Prediction of Protein Folding from Amino Acid Sequence over Discrete Conformation Spaces†

Gordon M. Crippen
College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109
Received August 27, 1990; Revised Manuscript Received December 21, 1990

ABSTRACT: Predicting the three-dimensional structure of a protein given only its amino acid sequence is a long-standing goal in computational chemistry. In the thermodynamic approach, one needs a potential function of conformation that resembles the free energy of the real protein to the extent that the global minimum of the potential is attained by the native conformation and no other. In practice, this has never been achieved with certainty because even with greatly simplified representations of the polypeptide chain, there are an astronomical number of local minima to examine. If one chooses instead a protein representation with only a large but manageable number of discrete conformations, then the global preference of the potential for the native can be directly verified. Representing a protein as a walk on a two-dimensional square lattice makes it easy to see that simple functions of the interresidue contacts are sufficient to globally favor a given "native" conformation, as long as it is a compact, globular structure. Explicit representation of the solvent is not required. Another more realistic way to confine the conformational search to a finite set is to draw alternative conformations from fragments of larger proteins having known crystal structure. Then it is possible to construct a simple function of interresidue contacts in three dimensions such that only 8 proteins are required to determine the adjustable parameters, and the native conformations of 37 other proteins are correctly preferred over all alternative conformations. The deduced function favors short-range backbone–backbone contacts regardless of residue type and long-range hydrophobic associations. Interactions over long distances, such as electrostatics, are not required.

Many globular proteins at equilibrium are at the global minimum of free energy over the kinetically accessible region of conformation space. (Anfinsen, 1973; Go, 1983) In the thermodynamic approach to predicting protein folding from amino acid sequence, one attempts to simulate this by choosing some kind of potential function of conformation and then searching for the conformation(s) having the global minimal value of this function. There are two interrelated issues that need to be solved in order to ensure the calculations are feasible and to get correct answers: how should conformations be represented, and how should the function be constructed? The standard approach to computational conformational analysis, molecular mechanics, represents each atom in the molecule as a point in three-dimensional Cartesian space, so that atoms can move in a continuous fashion by smoothly changing their x, y, and z coordinates. Sometimes the atomic Cartesian coordinates are calculated from specified internal coordinates, such as bond lengths, bond angles, and dihedral angles, but in any case, conformational movements are smooth and continuous. Potential functions of conformation are generally chosen as long sums of two-, three-, and four-atom interactions, where each term is some reasonable continuous function of atomic coordinates. The adjustable parameters within these terms are subsequently varied so as to reproduce some selection of experimentally observed conformations, crystal structures, known bond rotation barriers, enthalpies of sublimation, vibrational frequencies, etc. The intent is that the function should simulate the enthalpy of a molecule at room temperature (or sometimes at 0 K), and entropic effects, such as solvation and conformational variability at room temperature, must be simulated by lengthy molecular dynamics calculations using this same function. Even verification of the function with respect to energetically determined phenomena is typically limited. In particular, agreement with experimentally observed conformations means that the function should have a substantial local minimum relatively close (<0.1 Å) to the experimental value. Apparently any reasonable potential having interatomic attractions but also repulsions (simulating van der Waals overlaps, covalent bond compression, and so on) must have a number of local minima that increases rapidly with the number of atoms. Thus, molecular mechanics potential functions are generally tuned to agree well with the experimentally observed conformations of a selection of molecules in the vicinity of those conformations, but in a more global sense, there may be other much deeper local minima corresponding to very different conformations. I am especially concerned that standard functions do not necessarily favor the crystal structure of a protein over an alternative folding (Bryant & Amzel 1987; Novotny et al., 1988). My co-workers and I have devised potential functions having better global properties (Crippen & Snow, 1990; Seetharamulu & Crippen, 1991) in the sense that the functions are intended to simulate the free energy of solvated polypeptides and that the global minimum should be near the crystal structure for a selection of small proteins. This is a qualitatively different parameter-adjustment task, made extremely difficult because the number of local minima apparently increases exponentially with the size of the molecule (Crippen, 1975). Since the molecule is represented in terms of interacting particles positioned by smoothly varying Cartesian coordinates and the potential is a continuous function of these coordinates, even finding a local minimum requires converging on it with a nonlinear minimization algorithm, much less verifying that the near-native minimum is indeed the very lowest of all local minima.

