A Knowledge-Based Method for Protein Structure Refinement and Prediction

Shankar Subramaniam[†#*], David K. Tcheng[†] and James M. Fenton[†]
[†]Beckman Institute, National Center for Supercomputing Applications,
[*#]Center for Biophysics and Computational Biology &
Department of Molecular and Integrative Physiology,
University of Illinois at Urbana-Champaign, Urbana, IL 61801

Abstract

The native conformation of a protein, in a given environment, is determined entirely by the various interatomic interactions dictated by the amino acid sequence (1-3). We describe here a knowledge-based approach for protein structure assessment and prediction. Using a well-defined set of high-resolution protein structures, we have derived statistical potentials, in the form of atom-pairwise distance probability density functions. These provide a description of pairwise interatomic interactions of native proteins. When applied to highly randomized and noisy structures of proteins distinct from the basis set, native-like structures were obtained to very high precision (≤2Å). The examples tested include proteins of all sizes (from 38 up to 461 amino acids long) and diverse topological structures (alpha, beta and alpha-beta classes). The potentials appear to be sensitive enough to recognize subtle distortions from a native packing structure and in optimization of structures drive them consistently to a higher probability. Therefore they provide a powerful tool for refinement of X-ray and NMR derived structures at arbitrary degrees of initial precision.

Efforts to fold a protein from a random structure corresponding to its sequence have met with little success. The objective of a number of these efforts has been to minimize an energy or free energy function that describes interatomic interactions (4-6) to obtain the folded protein structure. These energy functions have been obtained from theoretical or phenomenological considerations. The direct energy function methods include use of mechanics force-fields (7) and semi-empirical force fields (8,9). The traditional molecular mechanics force fields use energy functions for bonds, angles, torsions and for pairwise nonbonded interactions. These have been employed for both local structure predictions and in conjunction with crystallographic data or distance-constraints obtained from magnetic resonance methods (10,11). Knowledge-derived potential functions have also been employed to fold protein sequences into structures to a limited degree of success. These include residue-based profiles (12,13), lattice models (14-16), threading methods (17-20) and homology models (21,22).

Knowledge-Based Potentials

The problem of describing a folded protein in terms of the optimal interatomic interactions can be inverted to obtain energy functions that are optimal for folded protein structures. A majority of these methods use high resolution protein structures to derive pairwise residue-level contact information or in a few cases atom level interactions (23,24). The knowledge of pairwise atomic contacts in known protein structures reflects the relative probability of finding atom pairs at specified distances and hence can provide a measure of the free energy of interaction. The theoretical foundations for this stem from the Boltzmann principle, which asserts that the probability of a given atomic configuration for a protein structure is related to the free energy. Minimizing the free energy of a protein is equivalent to maximizing concomitantly all the pairwise atomic distance probabilities. We note here that this is strictly true if and only if the probabilities are truly independent.

We have used the above principle to develop statistically-derived potentials for refining and predicting protein structures. We derive the statistical potentials as probability density functions (PDFs) that describe the distribution of distances between different groups of atoms. Each heavy atom in an amino acid is described as a group and the twenty amino acids yield 167 groups of atoms. The distance data for constructing the distributions is obtained from a database of 380 unique protein structures. The latter set is constructed from a larger database of protein structures by requiring each member of the unique set to have a resolution of less than 2.5 Å and any two members to have less than 50 percent sequence homology as defined by standard BLAST protocols (25). Distance examples are generated from this database for every pair of atomic contacts and these are used to construct the PDFs. The conditional pairwise distance PDFs take the form of Probability $(X| R_i,A_k,R_j,A_l,S_n)$, where, X is the distance, R_i and R_j represent residue indices, A_k and A_l atom indices, and S_n represents the sequential distance between the residues R_i and R_j. The total probability of pairwise distance contacts in a protein is given by combining the conditional probabilities,

$$\text{P(protein)} = \prod_{k,l,i,j,S_n} P(X|R_i,A_K,R_j,A_l,S_n)$$

where the indices run over all atoms, residues and the specified sequential distances. The sequential distance is used so as to preserve the sequentially contiguous interactions that give rise to secondary structure in proteins. The case where n=0 represents the intra-residue PDFs which are a measure of the configurational and conformational geometry of the amino acid considered. We observe that for PDFs of atom pairs separated by more than 3 residues there are no specific secondary structure interactions and we consider these as tertiary PDFs. A unique PDF is formed for each unique combination of R_i, R_j, A_k, A_l and S_n. For instance, P(X|Val,CG1,Leu,CD1,3) represents the PDF for Val CG1-Leu CD1 atom pairs for which the parent residues Val and Leu are sequentially in 1-3 positions. A key problem in deriving these potentials is the non-uniform distribution of pairwise distances in the distance space. To overcome this limitation statistically rigorous methods of kernel density estimation and maximum likelihood evaluation are used to construct the PDFs (26).

Methods

The January 1994 relase from the Brookhaven Protein Data Bank (3611 protein chain sequences) was used in building the non-homologous set of proteins. Each entire sequence was compared against all of the sequences in the database. Sets of homologous protein chains were created with each set containing proteins which had more than 50% identity. Amongst the homologous set, the highest resolution protein was chosen as a representative [28]. Thus a list of unique chains was selected.

