

MEETING ON CRITICAL ASSESSMENT OF TECHNIQUES FOR PROTEIN STRUCTURE PREDICTION

PREDICTION METHOD ABSTRACTS

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Asilomar, December 4 - 8, 1994

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Tim Cardozo, Maxim Totrov, Ruben Abagyan

Structural Biology, Skirball Institute of Biomolecular Medicine, New York University Medical Center, Tel. 212-2637048, FAX. 212-2638951, E-mail: abagyan@mcbl-18.med.nyu.edu

COMPARATIVE MODELING PROCEDURE - ABSTRACT

Five homology models were submitted for the first Critical Assessment of Techniques for Protein Structure Prediction Meeting in Asilomar, CA this December. The meeting has been organized around a blind prediction contest in which modelers from around the world have been provided with sequences and database information for molecules which were solved this year. Three categories were delineated: threading, ab initio and comparative (homology) modelling. This meeting represents the first organized assessment of the various methods for structure prediction.

METHOD- The global free energy optimization method used was the Biased Probability Monte Carlo (BPMC) conformational search using the ICM program (Abagyan and Totrov, J. Mol Biol., 1994; Abagyan and Totrov, Kuznetsov J. Comp. Chem., 1994). The steps in the procedure were as follows:

- 1) A multiple sequence alignment with solved structures was used as the starting point. The starting object consisted of the prediction sequence assigned standard geometry and regularized to the nearest related solved structure. The alignment and visual inspection revealed loop regions in which either insertions or deletions were present.
- 2) A model representing the starting structure with each of these loops deleted was constructed and optimized to provide a low energy environment against which loops or terminal extensions would be predicted. In the case of targets in which no insertions or deletions were present the modeling procedure ended here. Optimization at this step consisted of:
 - sampling of non-conserved side chains by the BPMC procedure followed by energy minimization. In other words, the backbone was fixed to the homologous side chain and the conformational space searched consisted of the aggregate of non-conserved side chains.
 - minimization of van der Waals clashes allowing small movements in the backbone from the homologous coordinates.
- 3) The loops were re-inserted into this optimized fragment and modeled by a local deformation BPMC procedure. Each loop was added individually and modelled separately in the absence of the others.

4) The best conformation from each loop modeling simulation was assigned resulting in a complete protein with each of the loops predicted. In the case of inter-acting loops (i.e. antibody combining site). Sequence conservation and visual inspection of related structures were used to select loops for fixation against which the least reasonable/conserved loop would be modelled. If the loops did not interact significantly, step 5) was immediately implemented.

5) A final optimization of the same nature as that in step 2 was performed using as the conformational space all the nonconserved side chains plus those in the loop regions. Energy was the discriminating function throughout the procedure, including terms for van der Waals interactions, hydrogen bonding, electrostatics, torsional energy and solvation energy (Abagyan and Totrov, J. Mol Biol., 1994).

SCREENING LIBRARY- Protein Data Bank

Submissions:

hpr.predict (His-phosphotransferase)
nm23h2.predict (Nuc-diphosphate kinase)
antibody.E5.2.pred
p450eryf.predict (cytochrome)
crabp.predict (retinoic acid binding)

The ETH Prediction Group

Michael B. Bolger and Sujit Basu

USC School of Pharmacy, 1985 Zonal Ave. PSC 700, Los Angeles, CA 90033

STRUCTURE PREDICTION OF E5.2 (ANTI-IDIO TYPE ANTIBODY TO ANTIBODY D1.3)

1. SEQUENCE ALIGNMENT AND ANALYSIS

The amino acid sequences of the variable region light and heavy chains of E5.2 were the starting points for the structure prediction. These sequences were compared with the Fab fragment amino acid sequences of known immunoglobulins. Sequence homology was checked using a multiple sequence alignment routine from the program PCGENE (Intelligenetics, San Jose, CA). Framework and hypervariable regions were determined by comparing E5.2 to known structures. The best X-ray structure for homology modeling was determined by comparing E5.2 to known canonical structures. The best starting structure for comparative homology modeling was determined to be mouse IgG Fab fragment R19.9 (anti-arsenate,

1FAI). The two protein sequences, E5.2 and R19.9, show a considerable degree of homology, 93.5% for light chain, and 58.7% for heavy chain. The canonical structure groups (Chothia et al., Nature, 342:877, 1989) for the six hypervariable regions were: L1 Group 2, L2 Group 1, L3 Group 1, H1 Group 1, H2 Group 2, H3 Not available.

2. COMPARATIVE HOMOLOGUE MODELING OF E5.2

All framework residues were substituted without modification of 1FAI protein backbone (Bolger et al. Methods in Enzymol., 203:21 1991). Likewise, hypervariable regions (L1, L2, L3, H1, and H2) from E5.2 were equal in length and canonical structure to 1FAI backbone and were not modified. Hypervariable region H3 was 2 residues shorter for E5.2 than for 1FAI. Consequently, the backbone for H3 of 1FAI was shortened by removing Glycine #102 H and Tyrosine #109 H using the program Hyperchem (Hypercube, Inc., Canada). A new bond was formed and the hypervariable loop alone was subjected to energy minimization to adjust the bond lengths. The whole structure of E5.2 was subjected to energy minimization using the program AMBER 4.0 (UCSF). The energy minimized structure was converted to PDB format and submitted.

3. DOCKING OF E5.2 ANTI-IDIOTYPE WITH D1.3 (ANTI-LYSOZYME)
Several modes of interaction are possible for binding of E5.2 to D1.3. We assumed that E5.2 hypervariable loops could recognize public or private epitopes of D1.3. Therefore, we attempted to use DOCK (Kuntz, et al., J. Mol. Biol. 161:269, 1982) to produce a family of geometrically plausible binding interactions. First, solvent excluded molecular surfaces were created for just the hypervariable region of E5.2 ("receptor") and the entire D1.3 Fv region ("ligand"). Then spheres of both receptor and ligand were generated using SPHGEN. The resulting set of dock spheres were too large for the program DOCK 2.0 to handle. Consequently, the spheres from SPHGEN were visualized using MIDAS (UCSF) and edited to suit DOCK 2.0. Also, the entire D1.3 Fv region was too large to handle with DOCK 2.0. Therefore, we attempted to "dock" the hypervariable regions. The resulting "docked" macromolecules were less than satisfactory. Only a single chain of D1.3 made contact with E5.2 (see poster at meeting for details). Reasons for the failure of DOCK 2.0 to work in our hands may be due to a lack of familiarity with the details of the software or it could be related to the fact that DOCK was designed to handle small molecule docking rather than macromolecular surfaces.

Marc Delarue

Unité d'Immunologie Structurale, Institut Pasteur, 25 rue du Dr. Roux, 75015 Paris, France (delarue@pasteur.fr).

Patrice Koehl

UPR 9003 (Cancerogenèse et Mutagenèse Moléculaire et Structurale) E.S.B.S. Blvd S. Brant, 67400 Illkirch-Graffenstaden, France (koehl@sulawesi.u-strasg.fr)

We have sent essentially four pdb files:

1. ndpk.pdb, which represents the predicted coords of Nucleotide Diphosphate Kinase, as modelled from an alignment with the *Drosophila melanogaster* sequence and structure. The structure predicted is essentially complete.
2. edn.pdb, which is the eosinophil derived neurotoxin modelled on pancreatic bovine ribonuclease A. The main difficulty was the insertion of 9 residues, and a deletion from 10 to 4 residues between two alpha helices in the N-term part of the protein. We had to suppose the the N-term residue of the second helix and the C term one of the first helix were not in a alpha helical conformation. Also, the last three C-term residues have not been modelled.
3. hpr.pdb (histidine containing phospho carrier): hpr1.pdb is modelled upon the *S. faecalis* structure and hpr2.pdb is modelled upon the *E.coli* one. The first and last residues have not been modelled in this last case.

Here follows a brief word about the methodology employed: The method is entirely automatic and uses as an entry the alignment of two sequences (the one to be modelled and the one of the known related structure) plus the coords of the model. Side chains are placed according to a recently published method based on a self-consistent mean field method (Koehl & Delarue, J. Mol. Biol. 239:249-275). A conformational matrix, whose $M(i,j)$ elements represent the probability of finding the i th rotamer at the j th position, is iteratively refined using a simple mean field theory: each residue "feels" the average of all possible rotamers (copies) of all its neighbours, weighted by their respective probabilities; in this way, energies for each possible rotamer are being evaluated for each position; these energies are then transformed into probabilities using Boltzmann formula, and used to update the conformational matrix before going on to the next cycle. The procedure converges in a few cycles. The backbone is essentially fixed and the energy used is only van der Waals. This is just a preliminary version of the program. More elaborated energies are being incorporated right now. The program also automatically takes care of insertions and deletions

by generating multiple copies of possible backbones joining the two ends of the loop, as found in the protein data base (in a method very similar to the one of Jones and Thirup). In principle, the method could use multiple copies of the entire backbone: it has been shown recently to converge towards the right result in test cases (Koehl and Delarue, *Nature, Structural Biology*, submitted). However, for computer time limitations reasons, this has not been used here.

Robert W. Harrison, Devjani Chatterji and Irene T. Weber.

Thomas Jefferson University, Dept. of Pharmacology, Philadelphia, PA 19107, harrison@asterix.jci.tju.edu

Rapid Homology Modeling with Molecular Mechanics and Dynamics.

Six of the comparative modeling targets (HPR, NDK, EDN, P450, CRABP and HALOFER) were predicted with the procedure we are developing for the fast and semi-automatic generation of homology models. The principle new features of this procedure are the use of an all-atom potential with no cutoff on the long range forces, the use of iterative distance geometry, and the use of different algorithms for optimization. The variations which were tested include the "genetic" algorithm and 4-dimensional embedding as well as combining molecular dynamics with conjugate gradients optimization. The molecular mechanics and dynamics program AMMP was used (Harrison 1993).

First, the amino acid sequence of the target is aligned with sequences of related proteins of known structure using GCG. Then the sequences and the alignments are examined in order to reposition insertions and deletions preferentially at the protein surface and between elements of regular secondary structure. Incorrect placement of insertions and deletions is a potential source of major error. Ideally the structure with the highest number of identical residues and the fewest insertions or deletions is used as the starting model. Also, it should be a well-refined structure at relatively high resolution. The starting coordinates are altered to represent the target sequence. The peptide backbone and the atoms which are identical in the side chains are kept and the new atoms are generated by AMMP. The identical atoms are constrained to their original positions and the new atoms are generated with iterative distance geometry. This procedure results in compact structures, but not

necessarily in chemically optimal structures. The structure is then minimized, first allowing only the new atoms to move, and then all of the atoms. The UFF all atom potentials with AMBER charges were used, and no cutoff radius was applied. (AMMP is able to run without cutoffs in times comparable to conventional programs which use cutoffs). We also include crystallographic waters when present and any common ligands because water molecules are often structurally important and functionally conserved in ligand binding sites. The most important parts of the model are the ligand binding sites, because their properties usually determine the biological activity of the protein. The final minimization of all atoms is usually done with a combination of conjugate gradients and short runs of molecular dynamics. The procedure can be repeated to test alternate positions of loops or specific residues. Alternative optimization strategies such as the "genetic" algorithm or 4-Dimensional embedding were also explored. The conserved regions can also be restrained to be near the starting positions during minimization. Manual adjustment of loops or side chains can be used if necessary.