One way of making the problem more tractable is to convert

† This work was supported by grants from the National Institutes of Health (GM37123) and the National Science Foundation (DMB-8705006).

0006-2960/91/0430-4232$02.50/0

from a continuous conformation space to a finite discrete one, mathematically speaking. Instead of allowing atomic coordinates to vary continuously so that important conformations correspond to local minima of the potential over the continuous space of all conformations, it would be much more convenient if there was just a (possibly long but) finite list of allowed conformations a protein could take on. Then, if one of these is designated to be the native conformation, the potential is merely required to have a lower value for it than for any other conformation in the list. Verification of the requirement amounts to simply evaluating the function for each conformation in the finite discrete conformation space. There are two plausible ways to construct such a discrete space. The traditional way in theoretical polymer chemistry is to model a protein as a walk on some sort of lattice, so that for short chains *all* conformations can be examined, subject to the constraints of certain allowed interresidue bond lengths and bond angles. The second way is to use pieces of the known protein crystal structures as examples of how polypeptides can fold. The advantage is that the conformations are obviously much more realistic, but since there are only so many protein crystal structures, not all possible conformations would be explored. In this paper, I will use the lattice approach to establish some important general conclusions about potential functions and the uniqueness of the native conformation. Then I will turn to the second approach to use these conclusions for producing a potential that has considerable global predictive power.

Others have been using both kinds of discrete conformation spaces for some time, particularly the lattice idea. However, it is important to note the sometimes subtle differences between related studies and this one. For instance, Covell and Jernigan (1990) looked at all conformations available to a few proteins corresponding to Hamilton walks on certain lattices and showed that a function of interresidue contacts previously derived from a survey of protein crystal structures could rank the native conformation among the best 2%. Lau and Dill (1989) found that only certain kinds of amino acid sequences can produce a unique global minimum of a particular extremely simple contact potential for two-dimensional square-lattice walks. Sippl (1990) constructed a continuous function of continuously variable interresidue distances by surveying the kinds of conformations seen in the protein crystal data for short segments. Here we treat a related but fundamentally different question: Given the finite set of all possible conformations and given one of these as the "native" structure, can one devise a potential function that clearly favors the native over all the rest? If so, does this function have any predictive power? What restrictions are there on the types of such functions and the kinds of conformations that can be globally favored? The next section explains how I first examined two-dimensional square-lattice models of proteins, because it is easy to explore all conformations and because the functional forms available for potentials are rather limited. Then the following section shows how to generalize the lattice results to a much more realistic representation of proteins in terms of three-dimensional atomic coordinates taken from protein X-ray crystal structures.

Square-Lattice Model

Represent a protein having n residues as a self-avoiding walk of $n - 1$ steps on a very large two-dimensional square lattice. Let the lattice spacing be unity. Sequentially adjacent residues are distance 1 apart. Each occupied lattice point has a sequence number $1 \leq i \leq n$ and a type $1 \leq t_i \leq 20$, corresponding to the 20 kinds of naturally occurring amino acids.

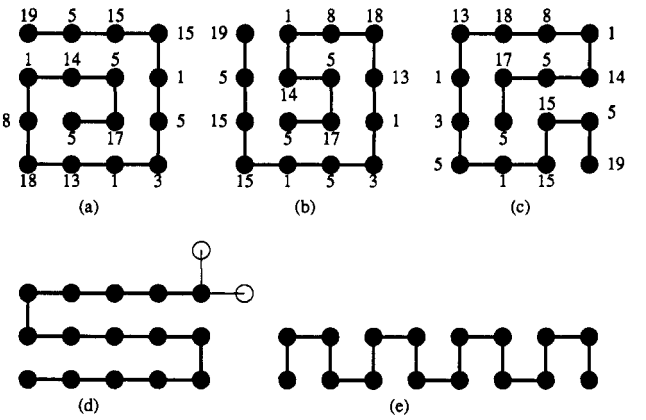

FIGURE 1: Some alternative conformations of a 16-residue walk on a square lattice. The arbitrarily chosen "native" conformation (a), showing the arbitrarily chosen sequence of residue types. The last residue in the sequence has type 19. Alternative rigid conformations are shown in b and c. Conformation d is not rigid, as indicated by the two alternative placements of its last residue. Conformation e is the lattice analogue of a helix.