The total number of atom types corresponding to the heavy atoms in the twenty amino acids is 167. The pairwise atomic distance PDFs are generated for intra residue, residues related by positions, n-n+1, n-n+2 and n-n+3 and each of the other pairs, within 10 Å form the tertiary PDFs. The n-n+1, n-n+2 and n-n+3 and tertiary PDFS are computed seperately for N to C and C to N terminal directions. The PDFs are assumed independent of each other. The only additional PDF used is one corresponding to the S-S bond in disulfides. The total number of PDFs thus amount to 112,226 types and the number of atomic distance pairs in the 380 proteins considered are 80,670,588. The compressed PDFs occupy approximately 115 MBytes of storage. The computational time required for constructing all the PDFs is 268 hours on a single R8000 SGI processor. The annealing of medium sized protein from a random structure takes about 40 hours on a single R8000 processor.

Kernel Density Estimation (KDE) coupled with Bias Optimization (BO) used to construct the distance PDFs. KDE algorithm employs a normal distribution for the "kernel" function. The algorithm first distributes the distance examples along the distance axis, and then slides the kernal function accross the distance axis while computing a weighted sum. The result of this convolution is then normalized so the area under the curve sums to 1.0. The final result is a PDF that estimates the probability of any pairwise distance. The height of the curve is proportional to the probability. In other words, a distance D1 is roughly twice as likely as distance D2 because the D1's probability P1 is twice as large as D2's probability P2.

The width of the kernel, sigma, is a critical parameter for obtaining optimal PDFs. If sigma is too small, the result will be a "jagged" distribution that "overfits" the data. Conversely, if the sigma is too large, important local changes in probability will be smoothed over which will "underfit" the data. The solution is to optimize the choice of sigma by selecting a range of different sigmas and to select the sigma that performs the best.

The precise definition of the performance (i.e., the objective function for optimization) is important. We need a measure that reflects the predicted performance of the system when operating on new (i.e., unseen) problems. In our method we measure performance by repeatedly selecting a random subset (e.g., 90%) of all examples for training (i.e., inducing the PDFs) and using the rest of the examples for testing (i.e., predicting distances). The process is repeated a number of times and the accuracy of predictions is averaged over all trials. The accuracy of a prediction is measured in terms of maximum likelihood principle (MLP). The MLP states that the best probability distribution function is the one that makes the joint probability of the examples most likely.

$$\text{performance} = \prod_{i=1}^{i=t} P(D_i)$$

For numerical considerations, this equation is expressed in terms of logs and is of the form:

$$\log(\text{performance}) = \sum_{i=1}^{i=t} \log P(D_i)$$

We form the PDFs using only the training examples and test the performance only on the test examples. In summary the choice of a sigma is dependent on the average likelihood of test examples when evaluated against PDFs formed from the training examples across multiple partitions of the training and testing set.

Further, squared distance division is carried out to obtain radial density normalization. In Fig. 1, we present exemplar PDFs for 4 categories. The height of the resulting normalized PDF curve is a measure of the relative probability of finding two atoms belonging to two residues at the defined distance. We note that all but the tertiary probabilities when normalized by squared distance go to zero before 10 Å and the tertiary probabilities reach a plateau value by 10 Å. We compute a mean probability from these truncated PDF curves. In evaluations of protein structures, we subtract the expected mean probability value corresponding to each PDF from the actual probability to assess how significantly better or worse that specific pair interaction is as compared to an average pair interaction of that type. Further, we can obtain either atomic or residue profiles of a protein, based on averaging over interactions of each atom or over each residue respectively and these profiles provide a relative measure of deviation from ideality of the interactions of the defined atom or residue in the context of the protein structure.

The total normalized probability summed over all pairwise atomic probabilities was computed for all the 380 proteins data set chosen for PDF construction. It was found that there was a correlation between the total logarithmic probability and the resolution of the structures. The Spearman rank correlation for normalized probabilities yielded a Z-value of -5.6 and a P-value less than 0.0001.

In the annealing procedure, atoms are incrementally moved in the direction which maximizes probability of all the atom's interactions, weighting each interaction equally. In each step, atoms are moved one at a time and the order in which each atom is moved is randomized each step. An atom is moved by defining a sphere of radius 0.2 Å around it, and randomly selecting 100 candidate points within the sphere using a uniform sampling distribution. The candidate points are evaluated one at a time and the difference between the candidate atom probability and current atom probability is calculated. The standard probabilistic simulated annealing acceptance criterion is used to determine when to move the atom.

The only additional constraint used in addition to the distance PDFs was a torsional term that was biased towards the correct chirality for each amino acid. For all the optimization experiments, 200 steps of the annealing procedure was employed. The control parameters for annealing, the tertiary interactions weight, the chirality weight, and the temperature used for simulated annealing changed each step based on linear schedules. The weight of tertiary interactions was varied from 0.0 to 1.0, the weight of the chirality term varied from 0.08 to 0.01, and the temperature in normalized probability units varied from 1/32 to 1/1000.

