via SAMT99 (CAFASP2 #15)). SECONDARY STRUCTURE PREDICTION: not used.
SIAP-PREDICTION: yes. CONFIDENCE IN TOPOLOGY PREDICTION: high.

T0105 - Protein Sp100b - SP100
---
Our prediction for is an ab initio assembly of predicted secondary structure elements (see separately submitted SS prediction), based primarily on two FUNCTIONAL SITE PREDICTIONS: (i) a putative cluster of Cys residues (Zn-binding?) in all homologues but drawn from different locations in the sequence; (ii) DNA-binding via a recognition helix. Constraints were derived from these prediction and used in combination with the predicted amphiphilic or internal characters of the beta strands. The result is a strongly twisted, half open and concave barrel structure which can be viewed almost as a cyclic permutated SH3-barrel. Combinatorial analysis suggests there might be other plausible topologies. Producing an approximate coordinate model representing this topology prediction was extremely difficult because of beta-strand twisting and bending. We hope that the intended topology is discernable, at least...

FOLD PREDICTION: combinatorial assembly of predicted secondary structures (ab initio). FOLD AGREEMENT WITH CAFASP2: no.
INCREASED THE NUMBER OF HOMOLOGOUS SEQS: yes (literature! (3)).
SECONDARY STRUCTURE PREDICTION: CAFASP2 servers + Benner&Gerloff.
SIAP-PREDICTION: yes. CONFIDENCE IN TOPOLOGY PREDICTION: medium-low.

T0121 - MalK - MALK [C-terminal 135aa only]
---
Our submissions include two different manual threading alignments (models 1 and 2) for the C-terminal 135 residues onto a duplicated OB-fold template. Besides this parent structure, we would also consider it possible to use a minimal Ig-like beta sandwich-core (strand connections A-(+3)-B-(-1)-C-(-1)-D-(+3)-E-(+1)) but preferred to concentrate on the alignment issue with 1b9m_A as the preferred parent structure. The function of the domain is basically unknown. We tried to derive FUNCTION SITE PREDICTIONS based on sequence conservation and literature (limited information about muta-genesis experiments, extracted from (4) and many other publications), e.g. an epitope function for the strongly conserved "GIRPED" sequence. In the alignment for model 1, we allowed these reflections to influence the sequence-to-structure alignment; the alignment for model 2 is based on sequence similarity and SIAP-prediction (and looks quite plausible, too).

FOLD PREDICTION: 2 plausible alternatives from CAFASP2 servers + combinatorial assembly of predicted secondary structures.

FOLD AGREEMENT WITH CAFASP2: yes (CAFASP FFAS server #10, for the
fold alternative used in the submissions). INCREASED THE NUMBER
OF HOMOLOGOUS SEQS: yes (genome projects via SAMT99 (CAFASP2 #15)).
SECONDARY STRUCTURE PREDICTION: CAFASP2 servers + Benner&Gerloff.
SIAP-PREDICTION: yes. CONFIDENCE IN TOPOLOGY PREDICTION: medium.

References used in individual predictions:
---
(1): M. Suzuki, S. E. Brenner, M. Gerstein and N. Yagi (1995).
Protein Eng. 8, 319-328.
(2): J. Jenkins, O. Mayans, R. Pickersgill (1998).
J. Struct. Biol. 122, 236-246
(3): T. J. Gibson, C. Ramu, C. Gemund and R. Aasland (1998).
Trends Biochem. Sci. 23, 242-244
(4): G. Schmees, A. Stein, S. Hunke, H. Landmesser, E. Schneider (1999).
Eur. J. Biochem. 266, 420-430

---

# fain , 469

number of submitted models: 30

**Prediction of structure of Alpha-Helical proteins.**

Boris Fain

*Stanford University*
*email:* bfain@stanford.edu

The procedure for prediction was as follows:

1) The sequence was submitted to the PhD secondary structure
server. Proteins that have none or very little beta-sheet
components were chosen for prediction.

2) Several possible secondary structure configurations were
chosen. (this is the step with the most human intervention).
The PhD prediction was chosen, and segments with low confidence
subsets were broken at that point to generate alternate
starting secondary structure assignments.