Define a contact to be when adjacent lattice points are occupied by two residues that are not sequentially adjacent: $d_{ij} \leq d_c = 1$ and $|i - j| > 1$. Since the walks are self-avoiding, a contact between residues i and j implies $|i - j| \geq 3$.

Two conformations, c and c', of the same protein are defined as different if corresponding residues cannot be superimposed by translation, rotation, and mirror inversion. Here, "corresponding" means that the two residues have the same sequence number.

Let S_c be the set of contacts occurring in conformation c. Identify the contacts by sequence separation, type of residue with lower sequence number, and type of residue with higher sequence number. Since the same residue type may occur more than once in the protein, the list of contacts may have more than one entry with the same description. We want a potential function (mimicking the free energy of a real protein as a function of conformation) that depends only on the contacts: $E(c) = E(S_c)$. This assumes that only two-body interactions are important. In particular, assume the effects of contacts are simply additive:

$$E(c) = \sum_{k \in S_c} f(k) \tag{1}$$

where f may take into account the types of residues involved, their sequence separation, and even which residue is higher in sequence number.

For n residues on an $m \times m$ lattice where $m \gg n$, there may be n_{rr} adjacent occupied pairs of points (residue–residue contacts plus $n - 1$ sequentially adjacent residue pairs), n_{rs} residue–solvent contacts, and n_{ss} solvent–solvent contacts. Then

$$\begin{aligned} n_{rr} &= n - 1 + x \\ n_{rs} &= 2n + 2 - 2x \\ n_{ss} &= 2(m - 1)m - 3n - 1 + x \end{aligned} \tag{2}$$

where always $x \geq 0$, $x = 0$ for a fully extended conformation, and $x \gg 0$ for globular conformations. Because all three types of contacts can be expressed in terms of a single parameter, it is sufficient to take into account the residue–residue contacts for defining E, even though solvation of real proteins is an important factor in their free energy.

We define a conformation c to be *rigid* if there is no $c' \neq c$ such that $S_c = S_{c'}$. For $d_c = 1$, that implies that rigid conformations tend to be rather compact. In Figure 1, con-

Table I: Contact Potential Function That Favors Figure 1a[a]

separation	class	e_{ij}
3	{all}	(0)
4–∞	1 = {1 – 18, 20}	$\begin{pmatrix} -2 & -1 \\ -1 & 0 \end{pmatrix}$
	2 = {19}	

[a] Classification of contacts is by sequence separation and by subsets of numeric residue type; the third column gives the matrix of empirically determined interaction parameters for contacts between residues of the various classes.

formations a, b, c, and e are all rigid, but conformation d is not, because there are two ways to place its last residue and the list of contacts would be exactly the same for both. One way distinguish the two alternatives for conformation d would be that in one conformation, one solvent "molecule" could be in contact simultaneously with residue sequence numbers 15 and 16, while in the other conformation, another solvent molecule would be in contact with 6 and 16. However, we have assumed that the potential function depends only on a sum of two-body contacts. Another way to distinguish the two alternatives would be to raise $d_c > \sqrt{2}$. Similarly for $d_c = 1$, the lattice helix analogue e is rigid, but an extended chain is not. Raising the cutoff to $\sqrt{2} < d_c < 2$ allows the extended chain to be rigid. Since in real polypeptides, a long isolated α-helix can be stable while a single extended strand is not, we keep the cutoff $d_c = 1$. If we want the native conformation c_{nat} to be the *unique* global minimum of E, i.e., $E(c_{nat}) < E(c)$ for all $c \neq c_{nat}$, then the native must be a rigid conformation.