Table 1. Comparison of Total Probabilities of Low and High Resolution Protein Pairs

Protein	PDB Name (Resolution in Å)	Total distance log(prob)	PDB Name (Resolution in Å)	^Total distance log(prob)
Pancreatic Trypsin Inhibitor	3pti(1.50)	-0.0043	9pti(1.22)	-0.0037
Alpha Bungarotoxin	2ebx(1.40)	-0.0230	3ebx(1.40)	-0.0178
Cytochrome B5	2b5c(2.00)	0.0215	3b5c(1.50)	0.0338
Plastocyanin	1pcy(1.60)	-0.0113	1plc(1.33)	-0.0082
Parvalbumin	1cpv(1.85)	-0.0052	4cpv(1.50)	0.0361
Cytochrome B562	156b(2.50)	-0.0165	256b(1.40)	0.0877
Pseudoazurin	1aza(2.00)	-0.0324	2aza(1.80)	-0.0230
Proteinase A	1sga(2.80)	-0.0472	2sga(1.50)	-0.0141
Dihydrofolate Reductase	1dfr(2.50)	-0.1858	3dfr(1.70)	-0.1221
Lysozyme	1lzm(2.40)	-0.0142	3lzm(1.70)	0.0421
Alpha-Lytic Protease	1alp(2.80)	-0.0510	2alp(1.70)	-0.0192
Actinidin	1act(2.80)	-0.2524	2act(1.70)	-0.0045
Acid Proteinase	1apr(2.50)	-0.1147	2apr(1.80)	-0.0064
Thermolysin	1tln(2.30)	-0.8905	3tln(1.60)	0.0082
Glutathione Reductase	2grs(2.00)	-0.1496	3grs(1.54)	0.0016

^ The total distance log (probability) is both mean value and r-squared normalized.

PDF Profiles of Protein Structures

The PDF curves represent local structure and packing interactions in proteins. An ideal protein would have every pair of atoms in the regions of high or highest probability and thus possess optimal interactions. In general higher resolution protein structures have higher pairwise atomic probabilities. In Table 1, we compare the total logarithmic probability scores averaged over all the pairwise interactions in the protein for 15 protein pairs, whose structures have been obtained at two resolutions. These proteins were not included in the 380 unique protein set used to construct the PDFs. In each case the higher resolution protein structure has a higher probability score. In one case, where the structures had the same resolution, the one with a lower crystallographic R-factor had

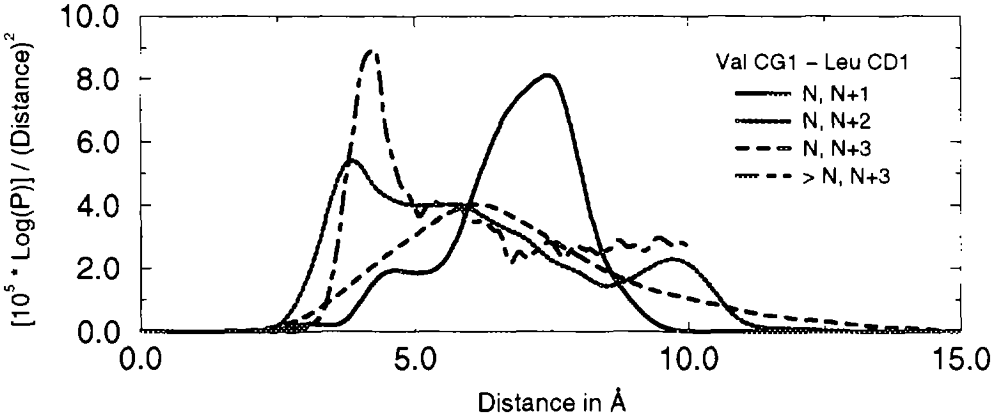


Figure 1. Representative PDFs for Val CG1-Leu CD1 atom pairs. N, N+1 represents adjacent Val-Leu residues. N,N+2 those seperated by one residue, N,N+3 those seperated by two and > N, N+3 stands for all other Val CG1-Leu CD1atom pairs. The PDFs are first normalized to unity and then by radial density.

