Learning About Protein Folding via Potential Functions

V.N. Maiorov and G.M. Crippen
College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109

ABSTRACT **Over the last few years we have developed an empirical potential function that solves the *protein structure recognition problem*: given the sequence for an *n*-residue globular protein and a collection of plausible protein conformations, including the native conformation for that sequence, identify the correct, native conformation. Having determined this potential on the basis of only some 6500 native/nonnative pairs of structures for 58 proteins, we find it recognizes the native conformation for essentially all compact, soluble, globular proteins having known native conformations in comparisons with 10^4 to 10^6 reasonable alternative conformations apiece. In this sense, the potential encodes nearly all the essential features of globular protein conformational preference. In addition it "knows" about many additional factors in protein folding, such as the stabilization of multimeric proteins, quaternary structure, the role of disulfide bridges and ligands, *pro*proteins vs. processed proteins, and minimal strand lengths in globular proteins. Comparisons are made with other sorts of protein folding problems, and applications in protein conformational determination and prediction are discussed.**

Key words: **globular proteins, tertiary structure, quaternary structure, folding determinants, disulfide bonds, polypeptide conformation, homology modelling, inverse folding problem**

DETERMINING THE POTENTIAL FUNCTION

In spite of recent attention to chaperonins and the problems of in vivo protein synthesis, export, and folding, the experimental fact remains that many proteins fold reversibly in vitro to a unique native conformation, apparently determined solely by their amino acid sequence.[1] To be able to predict the conformation from the sequence is possibly the most important current problem in molecular biology, especially now that so many genes are being sequenced. Surely a necessary step toward this long-range goal is to be able to distinguish between the native conformation and an assortment of alternative folds. Many researchers have addressed this and related questions.[2–8] Recently we have solved this "protein structure recognition problem" to a high degree of generality by devising a potential function of the amino acid sequence and three-dimensional structure that gives a clearly lower function value for the native than for any alternative.[9]

Figure 1 schematically summarizes our experimental knowledge about protein folding. Nature has sampled only an infinitesimal portion of all possible amino acid sequences and has tended to keep either fibrous, membrane, or globular proteins. While it is not at all clear what proportion of all sequences will fold up to unique globular structures,[10] many naturally occurring ones do (small circles in the figure) and even some carefully designed sequences do not (large circles). Substantially different sequences tend to fold up rather differently, with some remarkable exceptions,[11] but very similar sequences generally fold up to similar structures. This is the basis for all the activity in protein homology modeling. We are unaware of any example of two very similar sequences folding to very different, unique, globular structures.

We now have from X-ray crystallography and NMR more than 1000 examples of the native conformations of various water soluble globular proteins under roughly physiological conditions, along with their sequences. In spite of a vast effort on the part of many investigators, the repeatedly occurring structural patterns or motifs[12] do not correlate clearly enough with sequence to be able to deduce a set of rules for going reliably from sequence to local or secondary structure,[13–15] much less to the overall tertiary conformation. In fact, a five-residue segment of a given sequence may occur in quite different conformations in the context of different protein crystal structures, and the percentage of sequence identity between two segments required to give a high likelihood of conformational similarity decreases only gradually with increasing chain length.[16] The problem is that, just as a fragment of the polypeptide chain comprising a protein domain will not fold up, neither can one calculate the con-

Received February 25, 1994; revision accepted June 8, 1994.
Address reprint requests to G.M. Crippen, College of Pharmacy, University of Michigan, Ann Arbor, MI 48109.

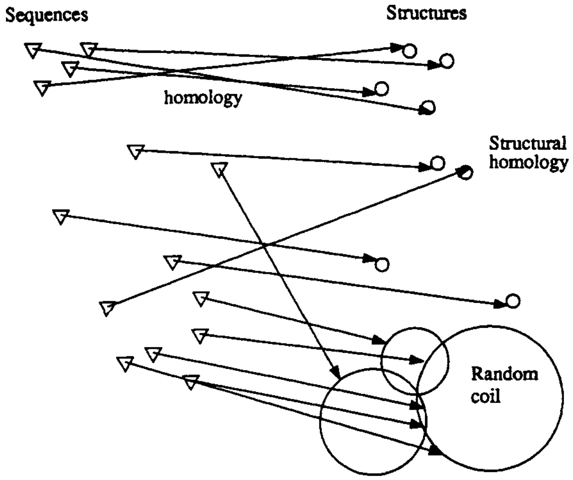

Fig. 1. Our current experimental knowledge about the correspondence between amino acid sequence and protein conformation. Triangles represent sequences, and their proximity corresponds to sequence similarity. Small circles represent precise globular folds, large circles are more disordered states, and proximity again implies similarity.