For a 16-residue chain on a square lattice, there are 802075 distinct conformations, not counting translations, rotations, and mirror inversions. Is it possible to build a potential E of the form in eq 1 such that the arbitrarily selected "native" conformation (Figure 1a) with its arbitrarily chosen amino acid sequence can be the unique global minimum of E? To eliminate any ambiguities about uniqueness, let f in eq 1 be an integer-valued function of the contacts. More specifically, f first groups contacts into a number of ranges of sequence separations, where the first range includes separations of 3 and perhaps more, and the last range covers all separations greater than the upper limit of the previous range. Then for each range, residue types are grouped into a number of mutually exclusive and comprehensive classes. Finally, each range has a table of integers e, so that $f(k) = e_{ij}$, where the contact k involved a residue of type class i in contact with a sequentially higher residue of type class j. The simple-minded way to determine f is to carry out a branch-and-bound depth-first search over the number of ranges, the minimal sequence separation of each range, the number of classes in each range, the assignment of residue types to the classes in each range, whether the interaction table should be symmetric, and finally what integer values should be put into the tables, starting with small absolute values. There are two checks along the way that keep the combinatorial explosion manageable. The first is that a given choice of ranges and their sequence separations may be eliminated if it fails to distinguish between the native and some alternative set of contacts even when residue types are fully separated into 20 classes in each range. The second check is that the chosen distribution of residues into classes in each range must distinguish between the native and each alternative. For example, exactly the same nine contacts occur in Figure 1 (conformations a and b) if there is only one range, i.e., the potential does not depend on sequence separations. The first solution found by the branch-and-bound search happens to be Table I, which might be paraphrased as "don't count hairpin turn contacts, but otherwise make as many contacts as possible, although contacts involving residue type 19 are not as favorable as others". Although this potential function is sufficient to make conformation a the global minimum over all 802074 alternatives, it can be viewed as having been determined by looking at only three of them! Alternative 1 is just conformation a with the sequence reversed; 2 is conformation b; and 3 is conformation c. Then in the search, 1 and 2 are active in setting up the classification of ranges and residue type classes, and 1 and 3 are active in determining the interaction tables. Of course the potential can be viewed as being determined by other sets of alternatives (just as there are many choices of basis in linear programming), but the important thing is that at least one set of critical alternatives is so small.

Outside of the rigidity of conformer a, there is nothing very special about it, and I have been able to produce potential functions that uniquely favor various other conformations with other sequences, or indeed, that simultaneously favor the respective native conformation of more than one "protein". The helical conformation e, for example, is uniquely favored by

$$\text{separation 3, class 1 = \{all\}, } e = (-1)$$
$$\text{separation 4–∞, class 1 = \{all\}, } e = (0) \qquad (3)$$

There is a simple procedure to automatically generate a small set of conformations that includes those critical for determining the potential function. Let every residue in the given protein be a node in a graph, and connect two nodes whenever they correspond to sequentially adjacent (i.e., bonded) residues or to residues in contact. Now any Hamilton walk on this graph corresponds to a reassignment of residue sequence numbers and hence residue types to the nodes, and all edges not traversed in the walk are contacts. Because contacts are distinguished only on the basis of sequence separation and the two residue types, two different walks may produce the same set of contacts. Thus, conformation a gives rise to 552 different walks on a 4 × 4 lattice, of which there are only 69 different sets of contacts, and these drive the combinatorial search to the same potential function as before. If conformation e is viewed as a homopolymer, there are 116 walks, yielding only 16 distinct sets of contacts, resulting in the same potential as before.

Protein Crystal Structures

Of course, the goal is to construct a potential function that can be used to predict the conformations of real proteins composed of atoms having three-dimensional, continuous, Cartesian atomic coordinates. The foregoing model studies suggest that some kind of residue–residue contact function would be adequate, even if solvent is not explicitly represented and the contact terms operated over only short distances. The key problems then are to define interresidue contacts for real protein structures and to contrast the native conformation with a large but manageable finite set of very plausible alternative conformations.