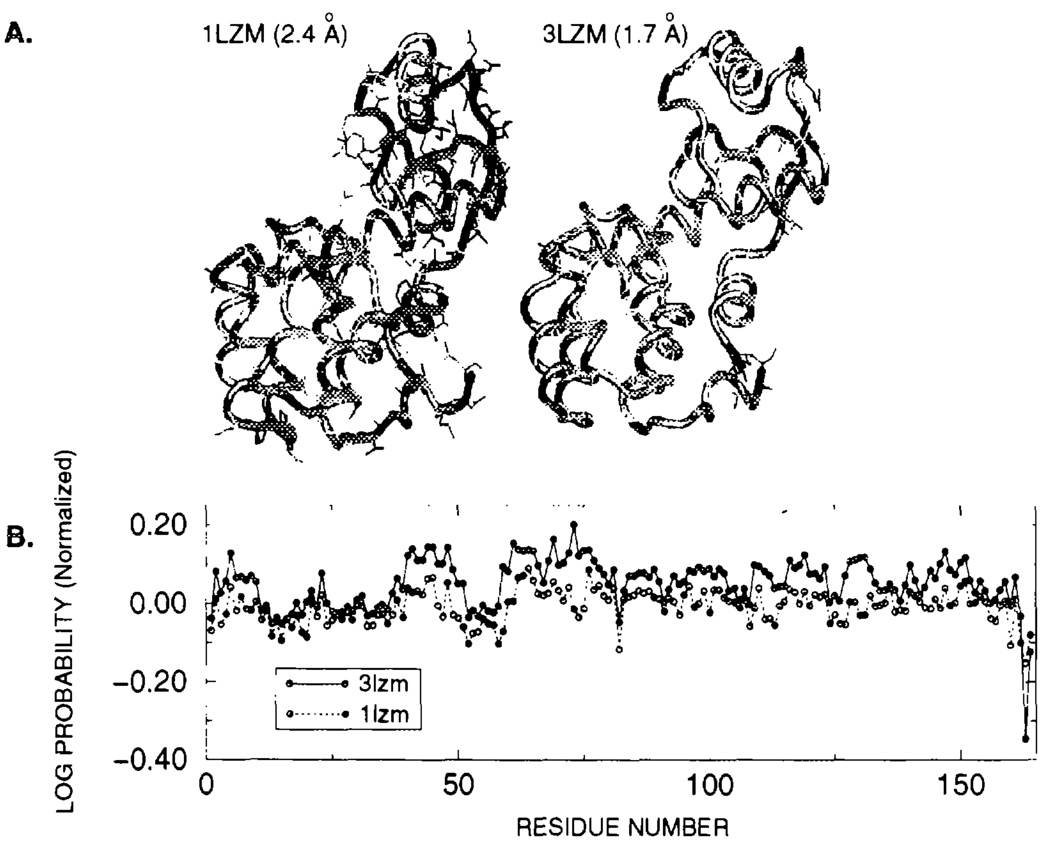


Figure 2. Comparison of (A) three dimensional structures of T4 phage lysozyme at 2.4 (1LZM) and 1.7 Å (3LZM) resolution, The darker and light shades represent low and high distance probability pairwise atomic contacts. (B) Residue profiles of 3LZM and 1LZM. The positive log(Probability) values indicate better pairwise atomic contacts averaged over all atoms for the residue. The C-terminal domain of 3LZM is better refined than that of 1LZM.

a higher probability score. To further illustrate the regions in the protein that have more ideal interactions as defined by the PDF scores, we show a comparison of the two structures and the residue-wise probability profiles of T4 phage lysozyme, 3lzm at 1.7 Å resolution and 1lzm at 2.4 Å resolution, in Fig. 2. The C-terminal domain is better resolved in 3lzm as reflected by the higher probability scores.

The x-ray structure of the single strand DNA-binding Gene V protein was solved by two groups of researchers and both used the x-ray data in conjunction with molecular mechanics methods to refine the structures (2GN5 and 0GVP). One of the structures (2GN5) had a phase shift error in the N-terminal domain of the protein although the overall topology of the two structures was similar. Despite the low molecular mechanics energies of both structures one of them was flawed by the phase shift. In Fig. 3, we compare the total and individual PDF residue profiles of the two structures. The overall residue profiles show clearly that the structure 2GN5 is not native-like owing to the large number of low distance probabilities. Analysis of the individual profiles reveals that the intra-residue and neighbour residue profiles obtained from pairwise interatomic interactions show that the PDF method is able to discriminate the non native-like local geometry of the residues. The structure 0GVP is consistent with native protein structures.

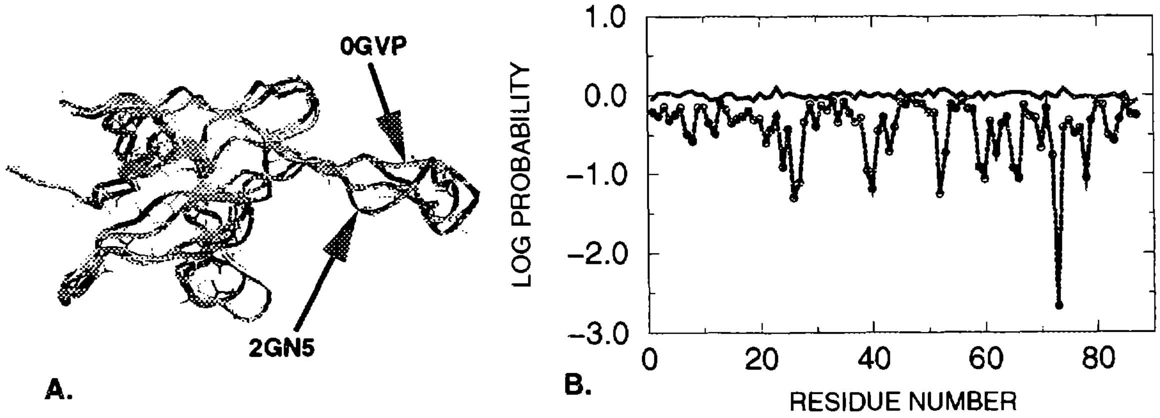


Figure 3. (A) Comparison of the three dimensional structures of the single-strand DNA binding Gene V protein (2GN5 and 0GVP). (B). Comparison of the residue profiles of the two structures.

Threading of a Sequence into Structural Motifs

In order to examine the value of PDFs in assessing structural fragments of naturally occuring proteins, we examined a structured fragment from the peptide GCN [28]. The structure of the native fragment is helix-like. We threaded the sequence into well-defined secondary-structural motifs extracted from high-resolution structures and energy-minimized the structures using molecular mechanics methods. The various structural motifs into which the sequence was threaded were analyzed for PDF profiles. Fig 4 presents the structural motifs and a comparison of the PDF profiles. Despite the very low molecular mechanics energies of the turn-like fragments, the PDF residue profiles show the highest probability for the native structure.

Annealing Noisy Protein Structures

An important use of the knowledge-based potentials is in the refinement of a non-native or poorly resolved protein structure to a native state. In conventional methods, optimization on an energy landscape is carried out to obtain the native protein structure. The potential functions that yield the energy landscape are based on either molecular mechanics-based, x-ray structure factors, NOE distance constraints or combinations of these (27). Molecular mechanics functions yield a very rough energy landscape and the resulting multiple minima render energy optimization complex. By virtue of being extremely specific, our statistical potentials, which describe each pairwise atomic interaction in accurate detail, yield a smooth and arguably unique minimum in the energy landscape and hence are ideally suited for folding non-native and noisy into native structures. We demonstrate the power of this method for diverse classes of proteins and suggest its possible use with low resolution x-ray or partial NOE data.