formation of a protein by looking at only a piece of it at a time. Certainly there are tendencies, and secondary structure prediction is about 70% correct, but the overall tertiary folding modifies the tendencies in the other 30% of the residues.

While most investigators have induced principles of protein folding from the experimental examples by focusing on the arrows of Figure 1, we have exploited in addition a rich source of information: the *lack* of arrows. It is certainly true that some sequences fold up to unique globular conformations, but it is just as true that those same sequences do not fold to any other of the known protein conformations. The key to the success of our potential function is that it is trained to give a lower value for the native conformation than for any alternative conformation. As we have described in great detail elsewhere,[9] we assumed that an adequate function could be defined as the sum of contributions of interresidue contacts, each of which depends on the sequence separation and amino acid types involved. Although the potential is defined to be nonlinear in the atomic coordinates, it is simply linear in the adjustable parameters, which are the weights assigned to the 112 different types of contacts we defined.[9]

Our unusual method for determining the parameters goes from a relatively small set of critical inputs to a potential that is in agreement with, and unaffected by, much larger sets of novel information about protein conformation. The reason is that every constraint that a particular training protein's native structure should have a lower contact potential value than that of some alternative conformation corresponds to a linear inequality in the space of the adjustable parameters. Such a set of linear inequalities may be inconsistent (which is fortunately not the case here), or there will be some feasible region, any point of which corresponds to a satisfactory potential function. Figure 2 illustrates our situation, where the feasible region turns out to be infinite but bounded on several sides by a relatively small number of inequalities. We choose the unique point that gives the smallest magnitudes of the parameters. The vast majority of inequalities are redundant, meaning they cannot affect the feasible region at all. The potential function is determined on the basis of a small training set of native proteins and a modest set of alternative conformations of these. Some relatively small subset of the corresponding inequalities outline the feasible region, and the rest are redundant. Introducing additional native and alternative conformations into the training set rarely slices a corner off the old feasible region, and even more rarely does it cut away the old optimal point. In other words, after the potential has been trained up to a certain level, novel protein folding motifs and newly determined crystal structures generally have no effect on the potential function at all. It has learned all it needs to know about protein conformations, even though some new structures may appear to us as unprecedented. On the other hand, if we try to require some erroneous conformer to be preferred over an intrinsically better alternative, we may cut away the entire feasible region. In contrast, least squares and mean field approaches[7,8,17–21] are more fault tolerant, but they always change the potential at least a little as the training set is increased. Another difference is that our potential makes no pretense of being based on statistical mechanical theory, so our potential values are not given in some units of energy.

The latest version of the potential was determined by a training set consisting of 58 compact, single chain proteins (50 X-ray + 8 NMR), and a few of their alternatives, totalling 6566 constraints. The native conformations are from the Protein Data Bank (PDB), and the alternatives are contiguous chain segments of the appropriate length cut out of larger PDB structures. Compact polypeptide conformers satisfy certain precise requirements of small radius of gyration and large numbers of interresidue contacts.[9] We do not demand the potential work for noncompact natives, but some noncompact alternatives are important in determining the parameters. Of course the potential satisfied these 6566 constraints, but we tested it on a superset of altogether 212 native proteins (176 X-ray + 17 NMR + 19 models) and 477 homologues of these, contrasted to a grand total of 1,627,714 alternatives. Of these 212, only 161 are compact (totalling 1,242,182 alternatives). Out of the noncompact proteins, just 21 (15 X-ray + 5 NMR + 1 model) had a few violations, and transforming growth factor α (4tgf, by NMR) had many. Six compact test proteins