The minimization procedure has been shown to give good agreement between calculated protein-ligand interaction energies and observed binding energies for complexes with similar solubility and entropy changes. With five crystal structures of HIV protease inhibitor complexes, minimization produced RMS differences of .48 to .66 Å on all Ca atoms and .67 to .92 Å for all atoms. These values are close to the amount of difference observed between two crystal forms of the same protein or structures refined in two different labs (.45 to .8 Åstroms for all atoms). This protein structure prediction experiment will allow us to evaluate this strategy and improve the modeling procedure.

W. Bret Church and David Kitson

Biosym Technologies, 1190 Saratoga Ave., Suite 210, San Jose, CA 95129
and Unit 17, Intec 2, Wade Rd., Basingstoke, RG24 ONE, U.K.

Construction of a model of the Histidine-containing Phosphocarrier protein from *Mycoplasma capricolum* (HPr) using homology modeling techniques.

A model for Histidine-containing Phosphocarrier protein from *Mycoplasma capricolum* (HPr) has been calculated using homology modeling techniques. The sequence of HPr was used

Comparative Modeling

with the Profiles-3D program [1] to search known protein structures as a method for identifying structural motifs compatible with this sequence. This analysis identified the mutant *Bacillus subtilis* Histidine-containing Phosphocarrier protein (2HPR), *Streptococcus faecalis* Histidine-containing Phosphocarrier protein (1PTF) and *E. coli* Histidine-containing Phosphocarrier protein (1POH) as possibly suitable models (in that order) and gave sequence alignments. Multiple sequence alignment of the 4 sequences was then performed. No sequence alignment suggested insertions or deletions so 2HPR was used to obtain the starting backbone coordinates. Identical side-chains to 2HPR in the multiple sequence alignment were also taken from this model. When an identity to one of either 1PTF or 1POH existed in the multiple alignment the model side-chain was positioned by minimizing the rms of the identical atoms and all the backbone atoms, producing a global rigid fit of the 2 proteins to construct the model.

An automated search of a rotamer library was used to find the minimum energy conformations of the other side-chains. This represented the starting point for 3 series of energy minimization.

All software used in this study was provided by Biosym Technologies. The Brookhaven Protein Data Bank was the source of all experimentally determined protein structures.

1. Luthy, R., McLachlan, A.D. & Eisenberg, D., Proteins, 10, 229-239 (1991)

Construction of a model of the Histidine-containing Phosphocarrier protein from *Mycoplasma capricolum* (HPr) using homology modeling techniques.

Homology modelling methods have been used to construct a model of the Histidine-containing Phosphocarrier protein (HPr) from *Mycoplasma capricolum*, based on its homology to known structures of this protein from other organisms. The sequence for the HPr protein [1] was compared with the sequences for the Histidine-containing Phosphocarrier proteins from *B. subtilis* [2] (PDB entry 2hpr), *S. faecalis* [3] (PDB entry 1ptf) and *E. coli* [4] (PDB entry 1poh). This comparison was made using both the FASTA program [5] and a multiple sequence alignment technique [Biosym Technologies]. The comparison indicated that all four sequences could be aligned without the need for insertions or deletions. The highest sequence identity (46% in a 74 amino acid region) was with the *B. subtilis* structure and this was chosen as the basis for construction of the *M. capricolum* structure.

The backbone coordinates from the *B. subtilis* structure were used. Residue Met-1 (missing from the *B. subtilis* structure) was built in the conformation found in the *S. faecalis* structure and Gly-89 was built in an extended conformation. Non-mutated side chains were taken directly from the *B. subtilis* structure. The catalytically important residues His-15 and Arg-17 exist in a variety of conformations in the three known structures and these were modelled in their conformation from the *B. subtilis* structure. A sulphate ion that is located between these residues was also used in the modelling process. Side chains for residues that are identical between *M. capricolum* and *S. faecalis* or *E. coli* (but that are different in *B. subtilis*) were built in the conformation from *S. faecalis* or *E. coli*, as appropriate (some were later manually adjusted). Side chains that are similar between *M. capricolum* and one of the reference proteins (for example, Ile vs. Leu) were built in conformations that were similar to those of the reference proteins. The remaining side chain conformations were adjusted using a combination of a manual scan of a side chain rotamer library for individual residues, and an automated procedure that attempts to minimise the energy of a group of side chains by selecting side chain conformers in a systematic fashion [Biosym Technologies]. Residue Phe-4 was adjusted so that the side chain occupied a "hole" left by the Phe-6 → Ala-6 replacement.

The structure was then relaxed using energy minimisation in several stages:

- 1) The three C-terminal residues were minimised while the rest of the protein was held fixed.
- 2) The water molecules from the *B. subtilis* structure were minimised. A 10 Å layer of water molecules was then added around the protein and these water molecules were minimised while the protein was fixed.
- 3) Side chains of about 70% of the residues (mainly surface residues, other than His-15 and Arg-17) were relaxed (tethered by a harmonic constraint).
- 4) All atoms were relaxed (tethered by a harmonic constraint to their starting positions). Amide torsion angles were weakly forced to be near to a planar trans configuration.

The structure was assessed using the Profiles-3D program [6] and geometric checks [Biosym Technologies]. These indicated that the protein was folded in a reasonable conformation. Finally, the protein structure was placed in a box of waters and more thoroughly energy minimised. A short molecular dynamics simulation was then run (all protein atoms, and the sulphate ion, were completely free to move). This was carried out in order to assess the usefulness of minimisation and

dynamics for refining model-built protein structures. The results of this study will be presented at the meeting.

1. Zhu et al., J. Biol. Chem. 268, 26531-26540 (1993)
2. Herzberg et al., Proc. Natl. Acad. Sci. USA, 89, 2499-2503 (1992)
3. Jia et al., Nature, 361, 94-97 (1993)
4. Jia et al., J. Biol. Chem. 268, 22490-22501 (1993)
5. Pearson and Lipman, Proc. Natl. Acad. Sci. USA, 85, 2444-2448 (1988)
6. Luthy et al., Nature, 356, 83-85 (1992)

Ram Samudrala, Jan T Pedersen, Hui-bel Zhou, Rui Luo, and John Moult

Center for Advanced Research in Biotechnology, University of Maryland, Rockville, MD 20850

Krzysztof Fidelis

Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551

Comparative modeling of histidine-containing phosphocarrier (HPr) protein from *Mycobacterium capricolum* (McHPr), the Eosinophil Derived Neurotoxin (edn), and the mouse cellular retinoic acid-binding protein (crabpl)

A method for homologous modeling of protein structures using a mixture of commercial, public-domain and "in-house" algorithms is presented. Models were evaluated using a battery of analytical methods. The following describes the steps in the modeling process: FASTA, OWL, ENTREZ, PROSRCH, SCOP, and PHD databases /programs were used to obtain sequences and structures that were related to the target protein. A multiple sequence alignment was generated with the AMPS package (developed by Geoffrey Barton) using the sequences obtained from the various databases. The alpha-carbon (C-alpha) root mean square deviations (RMSDs) were computed between the structure with highest identity to the target sequence and to the other known structures, using the Align program (developed by G.H. Cohen). A structural alignment was also computed using the G program (developed by Jan Pedersen) to identify discrepancies in the AMPS alignments.

A minimum perturbation (MP) technique (which preserves the backbone phi/psi and equivalent sidechain chi angles) was used to generate initial coordinates for the residues in the target protein, using the information contained in the structure with highest sequence identity. Mainchain stretches that were unfavorable based on the structural analysis described

below were improved by borrowing structural fragments from other homologues.

Insertions and deletions were built using a number of different ab initio loop building and database search methods. Short loops (5-7 residues) were built using the algorithm of Moult & James and Congen (developed by R.Brucoleri). Longer loops were built using a database search method with constraints derived from the framework of the parent structure.

In addition, for each position in the structure, different side-chain rotamers were generated using Insight (Biosym Technologies), Quanta (MSI) packages, and the Self Consistent Domain (SCD) method (developed by Krzysztof Fidelis and John Moult). Insight uses a MP method as described above but with a slightly different equivalence table. Quanta attempts to search conformational space available to the individual sidechain by spinning each of the chi angles, and evaluating the energy for each rotamer and selecting the lowest energy conformation. The SCD method is a locally developed systematic search algorithm that explores - within a residue based domain - all possible sets of sidechain rotamers.

The rotamers were then evaluated on the basis of their environment, electrostatic interactions, and hydrophobic burial. The rules used in the environment evaluation included packing (whether there was too much or too little space left after any change), favorable and unfavorable electrostatic interactions of sidechains and mainchain, van der Waals clashes, and burial or exposure of a residue. Electrostatic interactions were evaluated using the local Eneana program, and the goodness of the burial of a residue was evaluated using a conditional probability formalism (developed by R.Samudrala & J.Moult).

The set of rotamers and loops that were most favorable with respect to the above criteria were used to compose a final model. The final model was refined using careful energy minimization, during which electrostatic interactions were ignored. The minimizations were performed with the Discover (Biosym Technologies) and CHARMM (MSI) potentials.

Work at LLNL was performed under the auspices of the U.S. Department of Energy and supported by Contract W-745-ENG-48 and Laboratory Directed Research and Development Award 93-DI-003.

UCRL-JC-119286 ABS

Andrej Sali¹, Liz Potterton², Feng Yuan¹, Herman van Vlijmen¹, and Martin Karplus¹.

¹ Dept. of Chemistry, Harvard University, 12 Oxford St., Cambridge, MA 02138, USA.

² Molecular Simulations Inc, University of York, Dept. of Chemistry, York YO1 5DD, United Kingdom.

Comparative Modeling of Human Nucleoside Diphosphate Kinase, Mouse Cellular Retinoic Acid Protein 1, and Eosinophil Neurotoxin

Comparative modeling by satisfaction of spatial restraints [1-3] was used to calculate 3D structures of three proteins (human nucleoside diphosphate kinase, NM23H2; mouse cellular retinoic acid protein 1, CRABP1; and eosinophil neurotoxin, EDN). Most of the steps involved in modeling are implemented in program MODELLER [1-2].

First, proteins that have known 3D structure and sequence similar to the sequences being modeled had to be found. This was achieved by MODELLER, which can search a database of sequences representative of the whole Brookhaven Protein Databank. Each candidate sequence was evaluated by a dynamic programming algorithm for finding the optimal alignment between the candidate and target sequences.