We use an annealing procedure to predict the structures of the proteins listed in Table 2, starting from noisy structures. Each of the initial structures was created by adding noise to the x-ray structure. Noise was added by defining a sphere with a 10 Å radius around each atom and randomly relocating the atom within the sphere using a uniform sampling distribution, i.e., all points within the sphere were treated as equally likely.

Table 2 shows the r.m.s. deviations of the 15 proteins annealed from a randomized structure as compared to the x-ray structures. These proteins are chosen so as represent diversity in size, packing and fold and are excluded from the set from which PDFs are constructed. In all of the examples studied, a well-connected compact topological structure was formed in the early stages of annealing, i.e., within 50 steps of optimization and the resulting structures are similar to the x-ray structures. A large contribution to the small RMS deviations from the x-ray structures stem from the solvent exposed side-chain orientations. The PDFs do not take into account explicitly the protein-solvent interactions or crystal contacts. However, they improve the interactions in the protein interior so as to optimize packing. We wish to note, that optimization using covalent constraints alone yielded structures which were sequentially connected but lacking both secondary structures and tertiary packing arrangements.

Table 2. RMS Deviation and Total Probabilities of Refined Proteins. The RMSDs refer to comparison of the x-ray structure.

Protein	Number of Residues (Number of Atoms)	All Atom-RMSD in Å (random)	All Atom-RMSD in Å (refined)	Back -bone Atom-RMSD in Å	^Total dist. log(prob)
Pancreatic Trypsin Inhibitor(9pti)	58 (453)	7.826	1.998	1.557	-0.026
Alpha Bungarotoxin (3ebx)	62 (474)	7.849	2.040	1.496	-0.042
Cytochrome B5 (3b5c)	85 (692)	7.810	2.085	1.696	-0.001
Plastocyanin (1plc)	99 (737)	7.803	2.024	1.633	-0.030
Parvalbumin (4cpv)	108 (806)	7.791	2.180	1.636	0.001
Cytochrome B562 (256b)	106 (825)	7.782	1.913	1.475	0.036
Pseudoazurin (2aza)	129 (975)	7.784	2.128	1.558	-0.028
Proteinase A (2sga)	181 (1258)	7.751	2.129	1.657	-0.041
Dihydrofolate Reductase (3dfr)	162 (1293)	7.743	1.930	1.517	-0.017
Lysozyme (3lzm)	164 (1308)	7.741	2.081	1.610	0.006
Alpha-Lytic Protease (2alp)	198 (1390)	7.738	2.144	1.618	-0.040
Actinidin (2act)	218 (1645)	7.754	2.312	1.801	-0.038
Acid Proteinase (2apr)	325 (2402)	7.787	2.142	1.687	-0.027
Thermolysin (3tln)	316 (2431)	7.784	2.051	1.655	-0.014
Glutathione Reductase (3grs)	461 (3498)	7.777	2.157	1.652	-0.016

^ The total distance log (probability) is both mean value and r-squared normalized.