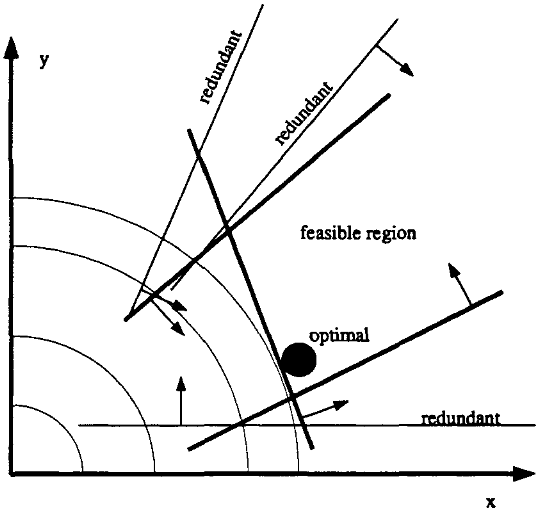

Fig. 2. Example of three redundant and three nonredundant linear inequalities in two variables, x and y, defining an infinite feasible region. Each inequality permits the half-plane indicated by the boundary line and arrow. One solution, labeled as optimal, satisfies all the inequalities, has only one active constraint, and otherwise minimizes the objective function, $x^2 + y^2$, which is indicated by circular level lines.

had in all 12 violations. Model structures (1mca.A, 1apk, and 2apk) account for 9 of these violations, but this says more about the calculations that determined them than about our agreement with experiment. One violation for pike parvalbumin (1pal) is due to a very similar structure from the homologous carp parvalbumin (5cpv), which is an example of the desirable property that very similar conformations have very close potential function values. Finally, there was one violation for each for myohemerythrin (2mhr) and chain 1 of bean pod mottle virus (1bmv.1). These two constraints need to be introduced into the training set when the next update of the potential is calculated. Still, two genuine errors out of 1,200,000 tests gives one confidence the potential has learned a lot about a wide variety of proteins.

LESSONS FROM THE POTENTIAL

The only factors determining the value of the potential for a particular protein's conformation are the amino acid sequence and coordinates of the backbone and C^{β} atoms. While it favors contacts between Cys residues as it does between hyrophobic residues in general, it does not need to know which are reduced and which are paired in disulfide bridges. Yet the potential correctly favors the native conformation over alternatives where disulfide bridges are broken or cysteines are incorrectly paired. This is in agreement with experiments that the correct crosslinks will form spontaneously for globular proteins under appropriate renaturing conditions. The general consensus reached from these sorts of experiments is that for single polypeptide chains over roughly 50 residues in length, the disulfide bridges are a stabilizing but not determining factor in the overall tertiary conformation.[22] Similarly, the potential disregards all prosthetic groups, organic ligands, and specifically bound metal ions. It is well known experimentally that adding these may raise the melting temperature considerably and reduce conformational fluctuations at room temperature, but since the potential considers only interactions between amino acid residues in correctly recognizing the native conformation, we conclude the general folding motif is encoded in the polypeptide chain alone.

Even though the potential was determined on the basis of single-chain proteins, it is applicable without modification to multichain aggregates. It is sufficient to count interactions between separate polypeptide chains the same as between residues on one chain that are separated greatly in sequence. The native conformations of 10 compact two-chain proteins (2ltn, 2fbj, 2fdl, 4fab, 3fab, 3hfm, 1f19, 2igf, 1mcp, and 2ig2) were correctly favored over the 500,000 to 2000 alternatives we could generate for each. Since we produce alternatives by taking pieces out of larger PDB structures, there are several large multimeric proteins (>400 residues) for which we have very few alternatives, making them weak tests of the potential. At the other extreme, a single melittin chain (2mlt) is a noncompact, 28 residue, bent helix, for which we can generate many alternatives. Not surprisingly, the potential favors a more compact alternative conformation. Even the native dimer is noncompact and has many preferred alternatives. However, the tretramer is compact and has no violations in the 1.28 million alternatives checked. This in agreement with the experimental fact that melittin exists as a tetramer in solution and in the crystal structure.[23] One cannot and should not expect the monomer to be preferred by the potential when it has so few internal contacts.*

Insulin is another "correct failure" of the potential function. In the crystal structure (3ins), the A and B chains are joined by disulfide bridges and together pass our criteria for compactness. However, alternatives are generated that keep each chain intact but pay no attention to the formation of any—much less the correct—bridges. We are easily able to locate many alternatives that the potential prefers over the native. The situation is not improved by considering the insulin dimer seen in the crystallographic

*On the other hand, we were unable to select a small oligomeric subset of the crystal structure for avian pancreatic polypeptide (1ppt) that was compact. In the absence of well-documented experimental studies on its aggregation state in solution, much less its solution conformation, we can only speculate that this may differ significantly from what is seen in the crystal structure.