A multiple alignment of all the structures with each target sequence was prepared. An initial alignment of the 3D structures was obtained by MODELLER, which implements multiple least-squares superposition. The multiple structural alignment was then aligned as one block with each target sequence, also using MODELLER. This initial alignment was edited manually, primarily to move gaps from helices and strands into the regions that are exposed, variable, and frequently located at the tips of the loops.

Matrices of pairwise sequence identities were calculated from the alignments and used to construct 'evolutionary' trees for the three families [4]. All significantly different structures in the cluster that contained the target sequence were used as templates in the subsequent model building. For NM23H2, nucleoside diphosphate kinase from *Drosophila melanogaster* (1NDL; 77%) and three forms of nucleoside diphosphate kinase from *Dictyostelium discoideum* were used (1NDC, INDP, 1NDK; 60%). For CRABP1, fatty acid binding protein (2HMB; 41%) and mouse adipocyte lipid-binding protein (1LIF; 38%) were used. For EDN, ribonuclease A (7RSA; 33%) was used. The Brookhaven codes and the sequence identities to the target are listed in parentheses.

The alignment and the list of templates were used by MODELLER, without any further intervention, to calculate 3D models for the three sequences containing all mainchain and sidechain heavy atoms. First, MODELLER derived many distance, angle, and dihedral angle restraints on the target sequence from its alignment with template 3D structures. Second, the spatial restraints and energy terms enforcing proper stereochemistry were combined into an objective function. Third, the models were obtained by optimizing the objective function. This optimization was carried out by the use of the target function method employing methods of conjugate gradients and molecular dynamics with simulated annealing. Five models were derived for each sequence by varying the initial structure. The representative model was that which had the lowest value of the objective function. An additional model was calculated for loop segments that fit on the two anchor regions to obtain several starting structures for energy minimization by CHARMM [6]; the submitted loop was the lowest energy structure found.

The models were evaluated in several ways. For self-consistency, the model had to satisfy most restraints used to calculate it, especially the stereochemical restraints. A similar test was performed by the PROCHECK program [5]. It was also useful to compare the models among themselves because those regions that were most variable were also likely to be most in error. The comparison with the actual X-ray structures showed that the model accuracy is within the expected ranges (sequence: DRMS for C-alpha atoms that superpose within 3.5A of each other, their number/total number, the percentage of 'correct' chi1 and chi2 dihedral angle classes). A comparison of NM23H2 chains R and U, which were refined independently, is listed in parentheses. NM23H2 (R.Williams): 0.340A (0.394A), 148/148, 76.2% (77.8%), 64.6% (70.7%). CRABP1 (G. Kleijwegt and A. Jones): 1.19A, 122/136, 64.4%, 70.9%. EDN (S.C.Mossmann and M.N.G.James): 1.39A, 91/134, 62.4%, 61.9%. The source of the largest error (N-terminus of EDN) was an error in the alignment between the EDN and template sequences.

1. A Sali, TL Blundell, *J. Mol. Biol.* 234, 779--815, 1993.
2. A Sali, JP Overington, *Protein Sci.* 3, 1582--1596, 1994.
3. A Sali, R Matsumoto, HP McNeil, M Karplus, RL Stevens, *J. Biol. Chem.* 268, 9023--9034, 1993.
4. J Felsenstein, *Evolution* 39, 783--791, 1985.
5. RA Laskowski, MW McArthur, DS Moss, JM Thornton, *J. Appl. Cryst.* 26, 283--191, 1993.
6. BR Brooks, RE Bruccoleri, BD Olafson, DJ States, S Swaminathan, M Karplus, *J. Comp. Chem.* 4, 187--217, 1983.

Mansoor Saqi

Bioinformatics Group, Dept of Biomolecular Structure, Glaxo Group Research, Greenford, Middx. UK

Molecular Modelling of Eosinophil Derived Nuerotoxin (EDN)

The starting point is an alignment of EDN with 1onc (PANCREATIC RIBONUCLEASE) and with 3rn3 (RIBONUCLEASE A). The sequence alignment required manual adjustment in places but, apart from this, the approach adopted was to generate a model using minimal human intervention. Two models were generated from the same alignment:

- (i) Model 1 was built using the Quanta (Molecular Simulations Inc.) software for modelling the backbone. Loops were generated using the Quanta fragment searching tool and assessed on the criteria of rms fit to the stem and on bad contacts. The sidechains were modelled, outside of Quanta, automatically, with the algorithm of Leach (J. Mol Biol 1994, 235, 345-56).
- (ii) Model 2 was generated from the same alignment, with the automatic modelling program ProMod (Manual Peitsch, Glaxo Institute for Molecular Biology, Geneva). ProMod can be accessed through SwissModel on the Expasy server (URL: <http://expasy.hcuge.ch/>) and requires no human intervention at all. This allows comparison of loop modelling and sidechain packing procedures with Model 1.

Mauno Vihinen

Center for Structural Biochemistry, Karolinska Institute, NOVUM, S-14157 Huddinge, Sweden

BLIND PREDICTION IN SWEDEN

Structures modeled:

hpr: Mycoplasma capricolum HPr (templates 1poh, 1pif, and 2hpr)
 um23h2: human nucleoside phosphate kinase (templates 1ndk, 1nlk, and 2nck)

Sequence analysis was performed with program packages GCG (Devereux et al., 1984) and MULTICOOP (Vihinen, 1990; Vihinen et al., 1992) and it was found to be straightforward in the case of these proteins. The template structures were taken from PDB and superimposed to study their relatedness. Locations of most drastic substitutions were inspected. One of the known structures, the one having highest sequence similarity to the one

to be modeled, was used as template, but all the structures were considered during modeling. Also the possibility of merging the template structures was considered. The modeling was performed with program InsightII (Biosym, San Diego, CA). The side chain rotamers in the substituted residues were taken into account and compared to all the template structures. The models were refined by energy minimization with CHARMM program (Brooks et al., 1993), while constraining the conserved residues. Validation was performed by using three techniques: PROCHECK (Laskowski et al., 1993), 3-d verify (Luthy et al., 1992), and Poldiag (Baumann et al., 1989) all of which indicated the models to have reasonable and structurally favourable conformations.

Baumann, G., Frimmel, C. and Sander, C. (1989) *Prot. Engin.*, 2, 329-334.
 Brooks, B. R., Brucoleri, R. E., Olafsen, B. D., States, D. J., Swaminathan, S., Karplus, M. (1983) *J. Comp. Chem.* 4, 187-217.
 Devereux, J., Haeblerli, P., and Smithies, O. (1984) *Nucleic Acids Res.* 12, 387-395.
 Laskowski, R. A., MacArthur, M. W., Moss, D. S. and Thornton, J. M. (1993) *J. Appl. Cryst.* 26, 283-291.
 Luthy, R., Bowie, J. U. and Eisenberg, D. (1992) *Nature* 356, 83-85.
 Vihinen, M. (1990) *Meth. Enzymol.* 183, 447-456.
 Vihinen, M., Euranto, A., Luostarinen, P. and Nevalainen, O., *Comput. Appl. Biosci.* (1992) 8, 35-38.

Gert Vriend

EMBL Meyerhofstrasse 1, 69117, Heidelberg, Germany

The program WHAT IF provides a multitude of tools that can aid with modeling, drug design, structure description and comparison, mutant prediction, inverted threading, etc. One of the many tools is automatic model building by homology from a given alignment with the sequence of a known structure. The three models submitted for the model building by homology competition were generated using this fully automatic method (only a few manual interventions were made, and those are listed in the submitted file headers).

Given a perfect alignment, this method normally leads to very good models, although many problems still have to be solved. We routinely use a set of 25 pairs of structures that can be modeled back and forth. This dataset has been selected to fulfill the following criteria:

very few insertions and deletions.
 equally spread over the 35-95% identity range

not too many co-factors
equally spread over all known folds
high quality

During the demonstrations several typical problem cases can be discussed. The quality of the model will never be better than the quality of the structure it was based on. Therefore the program WHAT IF provides extensive structure verification tools. These can be used to check the starting structures, but also to verify the final models. In several of the submitted models WHAT IF detected errors in the final model. I think (we will see it at the meeting...) that these are the result of poor sequence alignment. This is a different problem from the one I concentrated on, but it is nice to see that programs can detect these problems automatically. At the meeting we will see if these error detection techniques actually work or not. In any event, they can be discussed during the demonstrations.

Ps. WHAT IF is not a commercial product. It is distributed under restricted share-ware conditions.

Steve Bryant

National Institutes of Health, NCBI, NLM, Bethesda, MD 20894

Description of the Method

Threading runs were performed using the contact potential and alignment model described in (1). Segments in the trial sequence are aligned with explicitly-defined core elements in the folding motif, that is, with individual strands or helices. No gaps are allowed within core elements, but the lengths of intervening loops are free to vary within specified limits. Scoring is based on the sum of contact potentials for a given alignment, less the sum expected for random sequences with the same composition. No gap penalties are employed, and the threading score is based only on this sum of contact potentials. Note that this alignment model requires that all residue sites within a core motif be aligned with a residue from the threaded sequence. We thus search for a complete core structure or that of a pre-defined structural domain somewhere within the threaded sequence, but we do not attempt to identify matches with arbitrary structural fragments. We do, however, search for smaller structural domains within long sequences, by attempting to identify matches with all core motifs which have fewer residue sites than a sequence has residues. In this search the lengths of sequence segments C- or N-terminal to a core motif are neither constrained nor considered in scoring.

For the Asilomar contest we use two new techniques not described in (1). We first of all use a fast heuristic algorithm to search for favorable alignments, a Monte Carlo procedure based on the Gibbs Sampling algorithm. Subsequence blocks are initially aligned at random with core elements. We then sample allowed alignments of each element in turn, in the field defined by the others, and choose new alignments based on the Boltzmann probabilities of the alternative model structures generated in this fashion. This procedure is iterated, and in control experiments may be shown to converge to ensembles containing the correct native alignment. In the Gibbs Sampling algorithm we also allow the precise endpoints of core elements within the known structure to vary within, specified limits. Their values are again chosen stochastically, based on the Boltzmann probability of the model structure generated by extension or contraction of a core element. A second new technique is search of an explicitly defined database of core motifs. Large helices and beta strands in structures from the Protein Data Bank are identified by alpha-carbon distance templates, and trimmed from their ends until only a certain fraction of contacts with

other core elements remain. The minimal core motif defined in this way is intended to mimic a substructure one can expect to be conserved in any protein with the same "fold". In addition to minimal cores for complete structures in the Protein Data Bank, we define core motifs for large, chain-continuous globular domains, identified on the basis of chain breaks which produce a high ratio of intra- to inter-domain contacts. In threading we must align some segment from a contest sequence with a minimal core motif, but core elements may also be extended, via the Gibbs algorithm, to include more residues sites from the known structure. The precise boundaries of the core elements in a model structure are thus given as part of the threading results.

(1) Bryant, S.H., Lawrence, C.E., An Empirical Energy Function for Threading Protein Sequence through Folding Motif, *Proteins*, 16:92-112 (1993).