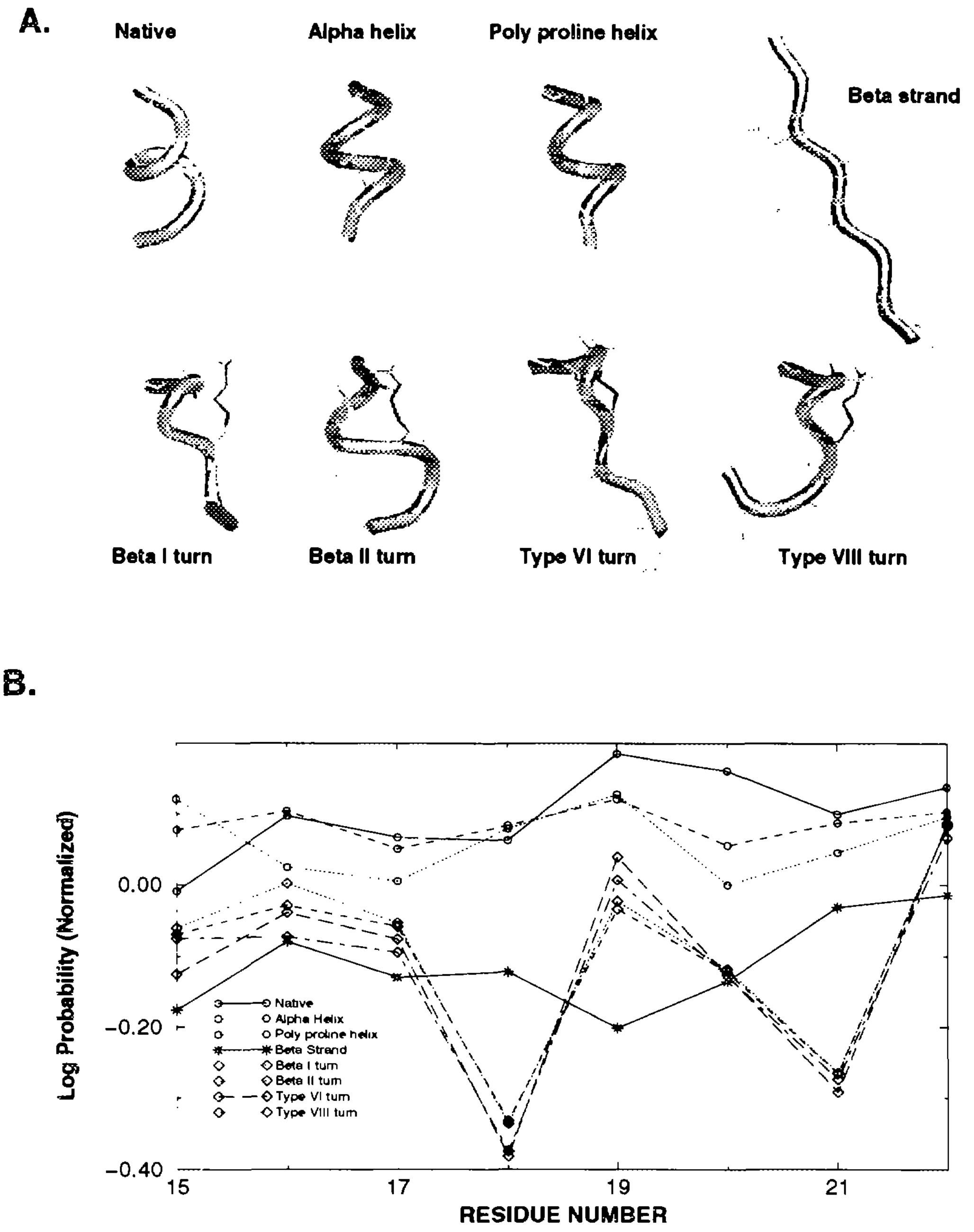


Figure 4. (A) The structural motifs into which the GCN peptide fragment is threaded. The structural motifs include native, standard alpha helix, poly proline helix, beta strand, beta I turn, beta II turn, type VI turn and Type VIII turn. (B) The residue profiles of the peptide in the above structural motifs are compared. The native structure has the highest overall profile.