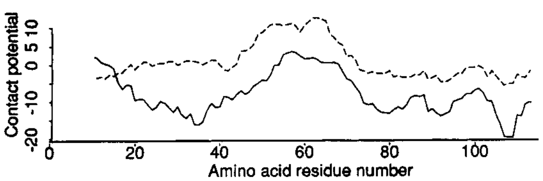

Fig. 3. RuBisCO contact potential vs. residue number, smoothed. Solid line: correct sequence on correct structure, $E = -894$. Dashed line: correct sequence on incorrect structure, $E = +22$.

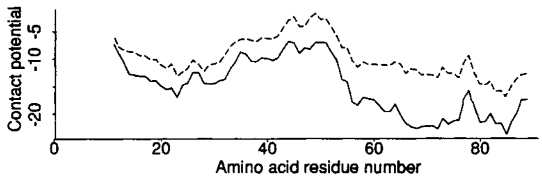

Fig. 4. HIV protease contact potential as a function of residue number, smoothed over a 21-residue window. Solid line: 5hvp sequence on 5hvp structure, $E = -1384$. Dashed line: 5hvp sequence on 1hvp structure, $E = -922$.

asymmetric unit. On the other hand, one must realize the closest corresponding experiment would be the reoxidation of the disulfide bridges of fully reduced insulin, which in fact produces a random mixture of different cysteine pairings.[24] Thus the potential correctly predicts that insulin cannot refold. However, if there were an X-ray or NMR structure available for *pro*insulin, we would expect the potential to favor the native over all alternatives.

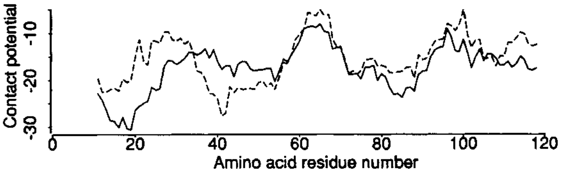

Fig. 5. Interleukin 4 contact potential vs. residue number, smoothed. Solid line: NMR sequence on NMR structure, $E = -2216$. Dashed line: NMR sequence on model structure, $E = -1996$.

APPLICATIONS TO STRUCTURE MODELING AND DETERMINATION

The potential function easily prefers the correct structure over the intentionally incorrect fold in all 12 published cases we could readily check.[25,26] A somewhat less artificial test is the comparison of the correct X-ray structure of RuBisCO[27] with the earlier incorrect chain tracing.[28] Of course, the decisive comparison is between the sums over all contacts for the two different conformations, but it is also interesting to see the contributions from different parts of the chain. As is typical for these sorts of potentials, the sum of the contacts for each residue varies abruptly for even sequentially adjacent residues because their amino acid types may be quite different. However, if we plot the mean contribution to the potential over a 21-residue sliding window centered on residue numbers 11 to n-11, one can sometimes observe local problems with a conformation. Moreover, we find that all 21-residue segments in the native have a net more favorable interaction (within the segment and between residues of the segment and residues outside it) than the corresponding segment does in any alternative. This is in agreement with the general observation that strands and helices constituting the core structural elements of globular proteins tend to be significantly shorter,[29] so that the 21-residue window looks beyond pure secondary structure to the beginnings of tertiary folding. In the case of RuBisCO (Fig. 3), the incorrect structure[28] is inferior to the correct one[27] almost everywhere uniformly down the chain. The main difference between the two is the initial assignment of amino acid sequence to the electron density was actually reversed from the correct fitting. Similarly, the X-ray structure of HIV protease (5hvp)[30] is everywhere an improvement over the earlier model structure (1hvp),[31] as shown in Figure 4.