Consider a set of 57 high-resolution (≤2.5 Å) small to moderate sized (≤250 residues) protein crystal structures selected from the Brookhaven Protein Data Bank (Bernstein et al., 1977) to avoid obvious duplication, large ligands, and substantial unresolved regions: 1mlt, 1ppt, 1crn, 3rxn, 1fdx, 2ovo, 4pti, 2mt2, 2ebx, 1sn3, 2abx, 2icb, 2pka, 351c, 1cc5, 1hip, 2b5c, 2gn5, 3fxc, 2pcy, 4cyt, 2fd1, 2cdv, 1rei, 3cpv, 1ccr, 3c2c, 1hmq, 2rhe, 1cy3, 155c, 1pp2, 1bp2 1rn3, 2ccy, 1aza, 1lz1, 1ecd, 4fxn, 2mhb, 2hhb, 2sns, 1fxl, 2lhb, 2sod, 1lh1, 3mbn, 4dfr, 1lzm, 3wga, 1gcr, 2stv, 3fab, 1ppd, 2act, 2cna, and 1tim (listed in order of increasing number of residues). Of each data set, only the first polypeptide chain was used, and it was read only up to the point of the first break in the

chain. In the case of 2pka, that meant only the first 77 residues were included, because there follows an uninterpretable portion of the electron-density map. Since that part of the chain comprises only some of a domain, 2pka was used for generating alternative conformations of small proteins, but when it comes to prediction, it is inappropriate to demand that a potential function should fold these residues correctly in the absence of the rest of the domain.

In order to devise a contact potential function to favor these crystal structures over all alternative conformations, we must define a contact for three-dimensional Cartesian coordinates of atoms. Consider only the backbone N and O atoms and side-chain C^β atom of each residue as forming contacts; construct the approximate position of an artificial C^β for each Gly residue. A backbone–backbone carbonyl-to-amide contact must have the O–N distance <3.2 Å and the C–N distance >3.9 Å. A backbone–side-chain contact is counted if the distance between N or O and the C^β is less than 5.0 Å; for side-chain–side-chain contacts, the cutoff is 9.0 Å. In addition, for side-chain–side-chain or backbone–side-chain contacts, there must be no other atom between the interacting pair closer than 1.4 Å to the line segment joining them. The idea is that much the same side-chain contacts could be made after point mutations by permitting small shifts in the conformation.

Using the above definition of a contact, one can convert a set of atomic coordinates into a list of contacts, where we enumerate only three items per contact: the sequence separation of the atoms in contact, and the types of the two amino acid residues involved. Only sequence separations of three or more are counted, and if a backbone atom is involved, its residue type is noted as Gly. The contact definition and enumeration scheme are crucial to the success of what follows. Looser definitions of what constitutes a contact fail to discriminate between the native and alternative conformations.

Instead of the use of walks on a lattice to generate a finite set of alternative conformations, the 57 crystal structures themselves can be used. For example, crambin (1crn) has 46 residues, while rubredoxin (3rxn) has 52. Use of the amino acid sequence of 1crn applied to the coordinates of 3rxn residues 1–46, 2–47, ..., 7–52 produces seven plausible alternative conformations of crambin. In this way we can produce 5330 alternative conformations for 1mlt from the coordinates of proteins 1ppt through 1tim, 4769 for 1ppt using 1crn through 1tim, and so on, down to 12 for 2cna. Of course, this scheme produces no alternative conformations for the largest protein, 1tim, and less than 100 each for the next three largest, but the average number of conformations per protein was 1621.