Neil Clarke

Johns Hopkins School of Medicine, Dept. of Biophysics and Biophysical Chemistry, Baltimore, MD 20205

A threading-type program written about five years ago was used to predict the structures of BphC and of synaptotagmin (Bowie JU, Clarke ND, Pabo CO, and Sauer RT. Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins* 7:257 (1990)). The program and the database were not changed in any significant respect since then.

The program uses the Needleman-Wunsch algorithm to align sequences to structures. Structures are represented as strings of solvent accessibility. Sequences are represented as strings of hydrophobicity. The elements for both the accessibility and hydrophobicity strings are assigned one of three values (e.g., high, medium and low accessibility; hydrophobic, in-between, hydrophilic). The cutoffs for assigning each residue to the appropriate class were determined empirically by maximizing the ability of globin sequences to identify globin structures. The scoring matrix and default gap penalties were determined in the same way.

The sequence strings used in these searches are actually a kind of consensus sequence. In the simplest case, at every residue position in a set of aligned homologous sequences we ask what the most hydrophilic amino acid is at that position and use that amino acid in assigning the hydrophobicity class for

R.R. Copley, C.D. Livingstone, R.B. Russell and G.J. Barton

Laboratory of Molecular Biophysics, The Rex Richards Building, South Parks Road, Oxford, OX1 3QU United Kingdom. Tel: 44 865 275368 E-mail: cd@bioch.ox.ac.uk rrc@bioch.ox.ac.uk
r.russell@europa.lif.linet.uk, gjb@bioch.ox.ac.uk, WWW URL: <http://geoff.bioch.ox.ac.uk>

Prediction of the folds of Synaptotagmin and the Beta subunit of Urease

The aim of our work is to identify the likely fold of a protein given a family of sequences. This is achieved in two steps: Secondary structure prediction (including prediction of buried residues) and Secondary Structure Mapping (identification of the fold). A secondary structure prediction is obtained from an accurate multiple sequence alignment for the family by identifying positions of insertions/deletions, conserved residues and hydrophobic conservation patterns, and combining this analysis with the results of a variety of conventional single sequence secondary structure prediction algorithms. This strategy has previously been used to generate accurate secondary structure predictions for the annexins, SH2 domains and protein-tyrosine phosphatases prior to the experimental determination of the structures [Barton et al., 1991; Russell et al., 1992; Livingstone & Barton, 1994]. Specifically, helices and strands were predicted by the methods of Lim, Chou and Fasman and Robson as implemented in the "Leeds" prediction suite [Lim, 1974a; Lim, 1974b; Chou & Fasman, 1978; Garnier et al., 1978; Elloupos, 1989]. Predictions of turn were made using the algorithms of Rose and Wilmot and Thornton [Rose, 1978; Wilmot & Thornton, 1988].

Regions of the multiple alignment where insertions and deletions are seen, or which are varied in composition across the aligned family of sequences were assigned to coil (non-core secondary structure). The intervening regions were assigned either to helix or strand by referring both to the results of the classical prediction algorithms and to patterns of residue conservation identified with the aid of the AMAS program [Livingstone & Barton, 1993]. The combined procedure yields a final three-state prediction. Secondary structure predictions by the method of Zvelebil et al. [1987] were used as a "casting vote" where ambiguity between strand and helix assignment could not otherwise be resolved.

For each prediction, a non-redundant set of sequences was extracted from the current version of the PIR [Sidman et al., 1988] and Entrez [Ostell, 1992] databases. Multiple sequence

that residue position. Since evolutionary replacement of a surface hydrophobic residue by a polar residue is more likely than the same substitution at a buried position, the most hydrophilic residue at a particular position in a sequence alignment should correlate better with solvent accessibility than would a particular residue in a single sequence. In practice, we found that throwing out one or a few of the most hydrophilic residues and then using the next most hydrophilic worked best. This is because (1) the sequence alignments can be ambiguous and (2) even in well aligned sequences subtle structural differences can allow a polar residue in one structure and not in another. Also we found that handling Arg and Lys in a special way helped things because of the amphipathic nature of these sidechains (see paper for details).

BphC: Six homologues of BphC were aligned, partly from published alignments and partly by eye. The program was run with various number of residue rejections (see above for how the consensus sequence is derived) and with a couple of different sets of gap penalties but the results consistently indicated a good match with catalase. For example, two rejections with the default gap penalties gave a Z score of greater than 5 for catalase. All other structures were between -3 and 2. With BphC I actually tried very hard to get the program to give a good alignment with glycolate oxidase (GO) rather than to catalase because circumstantial evidence makes me suspect that the protein actually looks more like GO. However, although GO consistently had a high raw score, the Z score was never as high as that of catalase.

Synaptotagmin: The PKC domain is found twice in synaptotagmins. I aligned the sequences of both domains from synaptotagmins of four species, for a total of eight sequences. Here again the search was performed using 1,2,3 or 4 rejected residues at each position. Unlike BphC, no one structure jumped out as being clearly better using any single set of rejections/gap penalties. However, when a consensus rating was generated by looking at the average ranking of each structure in the several different searches, it was found that two structures seemed significantly better than all others. The two were tobacco bushy stunt virus (chain C) and hemagglutinin. I exercised some subjective judgement in predicting the similarity to hemagglutinin because the biological function of the two proteins are related.

alignments were made using the AMPS package [Barton & Sternberg, 1987; Barton, 1990]. Poorly conserved regions within the resulting alignments were adjusted by eye in the locality of gaps.

Given the secondary structure prediction, prediction of buried residues and any additional knowledge, for example the location of catalytic residues, we searched a non-redundant set of protein domains to find folds consistent with the secondary structure prediction and other constraining data. This was achieved by the application of a newly developed technique known as Secondary Structure Mapping (SSM) [Russell et al., 1994]. SSM finds topologies consistent with secondary structure assignments and that are consistent with any experimental restraints (i.e. disulphide bonds, active site residues, etc.) available for a protein/protein family. First, all possible alignments of secondary structure elements are generated between a query (i.e. prediction) and a database (i.e. PDB) structure. These alignments (or maps) are then filtered to remove those structures which are 1) uncompact; 2) having ends of secondary structures too far apart to be bridged by predicted loop lengths; 3) having poor beta sheet hydrogen bonding; 4) lacking essential secondary structures (e.g. those with active site residues); 5) do not satisfy topological constraints (e.g. sequential beta strands are most often antiparallel); and 6) those which are unable to satisfy any experimentally derived distance restraints. The program is fast (able to search a representative set of PDB domains in about five minutes) and is able to go from a large number of initial maps (i.e. prior to filtering) to a few structurally sensible maps.

Proteins Predicted: Synaptotagmin C2 domain, Urease Beta subunit, The mystery protein (I)

- Barton, G. J. (1990). *Methods Enzymol.* 183, 403-428.
 Barton, G. J., Newman, R. H., Freemont, P. F. & Crumpton, M. J. (1991). *Eur. J. Biochem.* 198, 749-760.
 Barton, G. J. & Sternberg, M. J. E. (1987). *J. Mol. Biol.* 198, 327-337.
 Chou, P. Y. & Fasman, G. D. (1978). *Adv. Enzymol. Relat. Areas Mol. Biol.* Eliopoulos, E. (1989). The Leeds protein secondary structure suite.
 Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). *J. Mol. Biol.* Livingstone, C. D. & Barton, G. J. (1993). *Comput. Appl. Biosci.* 9, 745-756.
 Livingstone, C. D. & Barton, G. J. (1994). *Int. J. Pept. Protein Res.* Lim, V. (1974a). *J. Mol. Biol.* 88, 857-872.
 Lim, V. (1974b). *J. Mol. Biol.* 88, 873-894.
 Ostell, J. (1992). *Entrez sequences graphical user interface.*
 Rose, G. D. (1978). *Nature*, 272, 596-591.
 Russell, R. B., Bred, J. & Barton, G. J. (1992). *FEBS Lett.* 304, 15-20.
 Russell, R., Copley, R. & Barton, G. (1994). A method for fold prediction. The method is still under development.
 Sidman, K., George, D., Barker, W. & Hunt, L. (1988). *Nucleic Acids Res.*

Wilnot, A.C. M. & Thornton, J. M. (1988). *J. Mol. Biol.* 203, 221-232.
 Zvelebil, M. J. J. M., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). *J. Mol. Biol.* 195, 957-961.

Andrew F.W. Coulson

Biocomputing Research Unit, ICMB, Darwin Building, University of Edinburgh, Kings Buildings, Mayfield Road, Edinburgh EH9 3JR, United Kingdom, a.coulson@ed.ac.uk

Fold recognition by sequence similarity

Two sequence databases were created. 'U' consisted of the Swissprot sequence entries corresponding to Sander's representative set of structures. 'EU' consisted of the Swissprot entries for all the sequences referenced in HSSP. In other words, it contains all sequence entries any part(s) of which are so strongly similar to proteins of known structure that (these parts) can be unequivocally aligned with a known fold. Two other databases were also used - NRL3D and Swissprot.

Searches were made with each unknown sequence against one or more of these databases, using the Smith and Waterman 'Best Local Similarity' algorithm, implemented on the AMT Distributed Array Processor, or the MasPar MP-1. Scoring tables derived by the Dayhoff prescription for various evolutionary distances, expressed in 'PAM's' (1 PAM corresponds to 1 accepted point mutation per 100 residues of sequence). Conservative gap penalties were used, in a range in which experience shows that the appearance and length of gaps is not strongly dependent on the gap penalty value.

The output lists were scanned for significant similarities to proteins of known (or reliably inferred) structure; using the statistical criterion described in *Methods in Enzymology* 1990, 183, 474-486. Confirmation of potential positive hits was sought by repeated searches, using scoring tables with a varied range of PAM parameters, and by identification of a multiplicity of less significant hits on sequences of the same structure.

Finally, the sequence surroundings of a potential positive hit were examined by sequence comparison using a 'flat' comparison table (generated by a high PAM value), and the structure surroundings by molecular graphics and secondary structure prediction. Reprotoed predictions are cases in which a significant local similarity could be plausibly extended over a substantial part of the unknown sequence.

Assessment of the performance will require access to all the 'threading' example structures (not all predictions were

Sequence to Fold Assignment

submitted, either because they were negative, or because of pressure of time), but it is clear that none of the databases used was ideal. The most important improvement to the method would probably be by the construction of a better database, in which sequences were flagged by a structural classification.

Tim Hubbard and Jong Park

Centre for Protein Engineering, MRC Centre, Cambridge, UK, th@mrc-lmb.cam.ac.uk

Our entries to the structure prediction competition are for those sequences where there is no known homology with any sequence of known structure, with the objective of either recognising similar folds in the database of known structures (PDB) or predicting ab-initio a rough 3-D topology.

Both the fold recognition and ab-initio prediction algorithms used rely on the information contained in multiple sequence alignments, so predictions were not made in cases where the sequence family was very small or the alignments ambiguous. Initial multiple sequence alignments were generally obtained by mailing the target sequence to the PHD secondary structure prediction server [1] and were improved by adding additional related sequences obtained from running blast against a number of databases and using a number of other automatic and manual alignment methods.