The profile for the NMR structure[32] of interleukin 4 is more closely matched by that of the model structure[33] (Fig. 5), which is just the mirror image arrangement of the same 4-helix bundle, but the overall contact potential clearly prefers the native. In the case of muconate lactonizing enzyme (1mle)[34–36] vs. mandelate racemase (1mns),[37] which have very similar structures but little sequence similarity, each sequence prefers the more highly refined 1mns crystal structure, but by a small margin (Figs. 6 and 7).[†]

OTHER APPLICATIONS

Our experience is that the native conformation is preferred over substantially different folds almost invariably. Small variations on the order of 1–2 Å in backbone conformation from the native generally have similar potential values, but may drop slightly below that of the native. All other factors being equal, tightly packed, high resolution crystal structures are generally preferred over lower resolution structures or NMR structures. For other kinds of applications, one must keep in mind that the potential is designed to handle the structure recognition problem, and substantially different uses may correspond to quite different experimental situations. For example, one might try the potential on the "sequence recognition problem," also called the inverse folding problem. Here, one starts with some protein

[†] More precisely, what we compared was a 150 residue ungapped alignment having C^α root mean square coordinate deviation of 1.96 Å and 28% sequence identity.

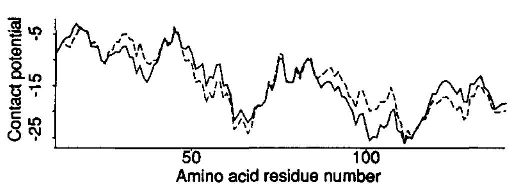

Fig. 6. Solid line: 1mle sequence on 1mle structure, $E = -2144$. Dashed line: 1mle sequence on 1mns structure, $E = -2177$.

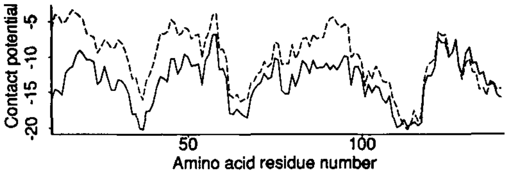

Fig. 7. Solid line: 1mns sequence on 1mns structure, $E = -2055$. Dashed line: 1mns sequence on 1mle structure, $E = -1482$.

of known native conformation and sequence, and tries to find other sequences that prefer that one, fixed conformation. The idea is to identify probable biochemical roles for new sequences even when they have insignificant sequence similarity to known proteins. Indeed, our potential usually prefers to thread the native sequence onto the native conformation over thousands of alternative contiguous subsequences taken from larger proteins, but it is not nearly as reliable as its ability to recognize the correct structure. As Sippl has pointed out, such a potential is built to have a unique global minimum at the native conformation as *conformation* is varied, holding the sequence constant, but there is no particular guarantee that it has a global minimum at the native sequence as the *sequence* is varied, holding the conformation constant.[38]