Given some classification scheme, as in Table I, there is a linear polynomial $E(c)$ in the e_{ij}'s corresponding to the set of contacts for each conformation c of a protein. Then every alternative conformation of every protein produces a homogeneous inequality $E(\text{native}) < E(\text{alternative})$, and the set of inequalities can be solved numerically (Jurs, 1986) for the set of e_{ij}'s. The initial classification used in these studies, shown in Table II, is based on conventional wisdom about grouping together helix-formers vs helix-breakers for short-range interactions, while residue types are grouped according to general hydrophobicity for the medium- and long-range interactions. It is easy to set up the corresponding 90 000 inequalities for the 55 proteins (the first 56 except for 2pka, as explained earlier) and solve them for the $4 \times 7 \times 6/2 = 84$ adjustable e_{ij}'s. If such a set of linear inequalities has any feasible solution at all, it is generally not unique but rather a convex region of parameter space. We can interpret a given solution in chemical terms as saying there are certain kinds of interactions that are particularly rare or particularly common in the native conformations of each protein compared to the respective alternative nonnative conformations. Other solutions within the convex feasible region may have substantially different interpretations as to what kinds of interactions are important. When residue types are grouped into many small classes, there will be many different kinds of interactions, each with its adjustable e_{ij} parameter. Then the solution of the set of inequalities tends to depend on some trivial combination of preponderances or lacks of some of these interactions in the native structures at hand, whereas the next protein may not fall into that particular pattern. Indeed, this is exactly what happened: the detailed residue classification of Table II leads easily to solutions that vary considerably, depending on which proteins were used to generate the set of inequalities, and the resulting potential functions seldom showed a preference for the native set of contacts over those of alternative conformations for any protein outside the training set.

Table II: The Standard Starting Classification Used To Eventually Find Broader Classifications with More Predictive Power[a]

separation	class
3	1 = {G}, 2 = {ALICMF}, 3 = {VHS}, 4 = {P}, 5 = {RDEQ}, 6 = {TKN}, 7 = {YW}
4	1 = {G}, 2 = {ALICMF}, 3 = {VHS}, 4 = {P}, 5 = {RDEQ}, 6 = {TKN}, 7 = {YW}
5–7	1 = {G}, 2 = {AV}, 3 = {LICMF}, 4 = {YHWST}, 5 = {KR}, 6 = {P}, 7 = {DNEQ}
8–∞	1 = {G}, 2 = {AV}, 3 = {LICMF}, 4 = {YHWST}, 5 = {KR}, 6 = {P}, 7 = {DNEQ}

[a] Amino acid residue contacts are grouped by sequence separation and by subsets of types of residue indicated by the single-letter residue code.

On the other hand, if residues are grouped into a few large classes, there will be a small number of kinds of contacts and a correspondingly small number of adjustable parameters. Findings a feasible solution tends to be more difficult, but if one exists, it tends to depend on more general trends in the kinds of contacts found in native proteins, and therefore the likelihood of successful predictions is greater. A very general way to find the potential with the greatest predictive power would be to start with one sequence separation range (3–∞) and all 20 residue types contained in a single class. Then systematically try all combinations of classification schemes in order of increasing detail until one yields a feasible solution to the corresponding inequalities. This was computationally tractable for the small square-lattice problems of the previous section but not for the more elaborate classifications required for these proteins. Instead I approached the problem in the opposite order: start with the very detailed classification in Table II, which leads easily to a solution of the corresponding inequalities; then, starting with the first separation range, attempt to redistribute the residues of one class into some combination of the other classes of that range, and check that a feasible solution can still be obtained. Eventually no further residue class in any sequence range can be eliminated. The process can consume literally weeks of CPU time on an Iris 4D/220, since there are many ways to reassign the classes of all the residues contained in the class to be eliminated, and sometimes thousands of ways had to be tried before either finding a feasible combination or eliminating all possibilities and going on to the next class for elimination. The result is Table III. Lumping all types of residues into a single class for the purposes of short- and medium-range interactions is partly an artifact of the classification simplification procedure, which starts with the short-range separations. In part, however, this seems to be a genuine trend inherent in the data.

Table III: Contact Potential Function That Favors 45 Proteins' Crystal Structures[a]

separation	class	e_{ij}			
3	{all}	(−0.008)			
4	{all}	(0.004)			
5–7	{all}	(0.021)			
8–∞	1 = {GYHSRNE}	−0.012	−0.074	−0.054	0.123
	2 = {AV}	−0.074	0.123	−0.317	0.156
	3 = {LICMF}	−0.054	−0.317	−0.263	−0.010
	4 = {PWTKDQ}	0.123	0.156	−0.010	−0.004

[a] Classification of amino acid residue contacts is by sequence separation and by subsets of types of residue indicated by the single-letter residue code. The third column gives the matrix of empirically determined interaction parameters for contacts between residues of the various classes.