For fold recognition hidden markov models (HMM) [2] were constructed from the multiple sequence alignment. Models were then used to search a subset of pdb chain sequences (pdb90), none of which has greater than 90% homology with any other. The sequence or alignment was also mailed to the PHD server to obtain a secondary structure prediction [1]. For each alignment a value was calculated measuring the degree of similarity between predicted secondary structure segments and those observed in the known structure (from DSSP [3]) using an algorithm similar to [4]. By considering the hmm score, the secondary structure overlap score and the ranking of similar folds in the list (using the fold classification of scop [5], which is incorporated into pdb90) a prediction of fold type was made. The predictions for xyla, kau, synapto, bphc and l14 were based mainly on the results from this approach.

In cases where a high beta sheet content was observed, an ab initio beta-strand pairing prediction was made [6]. The results of this prediction were combined with the PHD secondary structure prediction to identify strands most likely to pair.

Because the number of possible pairings is proportional to the square of the number of strands, whereas the number of observed pairs is linearly related, prediction generally becomes less reliable as the number of strands increases. The predictions of the small proteins, prosub and staufen3 were mainly made using this approach. In the case of large proteins, the method was used to guide the selection of the closest fold in PDB, based on the similarities in predicted contact maps observed between target and folds proposed to be similar.

For the cases of rtp and chmut, no fold could be recognised and the sequences predicted to be mainly helical, however certain sequence patterns suggesting leucine zippers were identified and speculative topology predictions submitted based on this.

Structures Predicted	Main Method Used
=====	=====
xyla	hmm
kau	hmm
prosub	hmm
synapto	hmm
staufen3	hmm
bphc	hmm + beta-strand topology
rtp	sequence patterns
chmut	sequence patterns
l14	hmm + beta-strand topology
=====	=====
Structures not predicted	Reason not predicted
bhted	too few homologous sequences
pcna	too short notice (2 days)
smnucecs	too few homologous sequences
podk	not enough time
pbdg	not enough time
mystery	suspected to be some sort of joke!

We thank Burkhard Rost, Reinhard Schneider and Chris Sander for access to the PHD secondary structure prediction server [1], the DSSP program [3], and the HSSP database, to Sean Eddy for use of HMM program suite [2], Erik Sonnhammer for the use of SWIR5, to Andrej Sali for use of Modeller and to TH's collaborators Alexey Murzin, Brenner and Cyrus Chothia for the development of SCOP [5]. TH is grateful to the MRC and ZENFCA for financial support.

- [1] B. Rost, C. Sander: Prediction of protein structure at better than 70% accuracy. *J. Mol. Biol.*, 1993, Vol. 232, pp. 584-599.
- [2] S. Eddy, "HMM": Hidden Markov Modeling of Proteins and Nucleic Acids", software freely available via anonymous ftp to celm.mrc-lmb.cam.ac.uk, in pub/sre, documentation from <http://celm.mrc-lmb.cam.ac.uk/>. Contact sre@mrc-lmb.cam.ac.uk for information.
- [3] W. Kabsch and C. Sander (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22, 2577-637.
- [4] B. Rost, C. Sander, R. Schneider (1994). Redefining the goals of protein secondary structure. *JMB*, 235, 13-26.

Sequence to Fold Assignment

Adam Godzik

Scripps Research Institute, La Jolla, California

"Topology fingerprint approach to inverse folding problem"

Available protein structures are divided into sequence families by clustering them based on the level of sequence similarity. The best quality structure from each cluster is then included into a database of representative protein structures. Each protein in this database is represented as a topology fingerprint: a contact map with additional information about the local secondary structure and side chain solvent exposure.

A topological fingerprint can be used to calculate a score, later called energy, of an arbitrary sequence "forced" to adopt this particular structure. Energy parameters are developed from contact statistics derived from an independent database of highly refined protein structures and include burial, two body and three body contributions. As shown on several examples, energy calculated in this way can be used as an indication of a protein structure quality. Series of additional simplifying assumptions are necessary to allow the introduction of gaps into the sequence being "forced" into a given structure. In this variant of that method, even weak homologies can be recognized and some cases structural similarities may be predicted.

D.T. Jones, R. Miller & J.M. Thornton

Biomolecular Structure and Modelling Unit, Dept. of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, United Kingdom

The listed predictions were carried out using an updated version of the method of Jones et al. (1992). A dynamic programming algorithm ('double' dynamic programming) was used to align the given sequence with the 'real' coordinates of each structure in a library of folds, taking into account residue pairwise interactions. Matching pairwise interactions is related to the requirements of structure comparison methods. The interaction environment of a residue i is defined as being the sum of all pairwise interactions involving i and all other residues $j \neq i$. This is a similar definition to that of a residue's structural environment, as described by Taylor and Orengo (1989). In the simplest case, the structural environment of a residue i is defined as the set of all

[5] A. Murzin, S.E. Brenner, T.J.P. Hubbard, C. Chothia (1994). The SCOP (Structural Classification of Proteins) database: available world-wide over the internet through the world wide web (WWW) at <http://scop.mrc-lmb.cam.ac.uk/scop/>.
[6] T.J.P. Hubbard (1994). Use of beta-strand interaction Pseudo-Potentials in Protein Structure Prediction and Modelling. In R.H. Lathrop (eds.), Proceedings of the Biotechnology Computing Track, Protein Structure Prediction Minitrack of the 27th HICSS. IEEE Computer Society Press, pp. 338-354. (machintosh Postscript file available by anonymous ftp to ind2.mrc-lmb.cam.ac.uk, as pub/th/beta/hicss-27.pa.z. Contact th@mrc-cpe.cam.ac.uk for information.

Danny Rice and David Eisenberg

University of California at Los Angeles, Institute of Molecular Biology, Los Angeles, California

The following 4 predictions were made using the Smith and Waterman alignment algorithm with a home-made residue-residue scoring table. The residue preference scores were derived from a database of pairs of aligned protein structures that had the same fold, and had sequence identity below 35%. The scoring table was used to align contest sequences against representatives of known structural folds. The score was determined for each alignment by randomizing the sequences 100 times, and calculating the number of standard deviations above the random mean the true sequence scored. This score for these four predictions varied between about 6 and 7.

- 1) Target Sequence to predict: (CME) Carboxymuconate lactonizing enzyme.
Detected protein: (6TAA) Alpha amylase.
- 2) Target Sequence to predict: (IPNS) Isopenicillin N synthase.
Detected protein: (2PMGA) phosphoglucosmutase
- 3) Target Sequence to predict: (PPDK) pyruvate phosphate dikinase.
Detected protein: (2PNI) phosphatidylinositol 3-kinase (p85-Alpha subunit SH3 Domain)
- 4) Target Sequence to predict: (KITP) Replication Terminator Protein.
Detected protein: (2HPDA) Cytochrome P450.

Inter-Ca distances between residue *i* and all other residues *j* in Taylor and Orengo developed a novel dynamic programming algorithm for the comparison of residue structural environments, and it is a derivation of this method that we therefore use for the effective comparison of residue interaction environments.

At the heart of the evaluation function used here is a set of pairwise potentials of mean force, determined by a statistical analysis of highly resolved protein X-ray crystal structures and the application of the inverse Boltzmann equation. In addition to the pairwise potentials, a solvation potential is also used. This potential is determined in a similar fashion to the pairwise potentials, except that the variable in this case is relative solvent accessibility rather than interatomic distance.

A modified set of potentials was used for all predictions apart from the first 3. The main difference between the old and new pairwise terms is a correction for the size of the proteins used to generate the potentials and the protein being predicted (D. Jones, paper in preparation). This correction allows more of the pairwise interaction information to be extracted than by simply truncating the potentials at 10 Å. The new solvation potentials are based on just 5 unequal relative accessibility divisions, rather than 10 or 20 equal divisions.

Target	No. of library folds	Potentials
xylanase	244	unscaled/10 solv. divisions
bhted	253	unscaled/10 solv. divisions
smarucacs	253	unscaled/10 solv. divisions
bphc	253	scaled/5 solv. divisions
ce-1	253	scaled/5 solv. divisions
urease	253	scaled/5 solv. divisions
114	266	scaled/5 solv. divisions
pbdg	266	scaled/5 solv. divisions
pxdk	266	scaled/5 solv. divisions
xtp	266	scaled/5 solv. divisions
stauten	266	scaled/5 solv. divisions
synapto	266	scaled/5 solv. divisions

References

1. Jones, D.T. Taylor, W.R. & Thornton, J.M. (1992) Nature 358, 86-89.
2. Taylor, W.R. & Orengo, C.A. (1989) J. Mol. Biol. 208, 1-22.

Yo Matsuo and Ken Nishikawa

Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan, matsuo@peri.co.jp, nishikawa@peri.co.jp

A Method for Evaluating Protein Sequence-Structure Compatibility

Four evaluation functions, side-chain packing (Fsp), solvation (Fsolv), hydrogen-bonding (Fhb), and local structure (Floc) functions, were used to evaluate the compatibility of an amino acid sequence with a structure [1-4]. They have the following general form:

$$Fx = -\log (\hat{f}(x;s)/\hat{f}(s)), \quad x = (sp, solv, hb, loc),$$

where *a* denotes the type of amino acid residue (for Fsolv and Floc) or residue pair (for Fsp and Fhb); *s*, the state of a (spatial) relationship between residues for Fsp, solvent-accessibility for Fsolv, hydrogen-bonded or not for Fhb, and local structure for Floc; $\hat{f}(x;s)$, the frequency of *a* in the state *s*; and $\hat{f}(s)$, the frequency of any residue or residue pair in the state *s*.

The side-chain packing function (Fsp) indicates the propensity of a residue pair (*a,b*) to be in contact in a particular spatial relationship. The spatial relationship between two residues was defined by the distance between their C α atoms and the angle between the residues. The angle between residues *i* and *j* was defined as the sum of the angles Cb(*i*)-Ca(*i*)-Cb(*j*) and Cb(*j*)-Ca(*j*)-Cb(*i*).

For the other three functions (Fsolv, Fhb, and Floc), the state of a residue or residue pair was defined as follows. Solvent accessibility of a residue was defined by the number of main-chain and C β atoms within a shell between 8 Å and 12 Å from the C α atom of the residue. Hydrogen bonds were defined by the DSSP algorithm [5]. Local structures (backbone conformations) of residues were classified into five classes: α -helical, right-handed helical, b-strand, extended, and left-handed helical.

A sequence was threaded onto a structure using the 3D-profile method [6,7]. The 3D-profile of a structure was constructed as in [7] using the above functions. A sequence was aligned with the 3D-profile using the Needleman-Wunsch algorithm [8]. A sequence was mounted on a structure according to the alignment thus derived. For a sequence mounted on a structure, scores S_x ($x = \{sp, solv, hb, loc\}$) were given by summing up the values of F_x over all residues or residue pairs. The S_x scores were then added up to give S_{tot} , which measured the compatibility of the sequence with the structure. A more negative score indicates better compatibility.