Clearly any successful sequence recognition algorithm must also treat insertions and deletions, just as it is essential for sequence alignment methods. Conventional wisdom holds that the native conformation of a globular protein is uniquely stabilized by interactions among the interior "core" segments of the polypeptide chain, as opposed to the more solvent exposed exterior "loop" and terminal segments. Since the sequence and three-dimensional structure of the core segments both tend to be conserved over a series of homologous proteins, while those of the loop segments are not, there must be some characteristic pattern of amino acid types in the core that determines the common folding motif. There are two different computer experiments that one might propose to explore this line of reasoning. Suppose we delete the loop portions of the sequence of a compact protein and try all different ways to slide the now disconnected core segments of the native sequence along the contiguous native conformational template.[7,8] Any part of the structure not matched to an amino acid of the core sequence strands would make no contribution to the potential function value. Alternatively, we could delete the loop segments of the native (or any other protein's) structure and try different alignments of the full native sequence onto the fixed spatial positions of the core segments. Any part of the sequence not currently matched to structure would in effect vanish temporarily. In either case, we find that our potential may not even prefer the native alignment of native sequence onto the native conformation over alternative alignments! There are three reasons why this is nevertheless the correct outcome. (1) In an examination of several small proteins, we note that for the full native sequence on the full native structure, core–core interactions account for anywhere between 30 and 80% of the total potential function value. Perhaps our potential is in error when it assigns significant stabilizing core–loop and loop–loop interactions, or perhaps the subjective definition of core vs. loop needs to be refined. Yet given the general success of the potential at structure recognition even for multichain proteins, the widely assumed importance of core–core interactions may be overestimated. (2) Think of the corresponding experiment: would those disconnected polypeptide strands spontaneously aggregate to form the core structure of the original protein? Given that even small deletions of chain can destroy a protein's ability to refold, it seems unlikely they would. If we had a truly universally correct potential function for the structure recognition problem, then the observation that it prefers a nonnative alignment over the native one is completely irrelevant. Instead, it should prefer even more strongly some (perhaps many) nonnative conformation for these independently moveable segments of polypeptide chain, and this structure should correspond to the experimentally observed state: a unique conformation, a molten globule, or random coil. (3) These alignment calculations have introduced degrees of freedom into the problem that the real protein does not have. Any part of the template structure not assigned some part of the sequence, or any part of the sequence not assigned to some position in the structure, effectively disappears in the calculation. On the other hand, if the real protein tries to fold by pushing the end of a core strand out onto the globule's exterior, these residues continue to interact with the new environment made up of more solvent and neighboring surface residues. Although it is impractical, the correct calculation would be an extension of homology modeling, where different alignments of sequence onto core segments are compared only after the loop segments have been optimally positioned in space.

Why then do profile methods, such as those of the Eisenberg and Sander groups, work for the sequence recognition problem?[2–5] In order to understand this, we must first remind ourselves of their general protocol, without getting into a lot of detail. From some sort of survey of known protein structures they first derive what we will term the "universal part" of a potential that encodes the preferences of different amino acid types for different environments (secondary structure assignment, solvent exposure, etc.). Then in order to recognize sequences that might fold to a particular core structure, they choose one accurately known protein structure of that type and calculate the detailed environment at each sequence position down the chain. The "specialized part" depends in general on the three-dimensional structure and even the sequence of the template protein chosen. Of course, the specialized part is quite different when looking for globins compared to seeking β-barrels, for example. We can think of solving the sequence recognition problem as a global search for the optimum value of some function as we vary some set of variables. The variables are the different amino acid sequences tried along with the choices of insertions and deletions relative to the template protein; the function is the sum of the (universal) preferences of the different amino acid residues for the (specialized) environments they find themselves in, given the particular alignment being tested. For one given sequence, it is feasible to carry out a full global search over the alignment variables and find the optimum function value. While it is not feasible to optimize also over all possible sequences, it is scientifically valuable to simply examine different members of a protein sequence database.

Comparing profile methods to a potential such as ours reveals three main differences. (1) Our potential has only a universal part and no counterpart to the specialized profile derived from a template protein. When our potential fails to recognize the native alignment of the native sequence on parts of the native structure, it does so in part because it does not have the same help from a specialized part. Conversely, applying a profile method to the structure recognition problem would be unlikely to work for a variety of proteins having different native folding motifs because the specialized part would bias the results toward whatever single template protein was used.[39] (2) Evaluating the function for a profile method can have less correspondence to a real physical situation than our potential has, and these nonphysical features bias the profile results toward success in the sequence recognition problem. For example, if an alignment fails to supply residues on nearby chains surrounding some position that was buried in the original template structure, the environment of that position nevertheless remains "buried." In contrast, our potential function would recognize the residue as now exposed, and would prefer putting a hydrophilic residue there, instead of the hydrophobic one found in the template protein's sequence. (3) The alignment variables and the mathematical form of the function optimized in profile methods have been chosen so that the global search is computationally feasible. In order to maintain the greater physical realism and universality of our potential, the global search over alignments in our case is more time-consuming. The situation becomes yet more difficult for the intended application of our potential, namely the structure recognition problem, because now the variables for the global search are polypeptide conformation parameters, a larger and more complicated space than gapped sequences.