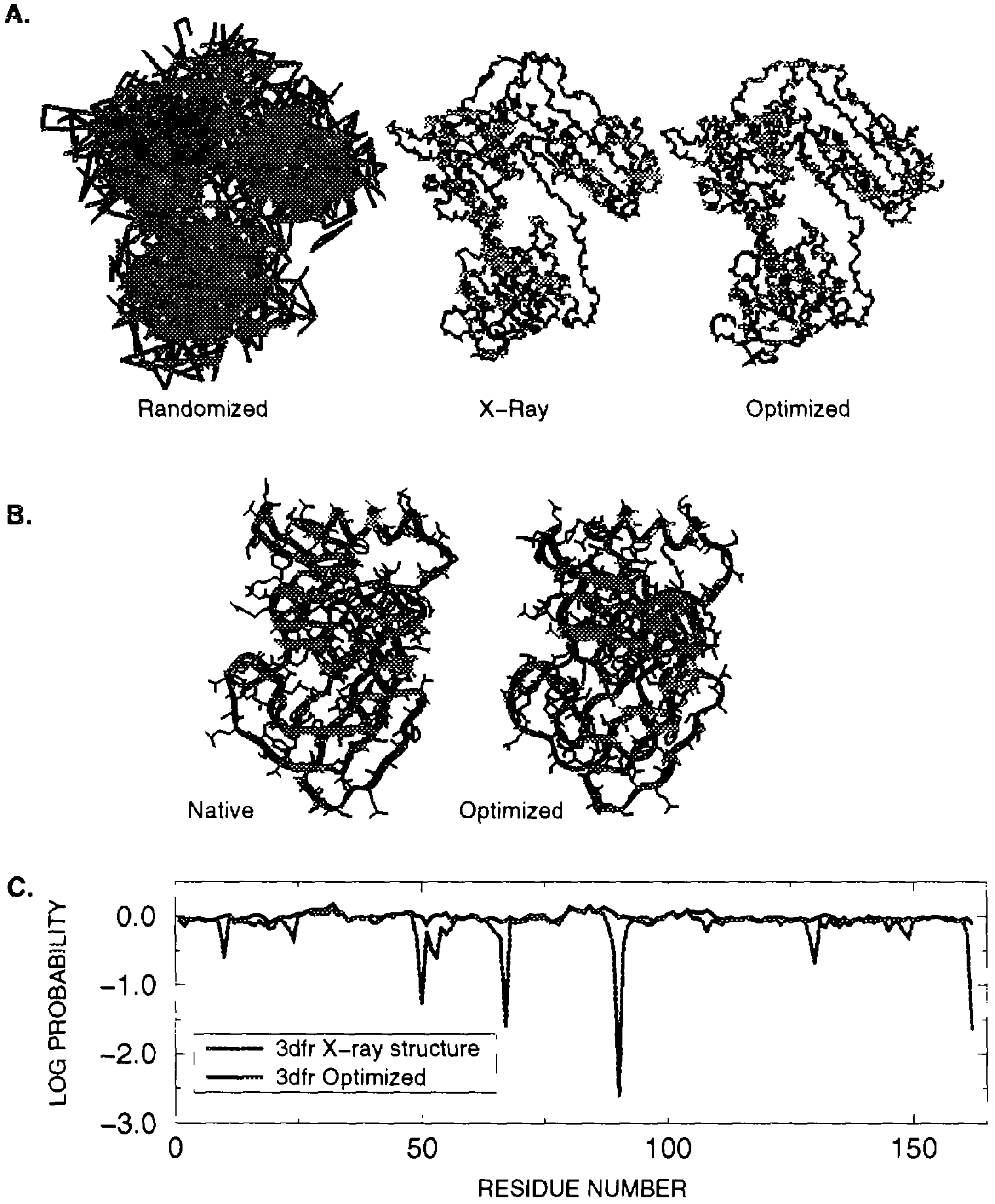
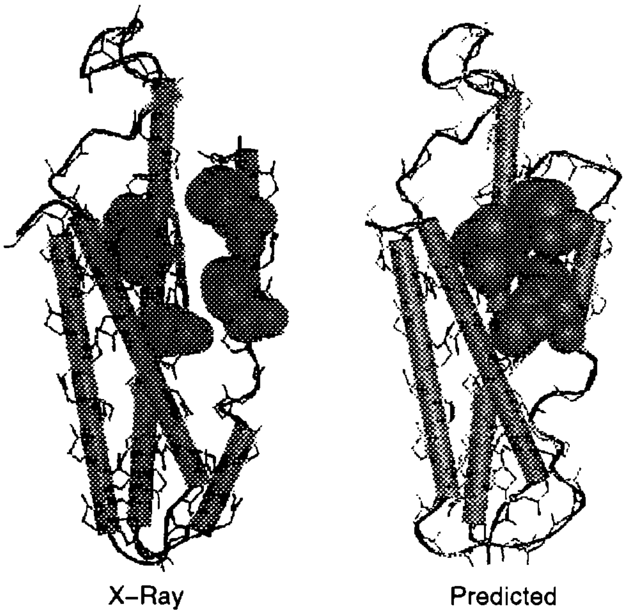
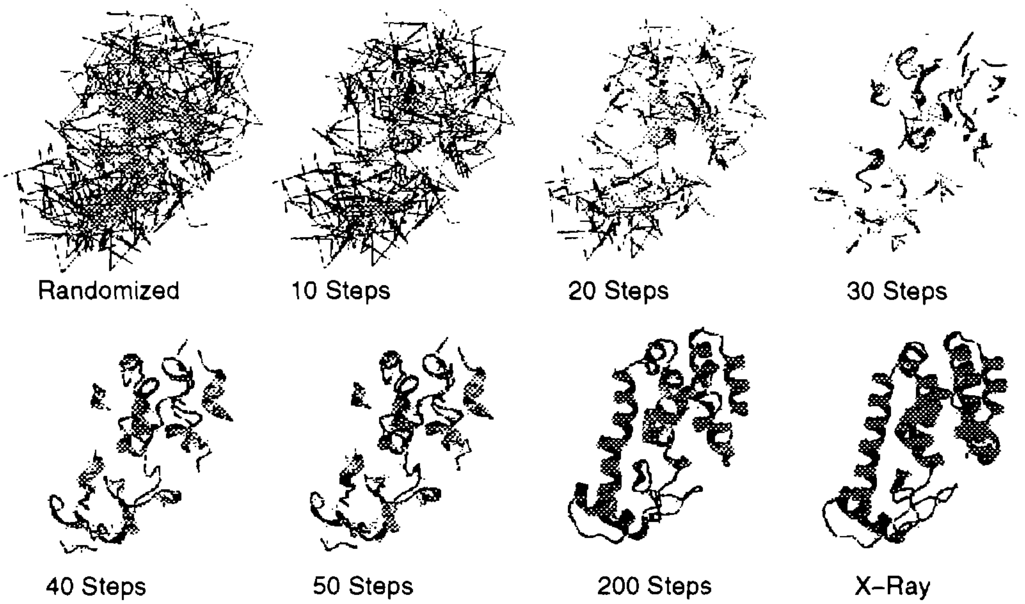


Figure 5. (A) Comparison of the randomized, annealed and x-ray structure of the 461 residue glutathione reductase (3grs). Only Ca atoms are displayed. (B) Comparison of the annealed and x-ray structures of the 162-residue dihydrofolate reductase (3dfr). The atoms are color coded such that the darker shade indicates lower pairwise atom distance probabilities. The x-ray structure has numerous improbable contacts (darker lines), which are corrected in the annealed structure (lighter lines). (C) Residue-wise probability profiles for the x-ray and refined dfr structures.



X-Ray Predicted

Figure 6. Comparison of the x-ray and annealed structures of cytochrome B562 (256b). Helices are represented by cylinders and the side chains thatrepack in PDF-based annealing are represented by spheres. The x-ray structure shows the heme group and the packing of the side chains and the annealed structure shows the packing reorganization; the heme group was not considered in the PDF-based annealing.



Randomized 10 Steps 20 Steps 30 Steps

40 Steps 50 Steps 200 Steps X-Ray

7. Structures of T4 phage lysozyme at different stages of annealing are shown. The color coding is darker (low) to lighter (high probabilities). The randomized structure anneals to a compact structure in the early stages and to secondary structures in the first 50 steps. The structure after 200 steps of annealing has achieved most of the secondary structure in the actual protein and has probabilities close to those in the x-ray structure.

In Fig. 5, we present the initial noisy, annealed and the x-ray structures of the two proteins, 461-residue glutathione reductase (3grs) and 162-residue dihydrofolate reductase (3dfr). In the latter, we also compare residue-averaged PDF profiles of the refined and the native x-ray structures. The x-ray structure has numerous improbable contacts, which are refined in the annealed structure.