A sequence was compared with a library of known structures. In the present work, 325 structures were taken from PDB. They have less than 30% sequence identity with one another. For the individual structures, compatibility scores S_{tot} were calculated. The scores were expressed in units of standard deviations from the mean (see [3] for details).

Manfred Sippl, Hannes Floeckner, Michael Braxenthaler

Center of Applied Molecular Engineering, University of Salzburg, Jakob
Haringer Str.1, A-5020 Salzburg

PREDICTION METHOD ABSTRACT

The results submitted to the prediction evaluation were obtained by two different methods: (1) Fold recognition and (2) Assembly of backbones from small overlapping fragments.

The fold recognition technique consists of the following parts (1) a knowledge based energy function (mean force potentials), (2) a technique to align sequences with structures (allowing gaps) and (3) quality assessment of the models obtained in the alignment (z-score).

The energy function consists only of C-beta-C-beta interactions and an approximation for solvent interactions. The energy function and z-score calculation is identical to the functions used in the PROTEIN Structure Analysis program PROSA-II available from gundl.came.sbg.ac.at, which can be used to detect incorrect folds or faulty parts in a structure (Sippl,M.J., Proteins, 17, 355, 1993).

The alignment technique used is still largely experimental and unpublished (some information on the strategies used has been described by Sippl,M.J., J.Comput.Aided Mol.Design, 7, 473, 1993 and Sippl,M.J. et al. in The Protein Folding Problem and Tertiary Structure Prediction, K.Merz and S.LeGrand (eds.), Birkhauser, Boston, 1994, pp 353-407).

Each target sequence was combined with a subset of structures obtained from the Brookhaven data base. The structure yielding the highest score was defined as the best model and the corresponding alignment was submitted as a prediction (in terms of backbone coordinates and the secondary structure assignments, helix or strand, obtained from the respective X-ray structure).

From our previous studies we found that in roughly one out of four cases the current implementation is able to recognise a related fold in the data base in the absence of significant sequence homology. This, of course, is only possible if a related fold exists in the data base. If no related fold exists the method often yields substructures similar to the target structure, but necessarily incorrect overall folds (in fact the extent of similarity is difficult to assess).

The second method employed is based on the combination of small fragments (Sippl et al. Protein Science, 1, 625-640, 1992). The method yields full backbone coordinates.

- [1] Nishikawa, K. and Matsuo, Y. (1993) Protein Eng., 6, 811-820.
- [2] Matsuo, Y. and Nishikawa, K. (1994) FEBS Lett., 345, 23-26.
- [3] Matsuo, Y. and Nishikawa, K. (1994) Protein Sci., in press.
- [4] Anano, T., Yoshida, M., Matsuo, Y. and Nishikawa, K. (1994), FEBS Lett., 351, 1-5.
- [5] Kabach, W. and Sander, C. (1983) Biopolymers, 22, 2577-2637.
- [6] Bowie, J.U., Luthy, R. and Eisenberg, D. (1991), Science, 253, 164-170.
- [7] Wilmanns, M. and Eisenberg, D. (1993), Proc.Natl.Acad.Sci.USA, 90, 1379-1383.
- [8] Needleman, S.B. and Wunsch, C.D. (1970) J.Mol.Biol., 48, 443-453.

List of Structures Predicted:

a	b	c
L14	L14 (prokaryotic ribosomal protein)	1GCR
bhted	beta-hydroxydecanoyl thiol ester dehydrase	1OFV
bphc	biphenyl-2,3-diol 1,2-dioxygenase	2LBP
ce-1	Chymotrypsin/Elastase Inhibitor-1	1TBPA
clmut	N-terminal of P-protein (Chorismate Mutase)	2REB
kau	Urease from Klebsiella aerogenes subunit A	8ACN
	subunit B	2MCM
	subunit G	2ER7E
mystery	A mystery sequence	2GBP
pbdg	6-Phospho-beta-D-galactosidase	3LADA
ppdk	pyruvate phosphate dikinase (PPDK)	2MNR
	domain 1	11PD
	domain 2	2AAB
	domain 3	2AAB
	domain 4	1PII
prosub	propeptide from subtilisin BPN'	8DPR
rtg	Replication Terminator Protein	1ABK
stmanu	extracellular endonuclease	2ER7E
staufn3	Domain 3 of Staufn	5TIMA
synapto	First C2 domain of synaptotagmin	1ATND

a. Abbreviation of the name of the protein
b. Name of the protein whose structure was predicted
c. PDB code of the structure which showed the best compatibility score.

1GCR, Gamma-II crystallin; 1OFV, flavodoxin;
2LBP, leucine-binding protein;
1TBPA, C-terminal 179 amino acids of TATA-binding protein;
2REB, RecA protein; 8ACN, aconitase; 2MCM, macromycin;
2ER7E, endothiapsin; 2GBP, galactose/glucose binding protein;
3LADA, dihydriolipoamide dehydrogenase; 2MNR, mandelate racemase;
11PD, 3-isopropylmalate dehydrogenase; 2AAB, rctin;
1PII, N-(5-phosphoribosyl)anthranilate isomerase; indol-3-glycerol-phosphate synthase;
8DPR, dihydrofolate reductase; 1ABK, endonuclease III;
5TIMA, triosephosphate isomerase; 1ATND, deoxyribonuclease I.

but uses only local energy terms. It can be used to predict the local structures but not the tertiary fold. In this prediction experiment we used the results in the following way: The backbones were assembled and the secondary structure assignments were calculated from the model. These results were then compared to the secondary structure assignments derived from the best model obtained from the sequence structure alignment technique. The consistency between the two results was used as an indication for the quality (in terms of local backbone geometry) of the predictions.

We submitted the best scoring models for 15 target sequences, although it was clear that the method has a chance to correctly predict a structure only if a closely related fold is contained in our data base of known structures.

Matthias Wilmanns

EMBL, Postfach 102209, D-69012 Heidelberg, Germany.

David Eisenberg

UCLA, 405 Hilgard, Los Angeles, CA 90024-1570, USA.

Inverted Folding by the Residue Pair Preference Profile Method.

The residue pair profile (R3P) method is an inverted folding method that combines the idea of environmental profiles and pair profile preferences. The method uses statistical preferences for residue pairs. Each pair is created from a profiled residue and a residue in the local environment of the profiled residue. All residue pairs are characterized by their

- (a) dihedral angles
- (b) secondary structure
- (c) number of neighboring residues (OOI numbers)

as function of residue types. Each residue pair preference is expressed for all 20 amino acids of the profiled residue and is weighted by the compatibility of the environment residue with its own local environment. The R3P method produces an initial profile/sequence alignment which is then refined by converting the initial profile into a profile of a target sequence threaded into the structure of the initial profile.

We have tested this method by evaluating alignments of sequences with known 3D structures using structural superposition as reference. R3P/sequence alignments are $\geq 50\%$ correct on the average for sequences whose 3D structure pairs

superimpose with an rms deviation of $\leq 1.97\text{\AA}$. The average improvement in correctness during this iterative refinement is 14%. The R3P/sequence alignments are compared to sequence/sequence and 3D profile/sequence alignments. When all three methods are combined on the average $\geq 50\%$ of the alignments are correct for pairs of 3D structures with $\leq 2.12\text{\AA}$.

Single sequences have been screened for compatibility against a R3P database (3044 entries). Examples for built 3D models from sequences of unknown 3D structure, detected with compatible but non-homologous R3Ps, will be presented at the meeting. Possible 3D models for sequences of the structure prediction contest will also be discussed.

The ETH Prediction Group

Dietlind L. Gerloff, Gareth Chelvanayagam, and Steven A. Benner

Department of Chemistry, ETH, CH-8092 Zurich, Switzerland

A Predicted Consensus Structure for the Protein Kinase C2 Homology (C2H) Domain, the Repeating Unit of Synaptotagmin

A Consensus Prediction of the Secondary Structure for the 6-Phospho-beta-D-Galactosidase Superfamily

The prediction method applies automated heuristics to assign surface, interior, active site (tertiary structural information), and parsing residues by analysis of patterns of conservation and variation among homologous protein sequences in light of evolutionary models that interpret amino acid substitutions as the consequence of neutral variation subjected to functional constraints. Secondary structural elements are assigned from patterns in the tertiary structural information.

Benner, S. A. Patterns of Divergence in Homologous Proteins as Indicators of Tertiary and Quaternary Structure. *Adv. Enzym. Regul.* 28, 219-236 (1989)
 Benner, S. A., and Gerloff, D. L. Patterns of Divergence in Homologous Proteins as Indicators of Secondary and Tertiary Structure: The Catalytic Domain of Protein Kinases. *Adv. Enzyme Regul.* 31, 121-181 (1991)
 Benner, S. A. Predicting de novo the Folded Structure of Proteins. *Curr. Opin. Struct. Biol.* 2, 402-412 (1992)
 Benner, S. A., Cohen, M. A., Gerloff, D. L., Correct Structure Prediction? *Nature* 359, 781 (1992)
 Gerloff, D. L., Jenny, T. F., Knecht, L. J., Gonnert, G. H., Benner, S. A. The Nitrogenase MoFe Protein: A Secondary Structure Prediction. *FEBS Lett.* 318, 118-124 (1993)
 Benner, S. A., Cohen, M. A., Gerloff, D. L., A Predicted Secondary Structure for the Src Homology Domain 3. *J. Mol. Biol.* 229, 295-305 (1993)
 Gerloff, D. L., Benner, S. A. Predicting the Conformation of Proteins: Man versus Machine. *FEBS Lett.* 325, 29-33 (1993)
 Gerloff, D. L., Jenny, T. F., Knecht, L. J., Benner, S. A. A Secondary Structure Prediction of the Hemorrhagic Metalloprotease Family. *Biochem. Biophys. Res. Commun.* 194, 560-565 (1993)
 Benner, S. A., I. Badcoe, Cohen, M. A., Gerloff, D. L., De Novo Prediction of Folded Structures: Assigning Interior and Surface Residues from Patterns of Variation and Conservation in Homologous Protein Sequences. *J. Mol. Biol.* 235, 926-958 (1994)
 Benner, S. A. Predicting the Conformation of Proteins from Sequence Data. in *Protein Engineering: A Guide to Design and Production*. C. S. Craik, J. Cleland, editors, in press (1993)
 Benner, S. A., Jenny, T. F., Cohen, M. A., Predicting the Conformation of Proteins from Sequences. *Progress and Future Progress. Adv. Enzyme Regul.* 34, 269-353 (1994)
 Jenny, T. F., Benner, S. A. Evaluating Predictions of Secondary Structure in Proteins. *Biochem. Biophys. Res. Commun.* 200, 149-155 (1994)
 Benner, S. A., Gerloff, D. L., T. F. Jenny, G. Chelvanayagam, Knecht, L. J., Gonnert, G. H. Transparent, bona fide predictions of protein secondary structure. *Nature, Structural Biology* submitted (1994)

Jenny, T. F., Benner, S. A. A Prediction of the Secondary Structure of the Pleckstrin Homology Domain. *Proteins: Struct. Funct. Genet.* 20, 1-4 (1994)
 Benner, S. A., Gerloff, D. L., Jenny, T. F. Predicting Protein Crystal Structures. *Science*, 265, 1641-1643 (1994)

S. G. Galaktionov and G. R. Marshall

Center for Molecular Design, Washington University, St. Louis, MO 63130

CALCULATION OF TERTIARY STRUCTURE OF SUBTILISIN BPN' PROPEPTIDE AND DOMAIN 3 OF THE PRODUCT OF DROSOPHILA GENE STAUFEN BASED ON PREDICTION OF THEIR INTRAGLOBULAR CONTACT MATRICES.