For example, I have obtained similar preliminary results using an alternative simplification procedure that tries to reduce the information theory entropy of the total classification in all separation ranges at once. Otherwise, Table III makes some degree of sense: helix formation is slightly favorable, whereas medium-range contacts are discouraged; the main factor is to form hydrophobic–hydrophobic long-range contacts while secondarily avoiding hydrophilic–hydrophilic contacts. It is probably an artifact of the fitting procedure that interactions between large hydrophobic residues and Ala or Val are singled out as being particularly important. Given that we see here just one point in a large feasible space and that the classification scheme and interaction parameters of Table III are quite dependent on which proteins were included in the study and on the proposed starting classification, it would be unwise to comment further on the physical significance of the result. Unlike least-squares fits, the solution found for a set of inequalities can be substantially shifted by adding one new protein or by suggesting a different starting classification.

At the solution shown in Table III, only eight proteins had one or more inequalities that were active: 4pti, 2ebx, 1cc5, 2gn5, 4cyt, 2cdv, 1hmq, and 2nlhb. Thus, a few of the alternative conformations available to only these eight proteins are what determine the adjustable interaction parameters e_{ij} in the table. Of the 56 proteins for which alternative conformations had been calculated at all, another 37 favored the native over all alternatives by substantial margins (>0.2). These correctly predicted proteins ranged in size from 1fdx with 54 residues to 2cna with 237 residues. That leaves 11 proteins where at least one alternative conformation had a function value lower than that of the native: 1mlt, 1ppt, 1crn, 3rxn, 2ovo, 2pka, 1hip, 2fd1, 3c2c, 1ecd, and 4fxn. The extremely small peptides melittin (1mlt) and avian pancreatic polypeptide (1ppt) seem to be difficult to bring into any general scheme, possibly because they are treated as isolated monomers in the calculations, whereas crystal packing and the tight tetramer of melittin may significantly affect their observed conformations. Porcine kallikrein (2pka) is not represented in these calculations as even a whole domain for technical reasons explained above, so failure to fit it is actually appropriate. Crambin (1crn) is an anomalously hydrophobic small protein crystallized in dilute ethanol, so failure here is perhaps excusable. The remaining seven mispredictions are not easy to explain away. They are in error because the best alternative conformation had a better contact function value than that of the native, but the margins of error are not significantly different from the margins of success for the 37 correctly predicted proteins (3rxn 0.7, 2ovo 0.3, 1hip 1.7, 2fd1 1.5, 3c2c 0.9, 1ecd 1.9, and 4fxn 14.1). Indeed, if the seven mispredicted proteins are included in the training set, another residue classification and interaction parameter set can be calculated, having similar numbers of residue classes and similar predictive powers. The main change is that a different 11 or 12 proteins are mispredicted. By including more proteins in the training set, it is certainly possible to fit all 52 proteins (leaving out only 1mlt, 1ppt, 1crn, and 2pka), but I have so far been unable to find a classification scheme that is as simple as that of Table III. Obviously this isn't the last word on the subject, and every attempt will be made to correctly predict the native conformations of more proteins.

Future use of this sort of contact potential would be for predicting the conformation of a protein having known sequence, but of course no known crystal structure in the Protein Data Bank. As long as there is some crystal structure of a larger protein where a contiguous chain segment of the right length approximately adopts the native structure of the novel protein, it is very straightforward to find this segment, and the search would require only an hour or two of computer time. Note that one needs no assumptions about sequence homology. Any completely unrelated protein of known three-dimensional structure is an eligible source of alternative conformations, and all you have to do is choose the conformation or conformations with lowest potential. Given that there seem to be a limited number of structural motifs, having a correct one in the Protein Data Bank is not unlikely. On the other hand, suppose the novel protein is so fortunate as to have a crystal structure for a closely homologous protein. It is still likely that there will be no completely contiguous portion of the known that matches the native of the novel protein, simply because of the usual insertions and deletions one sees on the surfaces of globular proteins. This means that a general prediction program built around one of these contact potentials would have to set up a correspondence between the novel sequence and the known structure that was not necessarily just a displacement of contiguous segments but must also allow for insertions and deletions. This problem is being actively explored.