3) Possible geometric arrangements of the helices were generated.
Subsequently the loops (and, alas, endloops) were added
using the program Segmod (written by M. Levitt). This step
was a fast and accurate way to produce loops; its drawback
is it's creativity with endloops.

4) The potential structures were scored by one function - an
optimized burial score. The details of optimization are
in press, IBM systems journal, special life sciences edition.
(a fuller version is submitted to JMB)
In brief, the function is optimized using 100 sets of decoys
for 100 sequences. The function measures ONLY the number of
neighbours of each residue (a neighbour has a CB within
10A). The trick is that the functional
form of the potential is completely arbitrary; this seems

to improve things a great deal.

Hydrogen bonds were not considered, which will probably result
in collapsed predictions for long extended helical bundles.

REMARKS:  there were many areas where I could have improved
the procedure; e.g. include other scoring functions. However
my main purpose in entering the experiment is to objectively
test the methods that I have developed in the past year.

---

# Chandonia-Cohen , 150

number of submitted models: 68

**New Methods for Accurate Prediction of Protein Secondary Structure**

John-Marc Chandonia and Fred E. Cohen

*UCSF*
*email:* jmc@cmpharm.ucsf.edu

The secondary structure predictions submitted to CASP4 were generated
by Pred2ary (1).  Exactly the same predictions were submitted by
my group and as server predictions for CAFASP-2.  The algorithm
used to generate secondary structure predictions was used exactly
as published (2), so I've included the abstract from that paper.

A primary and a secondary neural network are applied to secondary
structure and structural class prediction for a database of 681
nonhomologous protein chains. A new method of decoding the outputs of the
secondary structure prediction network is used to produce an estimate
of the probability of finding each type of secondary structure at every
position in the sequence. In addition to providing a reliable estimate of
the accuracy of the predictions, this method gives a more accurate Q3
(74.6%) than the cutoff method which is commonly used. Use of these
predictions in jury methods improves the Q3 to 74.8%, the best available
at present. An estimate of the overall Q3 for a given sequence is made
by averaging the estimated accuracy of the prediction over all residues
in the sequence. As an example, analysis is applied to the target
b-cryptogein, which was a difficult target for ab initio predictions in
the CASP2 study; it shows that the prediction made with the present
method (620f residues correct) is close to the expected accuracy (66%)
for this protein. The larger database and use of a new network training
protocol also improve structural class prediction accuracy to 86%,
relative to 800btained previously. Secondary structure content is
predicted with accuracy comparable to spectroscopic methods such as
vibrational or electronic circular dichroism and Fourier transform
infrared spectroscopy.

(1)   See the web site  http://www.cmpharm.ucsf.edu/~jmc/pred2ary/
 for information and a free download of the program.

(2) Chandonia & Karplus, "New Methods for Accurate Prediction of Protein
 Secondary Structure", PROTEINS 35, 293-306, 1999.

Supported in part by the NIH (1 F32 HG00200).

# Pred2ary/Chandonia , 151

number of submitted models: 42

**New Methods for Accurate Prediction of Protein Secondary Structure**

John-Marc Chandonia and Fred E. Cohen

*UCSF*
*email:* jmc@cmpharm.ucsf.edu

The secondary structure predictions submitted to CASP4 were generated by Pred2ary (1). Exactly the same predictions were submitted by my group and as server predictions for CAFASP-2. The algorithm used to generate secondary structure predictions was used exactly as published (2), so I've included the abstract from that paper.

A primary and a secondary neural network are applied to secondary structure and structural class prediction for a database of 681 nonhomologous protein chains. A new method of decoding the outputs of the secondary structure prediction network is used to produce an estimate of the probability of finding each type of secondary structure at every position in the sequence. In addition to providing a reliable estimate of the accuracy of the predictions, this method gives a more accurate Q3 (74.6%) than the cutoff method which is commonly used. Use of these predictions in jury methods improves the Q3 to 74.8%, the best available at present. An estimate of the overall Q3 for a given sequence is made by averaging the estimated accuracy of the prediction over all residues in the sequence. As an example, analysis is applied to the target b-cryptogein, which was a difficult target for ab initio predictions in the CASP2 study; it shows that the prediction made with the present method (620f residues correct) is close to the expected accuracy (66%) for this protein. The larger database and use of a new network training protocol also improve structural class prediction accuracy to 86%, relative to 800btained previously. Secondary structure content is predicted with accuracy comparable to spectroscopic methods such as vibrational or electronic circular dichroism and Fourier transform infrared spectroscopy.