The point of this lengthy comparison is that both sequence recognition and structure recognition are worthwhile scientific goals, but they are fundamentally different problems that must be attacked by different methods. Not only is it unreasonable to expect that a successful method for the one problem ought to perform well at the other, it is even an illogical expectation, once you examine the situation.

Aside from these interesting questions of structure recognition vs. sequence recognition, there are some entirely suitable applications of our potential that are immediately practical. Heretofore, protein engineering or ab initio protein design have focused on increasing the stability of a target native conformation by altering the amino acid sequence, relative to the starting sequence. The danger is that while some sequence alteration may stabilize the native, it may stabilize some very different conformation even more. Our experience with the contact potential strongly suggests that it is vital to increase the stability (i.e., decrease the contact potential value) relative to that of a large number of alternative conformations. Only in this way can one guard against designing a sequence that folds up in some undesired way or indeed has no unique native conformation at all. People tend to concentrate on a single target and find it difficult to keep in mind a universe of alternative outcomes. Our simple methods for generating large numbers of alternative conformations and evaluating their relative quality make this easy.

Clearly the contact potential has immediate uses whenever one needs to select one of several alternative conformations produced from theory or experiment: initial chain tracing when fitting low-resolution electron density maps in X-ray crystallography, ranking the ensemble of structures calculated from NMR conformational studies on proteins in solution, and protein conformation calculation by homology modeling. Since multimeric proteins are handled as well as those with only a single chain, the potential would be useful for choosing the best docking of two aggregating proteins, such as an enzyme and its small inhibitor protein, or for predicting protein quaternary structure in general. Given its surprising generality, there is a wide field of future applications of the potential.[39]

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (GM37123) and the National Institute on Drug Abuse (DA06746). We are indebted to all the crystallographers and NMR spectroscopists who deposited their protein structural data in the Protein Data Bank.