We used a modernized version of the approach previously described [1]. The approach use four main procedures: prediction of the secondary structure; prediction of the coordination number vector (i.e., the number of contacts for each residue); prediction of the contact matrix; reconstruction of the spatial structure (C-alpha's coordinates) on the basis of contact matrix. These procedures can be iterated in cyclic schemes to refine/edit concertedly the coordination number vector, the location of the elements of secondary structure, contact matrix and tertiary structure. The structures of both subtilisin BPN' propeptide (prosub) and domain 3 of the product of drosophila gene staufer (staufer) were predicted using essentially the same cyclic protocol.

Three known algorithms based on Bayes statistics, information theory and neural networks for SECONDARY STRUCTURE PREDICTION were used as implemented in SYBYL 5.5. The starting location of secondary structure was assumed at regions of complete consensus for all three predictions for prosub (helices, 18-24, 45-58, 67-74; beta-strand, 9-12) and 80% consensus for staufer (helices, 14-25, 56-70; beta-strand, 39-57).

The algorithm previously described [2] was used for PREDICTION OF COORDINATION NUMBER VECTORS starting with amino acid sequences and the data on the location of the secondary structure.

The core part of the procedure, the PREDICTION OF THE CONTACT MATRIX, includes routines for calculation of the starting matrix, bringing it into conformity with the set of stiff conditions (like symmetry) and satisfying the specific optimization criteria regarding the coordination number vector, the contact matrix, its powers and its eigensystems. The refinement routines are organized as an iterative procedure

removing the "weakest" contacts (using a given "goodness" matrix) and filling out subsequent vacancies in rows/columns at the most preferred positions. There are many possible ways to combine single operations and optimization matrices in this procedure. The protocol we applied for the first calculation of the contact matrix for both proteins was that providing the reconstruction of the spatial structure of 3FLX, a de novo designed protein, with the accuracy of 2.8 Å. 3FLX is of the same size as prosuB and staufer (79, 77 and 79 residues, respectively) and, like both predicted proteins, significantly helical.

For RECONSTRUCTION OF SPATIAL STRUCTURE on the basis of the contact matrix, we used our procedure [1]. Its parameters were tuned so that the mean intraglobular distance and its variance were close to the values characteristic for proteins of corresponding size (about 15.5 Å and 5.5 Å, respectively).

The first structure obtained was used for re-editing of the contact matrix (e.g., by removing the stressed contacts, etc.) and refinement of prediction of both secondary structure and coordination number vector, then followed a repeated structure reconstruction, and so on until near-convergence. In the resulting structures, most regions of regular secondary are significantly extended as compared with initially predicted locations.

1. Galaktionov, S. G. and G. R. Marshall. Proc. 27th Hawaiian International Conference on Systems Sciences, Biotechnology Computing, IEEE Computer Society Press, V:326-335 (1994).
2. Rodionov, M.A., Galaktionov, S.G. Mol. Biol., 26:777-783 (1992).

Dave Covell

FCRDC, Ft. Detrick, Frederick, MD 21702

Methods

Simulated annealing methods are applied to a simple cubic lattice alpha-carbon model of a 63 amino acid monomeric globular protein. Monomer units occupy only one lattice site and are connected along each lattice edge to their sequence neighbors located in adjacent lattice sites. All conformational changes are required to lie on this lattice. Each simulation begins with a randomly selected, expanded chain with a radius of gyration greater than six times that of comparably-sized folded proteins. At each simulation step the transition to an alternative

conformation is selected from a Boltzmann weighted distribution of chain energies for all possible single step transitions accessible to the current chain. Five types of self-avoiding, local moves are available for chain rearrangements, and all allowed moves are found at each simulation step. The energy of each chain is determined from considerations of number and type of nearest-neighbor non-bonded interactions, chain chirality and radius of gyration. Multiple simulations of greater than 100,000 steps are completed for each randomly chosen starting configuration. The lowest energy forms found in the collapsed states are compared to determine a consensus folding motif. This final configuration has been submitted to the competition.

Jean Garnier

Unité d'Ingénierie des Protéines, Biotechnologies INRA, 78352 Jouy-en-Josas, Cedex, France

1- Program SIMPA(J. Levin and J. Garnier, Biochim. Biophys. Acta, 1988, 28, 1177-92) when an homologous structure is known. SIMPA (SIMilarity Peptide Analysis) program is based on sequence similarity between a stretch of amino acids (17 amino acid long) of the test sequence and the sequences in a data base of proteins of known structure. Cross validated Q3 is 86% when the data base contains an homologous protein of known structure, if not the accuracy drops to 63-65%.

2- Predictions program SIMPA (J. Levin & J. Garnier, Biochim. Biophys. Acta, 1988, 28, 1177-92) with multiple alignments of homologous sequences (J. Levin et al. Protein Eng., 1993, 6, 849-54). When no homologous structure is known but homologous sequences are known, the cross validated Q3 can be improved from 63-65% to 69-70% by averaging the predictions at each position of a multiple alignment of the known homologous sequences (program CONSENSUS).

3- Program COMBINE (V. Blou et al. Prot. Eng. 1988, 2, 185-191) with multiple alignments of homologous sequences. COMBINE is an expert program of three secondary structure prediction algorithms: GOR III, SIMPA and Bit Pattern. This latter prediction program is based on hydrophobicity patterns and has a cross validated Q3 of 59%. COMBINE cross-validated Q3 is about 65%. COMBINE can be associated to CONSENSUS to yield an accuracy of 69-70% (V. Di Francesco, P.J. Munson and Garnier J., 28th Hawaii International Conference on System Science, 1995).

**Yukio Kobayashi¹, Nobuhiko Saito², Makoto Ohta³,
Hiroyuki Sasabe⁴, and Shigeki Mitaku³**

¹ Department of Information Systems Science, Faculty of Engineering,
Soka University, Hachioji, Tokyo 192, Japan.

² Department of Applied Physics, Waseda University, Shinjuku-ku,
Tokyo 169, Japan.

³ Faculty of Technology, Tokyo University of Agriculture and
Technology, Koganei, Tokyo 184, Japan.

⁴ Frontier Research Program, The Institute of Physical and Chemical
Research(RIKEN), Wako, Saitama 351-01, Japan.

Prediction of the Structures of GPCT and HPR by the Island
Model

We predict the structures of GPCT and HPR with ab initio
method based on the mechanism of protein folding, which is
referred to as the "island model" (Saito et al., 1988).

This method is formulated on the physicochemical basis, and
thus it is applicable to any type of proteins with low homology.
We have selected these proteins consisting of a small number of
residues in order that the structures of these proteins can be
predicted with our method in time
for the deadline of prediction.

We assume that folding starts first with the formation of
the secondary structures and then proceeds to assemble them
into the tertiary structure. The procedures of predicting a protein
structure are summarized as follows:

1. Determination of secondary structures. Perform the three-
state prediction with statistical mechanical method already
formulated before (Saito, 1987) to provide simultaneously the
probabilities of each residue in a-helix, b-strand or coil. This
formulation involves the following assumptions:
(1) At least one residue in coil exists between neighboring
secondary structures.
(2) Interactions are considered only among the residues within
four residues along the chain in the same secondary structure.
The above probabilities are calculated with statistical weights
for amino acid pairs in a-helix or in b-strand. We have estimated
the weights so as to minimize an objective function by referring
to 80 proteins for optimization which lack homology among
them as far as possible.

2. Model of protein chain.

- (1) Build a polypeptide chain with QUANTA (software
applications for modeling the structures and behavior of
molecular systems).

- (2) Replace side chains of amino acid residues by a sphere of van
der Waals radius.
- (3) Locate each sphere at the average distance of the center of
gravity on the direction of the b-carbon from the a-carbon.
- (4) Fix bond lengths and bond angles at the standard values.
- (5) Construct the initial conformation with the predicted
secondary structures and the extended conformations ($\phi=180^\circ$,
 $\psi=180^\circ$) for others.

3. Packing of secondary structures into tertiary structure.

- (1) Calculate energies by considering long-range hydrophobic
interactions and Lennard-Jones (12,6)-potentials between
relevant atoms from short distance pairs to longer ones step by
step.

(2) Pack the secondary structures through the local structure
formation by searching for the conformation of lower energy.

4. Refinement of the conformation with QUANTA. Generate the
atoms in the side chains and introduce the electrostatic
potentials ignored in the above step.

We represent the predicted conformations of GPCT and
HPR with the distance maps, where the conformations are
indicated on the basis of the distances between a-carbons.
These maps will be shown in the poster presentation.

Saito, N. et al. (1988) Proteins, 3, 199-207.

Saito, N. (1987) Cell Biophys., 11, 321-329.

Jan Pedersen and John Moul

CARB, University of Maryland Biotechnology Institute, Rockville, MD
20850, USA.

Determination of the Structure of Small Protein Fragments using
Torsion Space Monte Carlo and Genetic Algorithm Methods.

We have adopted an approach to determining structure
from sequence that seeks to understand protein folding
pathways (Moult and Unger, 1991), with the aim of then imitating
that pathway in folding simulations. So far, we are able to fold up
selected fragments of proteins to conformations close to that
found in the final full structures.

A torsion space representation of conformation is used,
together with an all atom model. The force field uses scaled
empirical point atomic charge electrostatics together with a
surface area term representing electrostatic and hydrophobic
solvation contributions. Both the electrostatics and surface

area terms have been parameterized by a potential of mean force analysis of protein structures (Avbelj, 1992; Avbelj and Moutl (a),(b)).

Initial conformations are generated starting from a random coil, and changing mainchain and side chain angles at random, drawing replacement angles from a library of observed values, and accepting or rejecting moves using the Metropolis criterion. Samples from the Monte Carlo run (Avbelj and Moutl (b)) are used to provide a starting population for a Genetic Algorithm. New generations are created by performing cross overs on the population, annealing the side chain conformations of the hybrids so formed, and then using the Metropolis criterion to decide whether to accept them.

In tests on small (12-16 residues) protein fragments selected for their independent folding properties, both Monte Carlo and Genetic Algorithm methods are able to generate conformations close to the experimental ones in most cases, although there are occasional convergence and potential problems.