(1) See the web site http://www.cmpharm.ucsf.edu/~jmc/pred2ary/ for information and a free download of the program.

(2) Chandonia & Karplus, "New Methods for Accurate Prediction of Protein Secondary Structure", PROTEINS 35, 293-306, 1999.

Supported in part by the NIH (1 F32 HG00200).

# Beverdige , 522

number of submitted models: 1

Y. Liu and D. L. Beveridge

*Wesleyan University*
*email:* yliu@wesleyan.edu

```
In this work we used the conformational optimization method based on multiple
copy  simulated annealing in torsional space and modified Generalized Born
Approximation.  Pair wise approximation GBA and analytical calculation of
solvent accessible area are  modified to include force field atom types
consistent with  AMBER united atom force field.  The multiple copy simulated
annealing in torsional space algorithm is devised for  molecular
conformational optimization.
```

---

# Akiyama , 252

number of submitted models: 7

**Ab initio Protein Structure Prediction in Water using an Accurate Parallelized Tree-code Molecular Dynamics Program (MolTreC) with Some Artificial Acceleration Terms for Folding**

Yutaka Akiyama, Tamotsu Noguchi, Kiyotaka Misoo

*Electrotechnical Laboratory (ETL), Japan*
*email:* yakiyama@etl.go.jp

```
An original molecular dynamics simulator called the MolTreC is used.
The MolTreC was developed for MD simulation of proteins in water
without cutoff approximation on Coulomb interactions. The MolTreC is
fully parallelized and is running on SGI Origin, Hitachi SR2201 and
Linux PC clusters. The program can rapidly calculate the Coulomb
potentials of all atom pairs based on Barnes-Hut tree algorithm.
The CASP4 submitted models were mainly calculated on SR2201 (256pu).
For CASP4 we added a special functionality on the MolTreC, that is,
introduction of artificial energy terms to accelerate secondary
structure (especially helix) formation. The terms are interactively
controlled through the time course of a simulation, based on the
DSSP analysis and/or insights by human experts. The detail of
introduced artificial terms depends to a specific simulation case.
The following paper describes the basic function of the MolTreC.
K. Misoo, Y. Akiyama, Y. Shizawa and M. Saito:
"Development of Molecular Dynamics Programs for Protein with a
Parallelized Barnes-Hut Code", Proc. of the Fourth International
Conference on High-Performance Computing in Asia-Pacific Region
```

(HPC-Asia 2000), Vol.2, pp.1103-1111 (2000).

# Sorry, this is just tentative abstract submission
# and is cut & pasted from our METHOD field in the submission form.

---

# PDG-contact-pred/Valencia , 424

number of submitted models: 7

**Residue contact prediction based on correlated mutations**

Florencio Pazos Alfonso Valencia

*National Center for Biotechnology (Spain)*
*email:* valencia@cnb.uam.es

Our contact prediction server takes as input a single sequence and predicts contacts between residues in the three-dimensional structure of the protein. Contact predictions are based on an "old" application of "correlated mutations" (Goebel et al. 94; Olmea & Valencia, 97; Pazos, Olmea & Valencia, 97) and a new way of calculating tree-determinat residues (Casari, Sander & Valencia, 95; Andrade et al., 97: Pazos et al., 97b; Del Sol, Pazos & Valencia -in prep.-). Contacts among hydrophobic and conserved residues are also extracted, they can be considered as a base-line for predictions. For the calculation of correlated mutations, tree-determinants and conserved residues a multiple sequence alignment is built in the following way:

1. BLAST (Altschul et al., 97) is used to search for homologous proteins in a non-redundant database.
2. CLUSTALW (Thompson et al., 94) is used to align those homologous sequences.
3. The alignmet is filtered to avoid redundance: pairs of sequences more similar than a given value are reduced to one sequence.
4. Divergent sequences (very distant homologous) are eliminated.
5. Small fragments are also eliminated from the alignment.