REFERENCES

1. Seckler, R., Jaenicke, R. Protein folding and protein refolding. FASEB J. 6, 2545–52, 1992.
2. Bowie, J. U., Lüthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
3. Wilmanns, M., Eisenberg, D. Three-dimensional profiles from residue-pair preferences: Identification of sequences with β/α-barrel fold. Proc. Natl. Acad. Sci. U.S.A. 90:1379–1383, 1993.
4. Lüthy, R., Bowie, J. U., Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature (London) 356:83–85, 1992.
5. Ouzounis, C., Sander, C., Scharf, M., Schneider, R. Prediction of protein structure by evaluation of sequence-structure fitness. J. Mol. Biol. 232:805–825, 1993.
6. Godzik, A., Skolnick, J. Sequence-structure matching in globular proteins: Application to supersecondary and tertiary structure determination. Proc. Natl. Acad. Sci. U.S.A. 89:12098–12102, 1992.
7. Jones, D. T., Taylor, W. R., Thornton, J. M. A new approach to protein fold recognition. Nature (London) 358: 86–89, 1992.
8. Bryant, S. H., Lawrence, C. E. An empirical energy function for threading protein sequence. Proteins 16:92–112, 1993.
9. Maiorov, V. N., Crippen, G. M. Contact potential that recognizes the correct folding of globular proteins. J. Mol. Biol. 227:876–888, 1992.
10. Shakhnovich, E., Farzitdinov, G., Gutin, A. M., Karplus, M. Protein folding bottlenecks: A lattice Monte Carlo simulation. Phys. Rev. Lett. 67:1665–1668, 1991.
11. Pastore, A., Lesk, A. Comparison of the structures of globins and phycocyanins: Evidence for evolutionary relationship. Proteins 8:133–155, 1990.
12. Chothia, C., Finkelstein, A. V. The classification and origins of protein folding patterns. Annu. Rev. Biochem. 59: 1007–1039, 1990.
13. Rooman, M. J., Rodriguez, J., Wodak, S. J. Relations between protein sequence and structure and their significance. J. Mol. Biol. 213:337–350, 1990.
14. Rooman, M. J., Kocher, J. P. A., Wodak, S. J. Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. Biochemistry 31:10226–10238, 1992.
15. Rackovsky, S. On the nature of the protein folding code. Proc. Natl. Acad. Sci. U.S.A. 90:644–648, 1993.
16. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
17. Oobatake, M., Crippen, G. M. Residue-residue potential function for conformational analysis of proteins. J. Phys. Chem. 85:1187–1197, 1981.
18. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, M. J. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J. Mol. Biol. 216: 167–180, 1990.
19. Hinds, D. A., Levitt, M. A lattice model for protein structure prediction at low resolution. Proc. Natl. Acad. Sci. U.S.A. 89:2536–2540, 1992.
20. Miyazawa, S., Jernigan, R. L. Estimation of interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18:535–552, 1985.
21. Godzik, A., Kolinski, A., Skolnick, J. Topology fingerprint approach to the inverse protein folding problem. J. Mol. Biol. 227:227–238, 1992.
22. Creighton, T. E. Disulphide bonds and protein stability. BioEssays 8:57–63, 1988.
23. Terwilliger, T., Eisenberg, D. The structure of melittin. I. Structure determination and partial refinement. J. Biol. Chem. 257:6016–6022, 1982.
24. Srinivasa, B. R. Sulfhydryl oxidation of reduced insulin in dilute solution, Biochem. Int. 9:523–529, 1984.
25. Novotny, J., Rashin, A. A., Bruccoleri, R. E. Criteria that discriminate between native proteins and incorrectly folded models. Proteins: Struct. Funct. Genet. 4:19–30, 1988.
26. Holm, L., Sander, C. Evaluation of protein models by atomic solvation preference. J. Mol. Biol. 225:93–105, 1992.
27. Curmi, P., Cascio, D., Sweet, R., Eisenberg, D., Schreuder, H. Crystal structure of the unactivated form of ribulose-1,5-bisphosphate carboxylase/oxygenase from tobacco refined at 2.0-Å resolution. J. Biol. Chem. 267:16980–16989, 1992.
28. Eisenberg, D. Personal communication.
29. Dill, K. A. Dominant forces in protein folding. Biochemistry 29, 7133–7155.
30. Fitzgerald, P. M. D., McKeever, B. M., vanMiddlesworth, J. F., Springer, J. P., Heimbac, J. C., Leu, C.-T., Herber, W. K., Dixon, R. A. F., Darke, P. L. Crystallographic analysis of a complex between human immunodeficiency virus type 1 proteins and acetylpeptstatin at 2.0-angstroms resolution. J. Biol. Chem. 265:14209–14219, 1990.
31. Weber, I., Miller, M., Jaskolski, M., Leis, J., Skalka, A., Wlodawer, A. Molecular modeling of the HIV-1 protease and its substrate binding site. Science 243:928–931, 1989.
32. Smith, L., Redfield, C., Boyd, J., Lawrence, G., Edwards, R. G., Smith, R. A. G., Dobson, G. M. Human interleukin 4: The solution structure of a four-helix-bundle protein. J. Mol. Biol. 224:899–904, 1992.
33. Cohen, F., Presnell, S. Personal communication.
34. Rice, P. A., Goldman, A., Steitz, T. A helix-turn-strand structural motif common in alpha-beta proteins. Proteins 8:334–340, 1990.
35. Goldman, A., Ollis, D. L., Steitz, T. A. Crystal structure of muconate lactonizing enzyme at 3 angstroms resolution. J. Mol. Biol. 194:143–153, 1987.
36. Sander, C., Holm, L. Muconate lactonizing enzyme heavy atom coordinates calculated from available X-ray C^α coordinates. Personal communication, 1993.
37. Landro, J. A., Gerlt, J. A., Kozarich, J. W., Koo, C. W., Shah, V. J., Kenyon, G. L., Neidhart, D. J., Fujita, S., Petsko, G. A. The role of lysine 166 in the mechanism of mandelate racemase from Pseudomonas putida: Mechanistic and crystallographic evidence for stereospecific alkylation by (r)-alpha-phenylglycidate. Biochemistry 33:635–643, 1994.
38. Sippl, M. J. Boltzmann's principle, knowledge-based mean fields and protein folding, an approach to the computational determination of protein structures. J. Comp.-Aided Mol. Design 7:473–501, 1993.
39. Cohen, F. E., Ring, C. S., Presnell, S. R. Protein structure prediction: Recent successes and failures. Protein Eng. 6:121, 1993.