In Fig. 6, we present the annealed and the x-ray structures of the protein cytochrome B562, (256b). We choose this protein to test the PDF potentials in assessing proteins that contain prosthetic groups. In the cytochrome B562, the non-inclusion of the prosthetic heme group in the PDF optimization does not appear to affect the overall structure. Fig. 4 shows that the side chain atoms of residues in contact with the heme, Met 7, Asn 11, Phe 65, Arg 98 and Arg 106, reorient to provide better packing in the optimized structure which does not contain the heme group. Helices 3 and 4 in the native protein move towards each other so as to yield better packing. Despite the deviations in the local region near the prosthetic group, the rest of the protein is annealed to the native structure.

We examined the annealing of the protein structures with the PDF method, by following the folding process in bacteriophage T4 lysozyme. In Fig. 7, we present different stages of annealing of a noisy structure, with the colors representing the transition of the structure from low probability and consequently high energy to high probability interactions. The early stages of annealing produces compactness, while the secondary structures are formed within circa 50 steps. At 200 steps of annealing the protein has optimized close to the actual structure. The folding process shows the specificity and hence the accuracy of these statistical potentials for native protein structure.

In summary, we have developed statistical potentials that describe the folded state of proteins accurately. These potentials are independent of protein sequence homologies, secondary structures or folds and contain information at the fundamental level of pairwise atomic interactions specific for each pair of residues. The distance-based statistical potentials appear to be ideally suited for combining with x-ray and NMR structural refinement methods.

Acknowledgements

We thank Drs. E. Jakobsson, S. Sligar, A. Crofts and C. Wraight for valuable discussions. We acknowledge a metacenter computer allocation and computer resources at the National Center for Supercomputing Applications.

References

1. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science* **181**, 223-230 (1973).

2. Dill, K.A. Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155 (1990).

3. Lattman, E.E. and Rose, G.D. Protein folding - What is the question? Proc. Natl. Acad. Sci. U.S.A. **90**, 439-441 (1993).

4. Levitt,M. and Warshel. A. Computer simulation of protein folding. *Nature* **253**, 694-698 (1975).

5. Godzik, A., Kolinski, A. and Skolnick, J. Are proteins ideal mixtures of amino acids - Analysis of energy parameter sets. *Protein Science* **10**, 2107-2117 (1995).

6. Elofsson, A., Le Grand, S.M., Eisenberg, D. Local moves - An efficient algorithm for simulation of protein folding. *Proteins: Struct. Funct. Genet.* **23**, 73-82 (1995).

7. McCammon, J.A. and Harvey, S.C., *Dynamics of Proteins and Nucleic Acids*. (Cambridge: Cambridge University Press, 1987).

8. Ponder, J.W. and Richards, F.M. Tertiary templates for proteins - Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791 (1987).

9. Srinivasan, R. and Rose, G.D. LINUS: A hierarchic procedure to predict the fold of a protein. *Proteins: Struct. Funct. Genet.* **22**, 81-99 (1995).

10. Brunger, A.T., Kuriyan, J. and Karplus, M. Crystallographic R-factor refinement by molecular dynamics. Science 235, 458-460 (1987).

11. Branden, C.I. and Jones, T.A. Between objectivity and subjectivity. *Nature* **343**, 687-689 (1990).

12. Crippen, G.M. Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry* **30**, 4232-4237 (1991).

13. Luthy, R., Bowie, J.U. and Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85 (1992).

14. Sali, A., Shakhnovich, E. and Karplus, M. How does protein fold? *Nature* **369**, 248-251 (1994).

15. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D. and Chan, H.S. Principles of protein folding - A perspective from simple exact models. *Protein Sci.* **4**, 561-602 (1995).

16. Karplus, M. and Sali, A. Theoretical studies of protein folding and unfolding. *Curr. Opin. Struct. Biol.* **5**, 58-73 (1995).

17. Sippl, M.J. and Weitckus, S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins: Struct. Funct. Genet.* **13**, 258-271 (1992).

18. Jones, D.T., Taylor, W.R. and Thornton, J.M. A new approach to protein fold recognition. *Nature* **358**, 86-89 (1992).

19. Holm, L. and Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-138 (1993).

20. Bryant, S.H. and Lawrence, C.E. An empirical energy function for threading protein sequence through the folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92-112 (1993).

21. Hearst, D.P. and Cohen, F.E. GRAFTER - A computational aid for the design of novel proteins. *Prot. Eng.* **7**, 1411-1421 (1994).

22. Sali, A. and Blundell, T.L. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815 (1993).

23. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859-883 (1990).

24. Sun, S. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Sci.* **2**, 762-785 (1993).

25. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. *J. Mol. Biol.* **215**, 403-410 (1990).

26. Hogg, R.V. and Tanis, E.A. *Probability and Statistical Inference* (New York: MacMillan Publ. Company, 1988).

27. Brunger, A. T. and Nilges, M. Computational challenges for macromolecular structure determination by x-ray crystallography and solution NMR-spectroscopy. *Q. Rev. Biophys.* **26**, 49-125 (1993).

28. Walsh, L.L. *An annotated guide to the Brookhaven Protein Databank, classification and comparisons of protein structures.* (Doctoral Dissertation Submitted to the University of Illinois, 1994).

29. The color-coded versions of figures in the manuscript can be found on the WWW, whose URL is http://bioweb.ncsa.uiuc.edu/doc/ISMB96.html.