Contributors
- O. Olmea. For the research on correlated mutations carried out in the group.
- J. M. Fernandez and F. Abascal. For setting up the BLAST system and for maintaining the databases.
- F. Abascal. For the program used to retrieve sequences by ID.

References
- Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997). Nucleic Acids Res. Sep 1; 25(17):3389-402. Review.
- Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTALW: improving the sensitivity of progresive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acid Res., 22, 4673-4680.
- G?el, U., Sander, C., Scheneider, R., Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins 18:309-317.
- Olmea O, Valencia A (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Folding & Design 2, S25-S32
- Pazos, F., Olmea, O. and Valencia, A.(1997). A graphical interface for

correlated mutations and other structure prediction methods. CABIOS. 13(3):319-321.
- Casari, G. Sander, C., Valencia, A. (1995) A method to predict functional residues in proteins. Nature S. B. 2: 171-178.
- Andrade M A, Casari G, Sander C, Valencia A (1997) Classification of protein families and detection of the determinat residues with a self-organizing neural network. Biol. Cybern. 76, 441-450.
- Pazos F, Sanchez-Pulido L, Garc?-Ranea J A, Andrade M A, Atrian S, Valencia A (1997). Comparative analysis of different methods for the detection of specificity regions in protein families. In D. Lundh. Olsson, B, Narayanan, A. (Eds.), Biocomputing and Emergent Computation (pp. 132-145). Singapore, New Jersey, London, Hong Kong: World Scientific.

---

# FCLD , 489

number of submitted models: 22

**Assessing Prediction Quality of Low to High Resolution Models**

K. W. Foreman, J. Chodera, M. R. Lee, and K. A. Dill

*University of California, San Francisco*
*email:* kw@francisco.compchem.ucsf.edu

We tested a compound ab initio structure prediction algorithm. This algorithm uses a simple energy function designed to give rapid, approximate predictions of protein fold. The top predicted structures of the quick but inaccurate potential were then handed to a potential considered to be highly accurate for structural refinement and selection. In more detail, we pass the given sequence through a preprocessing step, where consistent structural themes from the current PDB are extracted and roughly fixed in the model. The model uses both canonical side-chain conformations and omega angles, thus allowing only phi/psi to change. The fixed portions of the structure allow motion, but only in a coordinated fashion which morphs between the structures extracted from the PDB. This reduction in the number of degrees of freedom is essential for considering moderate sized proteins (<200 residues) in a reasonable amount of time (<1 day). Once the preprocessing is complete, we search over a folding landscape for the global minimum which should correspond to the native state if the potential were accurate. The potential consists of half springs (either attractive or repulsive) that contribute to a van der Waals attraction and repulsion, long range hydrophobic collase, short range hydrophobic collapse, electrostatic attraction (including hydrogen bonding) and repulsion. An additional term is available to include known constraints such as disulphide bonds. The weights for these terms roughly follows those in Yue and Dill [Protein Science 5:254-261 (96)]. Finally, a penalty for entering disallowed regions of the Ramachandran map is applied for relevant phi/psi pairs [Dill et al., I. M. Bromze et al. (eds.), Developments in Global Optimization, 217-234 (97)]. The search over the landscape produced by this potential is performed by the CGU algorithm [Foreman et al., in preparation;

Foreman et al., J. Comput. Chem. 20:1527-1532 (99)]. The CGU uses all known minima on the landscape to select a smaller region of the landscape in which to search. If the landscape is funnellike, then the new search region is likely to contain the global minimum. The refinement stage employed mostly the EEF1 potential of Lazaridis and Karplus [Proteins 35:133-152 (99)] or very rarely, the MMPB/SA method of Lee et al. [Proteins 39:309-316 (00)]. Simulations were performed from between 10 and 30 ps followed by minimization of the top ten scoring structures in the simple potential. The best scoring structure was then submitted.