As part of the protein structure prediction experiment, we attempted to use the Genetic Algorithm method to determine the conformation of three peptides: residues 7 to 22 of the subtilisin propeptide (prosub), residues 1 to 15 of Eosinophil Derived Neurotoxin (EDN), and the membrane binding domain of the C2 domain of human coagulation factor VIII (membind), a 22 residue peptide. Results of these experiments will be presented.

Avbelj, F. Biochemistry 29 2403-2408 1992
Avbelj, F. and Moutl, J. (a) Biochemistry (in press)
Avbelj, F. and Moutl, J. (b) (submitted)
Moutl, J. and Unger, R. Biochemistry 30 3816-3823 1991.

Peter J. Munson and Valentina DiFrancesco

Analytical Biostatistics Section, Laboratory of Structural Biology,
Division of Computer Research and Technology, National Institutes of
Health, Bldg 12A, Room 204, Bethesda, MD 20892-5626,
(munson@helix.nih.gov)

Maximum Likelihood Periodic Quadratic-Logistic Profile Predictions

We have submitted secondary structure predictions for the proteins: ipns, pbpg, ppdk, prosub, 114, staufen3, and mystery. Our methodology is called a maximum likelihood quadratic logistic (QL) discrimination model based on profiles

[1,2]. Briefly, we have calibrated a logistic model for a three state prediction using the maximum likelihood principle assuming that secondary structural state obeys an independent trinomial probability model. The logistic model includes linear or "main-effect" terms for every amino acid residue within a 17 residue window of the state to be predicted, together with certain quadratic or "pair-wise" effects. Namely, we assume a 3.6 residue period for the helix component, and a 2.0 residue period for strand, and multiply the residue pair-wise term by $\cos(2\pi k/3.6)$ or $\cos(2\pi k/2.0)$. The $2 \times 20 \times 20 = 800$ residue pair preference parameters are estimated along with the main-effect terms, using a penalized maximum-likelihood technique. Crossvalidated prediction rates for this method are seen to be 62.5% using on single sequences.

The profile method begins with a set of aligned homologous sequences, and rather than representing the sequence elements by a 20 vector of zeros and ones (dummy variables), uses the proportions of each residue seen at an aligned position, giving a 20 vector of proportions. For quadratic terms, we replace the 400-vector of dummy variables representing the residue pairs observed within a window with the corresponding 400 vector of proportions. Alignments are done by first choosing homologues from Swiss-Prot or PIR with greater than 20% homology on stretches longer than 80 residues, and using either pairwise or multiple alignment (CLUSTAL, PILE-UP) programs to determine alignments. Alignments were reviewed manually to remove obvious spurious homologues or spurious portions of the alignment. Areas with gaps in the final alignment were arbitrarily assigned a high coil probability. The expected percent correct figure for Q-L using profiles is 67% to 69%, in two separate crossvalidated tests.

[1] Munson, P. J., V. Di Francesco and R. Porrelli. Protein Secondary Structure Prediction using Periodic-Quadratic-Logistic Models: Theoretical and Practical Issues. 27th Annual Hawaii International Conference on System Science, 5: 375-384, 1994.

and updated in:

[2] Di Francesco, V., P.J. Munson, J. Garner. Use of Multiple Alignments in Protein Secondary Structure Prediction. 28th Hawaii International Conference on System Sciences, (accepted), 1995.

D.J. Osguthorpe

Molecular Graphics Unit, University of Bath, Claverton Down, Bath, BA2 7AY

Ab Initio Protein Folding

The protein folding methods I am using have started from a simplified model of protein structure with potentials developed to reproduce the physical behaviour of atoms rather than from statistical analysis of the database. This is a fully flexible model for folding as molecular dynamics is being used as the conformational space search technique. It is also an evolving model in that changes will be made when it appears the model cannot reproduce protein structures. Currently, this model does not involve backbone hydrogen bonding groups, it uses the classical virtual C_{α} bond approximation, with an atom positioned on each C_{α} atom. The side chain representation is also simplified but has evolved from a single atom centre per side chain to the current 1 to 3 atom model. An early decision was to retain the direction of the C_{α} -C β bond, as this simplified the generation of internal potentials around this bond.

The potentials are currently based on 4 major components, internal potentials (bond lengths, valence angles, torsion angles and out of planes), non-bond potentials, surface area solvation potentials and helix/sheet potentials. The basic internal potential around the C_{α} atom is based on a full-flexible phi,psi map of a 2 residue peptide, plus analysing the geometry of known proteins. The potential forms vary from simple harmonic terms to quartic polynomials and sums of gaussian functions. In analysing known protein structures only C_{α} 's which were NOT assigned the main secondary structures by the DSSP algorithm were included in geometry histograms, to remove the bias of the essentially fixed geometry of the main secondary structures. The out of plane potential was used to constrain the geometry around the C_{α} to the appropriate non-planar geometry for L amino acids. The nonbond potential is a Lennard-Jones 10-6 potential, with parameters based on two factors, the energetics of the approach of full atom side chains to each other for like pairs, using a full atom non-bond potential plus partial charges, and least square fitting of the r^* radius to minimise the deviation of the coordinates of 4 known proteins from their near X-ray positions. The near X-ray positions were generated by template forcing runs from 0.1 A RMS to 0.3 A RMS to reduce initial clashes due to the conversion to the model representation.

The solvent potential is based on the first shell relative surface accessibility (RSA) but it includes a linear function of RSA and RSA raised to selected powers. In all cases the solvent potential favours the exposure of the atom grouping, the difference between charged and "hydrophobic" amino acids being the power function which determines the rate at which the energy goes to 0 as RSA goes to 0. The energy value for the potential is taken from partial atomic energy values for the non-bond energy such that the solvent energy will roughly counteract this energy.

The helix/sheet potentials are based on energy functions which stabilise the distances and orientations between C_{α} atoms in the helix and sheet geometry, using gaussian functions of distance differences which vary from 0 to 1 when the distances exactly match. There is no sequence dependence in this potential, i.e. it is the same for all C_{α} which can form 2 hydrogen bonds, Pro being treated differently as it cannot form certain hydrogen bonds.

So far, the potential seems to be able to reproduce the geometry and secondary structures of known proteins i.e. proteins do not move rapidly away from the X-ray structure at low "temperatures". However, folding simulations lead to non-native collapsed structures which have similar amounts of secondary structure energy but too buried "hydrophobic" residues. The folding protocol currently follows a simple start from extended structure at 600 K and cool gradually, with extended periods at a fixed temperature after 500 K is reached at every 50 K, i.e. 500, 450, 400 etc. With an infinite non-bond cutoff proteins collapse at 450 K or so. Using a cutoff of 9.5 A leads to much less compact structures with a significant increase in the amount of secondary structure formed. The collapse in this case is at much lower temperatures, 350-300 K, and even the "collapsed" structures are larger than those with an infinite cutoff.

These problems are demonstrated in 3 folds of the staufer protein, one with an infinite cutoff from extended chain, two with a 9.5 A cutoff, one from extended and one from an all helical conformation. The extended runs started from a temperature of 600K, the helical run involved a cool to 0.1 K followed by a heat up to 400 K and then cooling.

Sander Group

EMBL Heidelberg, Europe

Prediction of 3D structure for the Aslomar contest. Dec. 4-8, 1994

Secondary structure was predicted for all proteins using the neural network method that uses sequence profiles as input (Rost and Sander, 1993, 1994).

The method employed for 3D prediction for the contest makes use of broad evolutionary, biochemical and structural knowledge and uses hydrophobicity as the main numerical criterion by which to optimize the alignment of the model sequence to a known 3-D structure. We typically built only one model using a template singled out using intuition.

Step 1: collect a sequence family alignment using MaxHom (Sander and Schneider, 1990). If possible, families are extended by pattern searches (Rohde and Bork, 1993, CABIOS 9, 183-189). In one prediction case (cytidyltransferase), a motif similarity to a protein already in the PDB was identified. At least two other remote homologues were missed: urease with adenosine deaminase and beta-galactosidase with beta-amylase, both detected by Dali structure comparison (Holm and Sander, JMB, 1993).

Step 2: secondary structure prediction after multiple sequence alignment using the PHD program (Rost and Sander JMB 1993, 232, 584-599, Proteins 1994, 19, 55-72) and by looking at conserved hydrophobicity patterns. If the reliability index of the PHD prediction was low, we occasionally allowed ourselves to change the secondary structure assignment.

Step 3: figure out a compact 3-D fold for the secondary structure elements by drawing diagrams on paper. Very hydrophobic elements should be interior and invariant residues which are likely to cluster around the active site should be close in space. If a protein has beta strands, they must form sheets. There are a few favoured folding motifs (Holm and Sander 1993, JMB 233, 123-138) for alpha/beta proteins. Helices may associate as bundles or polyhedra. In the prediction game we proceeded to 3-D model building with two presumed TIM barrel proteins.

In the second case (beta-galactosidase) we listed a number of fuzzy reasons that favoured a TIM barrel topology based on sequence conservation in the protein family. The reasons were

- (i) high frequency of TIM barrel topologies among glycosyl hydrolases,

- (ii) large size of the protein,

(iii) typical helix-Gxx-strand pattern which is observed also between TIM barrels that are unrelated in evolution,

(iv) conserved residues clustered at the C-terminal side of the strands (all TIM barrels have the active site on this side),

(v) short loops at the bottom and long loops at the top (side of active site) is a characteristic seen in the structure comparison of TIM barrels.

Step 4: generate a 3-D model by substituting the side chains but keeping the backbone of a known structural template (program MaxSprout, Holm and Sander, JMB 1991, 218, 183-194). The sequence-structure alignment was made iteratively. The initial alignment tries to conserve hydrophobic patches between template and the modelled sequence family (sequence between template and the modelled sequence family (sequence is better conserved in the hydrophobic core than on the protein surface). The model is then evaluated using atomic solvation preference (Holm and Sander, JMB 1992, 225, 93-105). Solvation preference profiles suggest places where the alignment should be improved. Native proteins and deliberately misfolded models have a very large gap in solvation preference. Credible models have solvation preferences closer to native proteins than to misfolded ones. The alignment is iteratively improved to optimize solvation preference. Loops /insertions are excluded from the 3-D model and backbone remains fixed. The sidechain optimization in rotamer space also tries to minimize clashes in the core. Few clashes is also an indication of a credible model.

In the first case (xylanase) we obtained a good 3-D model on a TIM barrel template as evaluated by solvation preference criteria but also saw some implausible features in the model and did not believe it.

Having screwed up the first case, we put more weight on the reasonable solvation preference in the second prediction (beta-galactosidase) although this model also had some errors by visual inspection, and were right to do so.

Our sequence-structure fitness program for threading (FosFos = fitness of sequence for structure; Ouzounis et al. JMB 1993) was not used for the contest as we do not consider it sufficiently reliable in its present form.

Contributions from Lilisa Holm, Burkhard Rost, Peer Bork and Chris Sander. Abstract by Lilisa Holm, edited by Chris Sander.