Conclusions

As long as two-body interactions are sufficient to approximate the free energy of a protein, then a potential designed as a function of only residue–residue contacts can adequately include solvation effects implicitly. It is no coincidence that the proteins that appear to have unique native structures are globular. If the major contributions to a protein's free energy as a function of conformation act effectively over only short distances through space, then a protein's native structure must be globular. Indeed a carefully chosen short-distance definition of contacts in protein crystal structures can be used to devise a simple function that correctly prefers the native conformation of several proteins over thousands of alternative structures and can correctly predict the native to be optimal for many more proteins. Key to this success are (1) a definition of a residue–residue contact that is relatively unaffected by changing amino acid side chains and (2) use of a finite discrete space of alternative conformations derived from other protein crystal structures. Refinements could eventually produce a potential capable of preferring the X-ray crystal structure of most proteins over alternative conformations. Then when presented with a novel amino acid sequence having little or no resemblance to those found in the Protein Data Bank, current methodology would choose the contiguous segment out of all the known protein structures that the contact function prefers. If there is a known protein having a domain of the correct conformation (structural homology, but not necessarily any sequence homology), then the potential function would make this its prediction and be correct in doing so. If the only structural match in the database involved substantial insertions

and deletions, then the current methods would have to be improved in order to make a correct prediction.

ACKNOWLEDGMENTS

I thank Andrew Smellie for helpful discussions in developing these ideas.

REFERENCES

Anfinsen, C. B. (1973) *Science 181*, 223-230.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535-542.
Bryant, S. H., & Amzel, L. M. (1987) *Int. J. Peptide Protein Res. 29*, 46-52.
Covell, D. G., & Jernigan, R. L. (1990) *Biochemistry 29*, 3287-3294.
Crippen, G. M. (1975) *J. Comput. Phys. 18*, 224-231.
Crippen, G. M., & Snow, M. E. (1990) *Biopolymers 29*, 1479-1489.
Go, N. (1983) *Annu. Rev. Biophys. Bioeng. 12*, 183-210.
Jurs, P. C. (1986) *Computer Software Applications in Chemistry*, pp 198-199, John Wiley and Sons, New York.
Lau, K. F. & Dill, K. A. (1989) *Macromolecules 22*, 3986-3997.
Novotny, J., Rashin, A. A., & Bruccoleri, R. E. (1988) *Proteins: Struct., Funct., Genet. 4*, 19-30.
Seetharamulu, P., & Crippen, G. M. (1991) *J. Math. Chem. 6*, 91-110.
Sippl, M. J. (1990) *J. Mol. Biol. 213*, 859-883.

Contribution to the Thermodynamics of Protein Folding from the Reduction in Water-Accessible Nonpolar Surface Area†

Jeff R. Livingstone,‡ Ruth S. Spolar,§ and M. Thomas Record, Jr.*,‡,§
Departments of Biochemistry and Chemistry, University of Wisconsin—Madison, Madison, Wisconsin 53706
Received October 17, 1990; Revised Manuscript Received January 2, 1991

ABSTRACT: Protein folding and the transfer of hydrocarbons from a dilute aqueous solution to the pure liquid phase are thermodynamically similar in that both processes remove nonpolar surface from water and both are accompanied by anomalously large negative heat capacity changes. On the basis of a limited set of published surface areas, we previously proposed that heat capacity changes (ΔC_p°) for the transfer of hydrocarbons from water to the pure liquid phase and for the folding of globular proteins exhibit the same proportionality to the reduction in water-accessible nonpolar surface area (ΔA_{np}) [Spolar, R. S., Ha, J. H., & Record, M. T., Jr. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 8382-8385]. The consequence of this proposal is that the experimental ΔC_p° for protein folding can be used to obtain estimates of ΔA_{np} and of the contribution to the stability of the folded state from removal of a nonpolar surface from water. In this paper, a rigorous molecular surface area algorithm [Richmond, T. J. (1984) *J. Mol. Biol. 178*, 63-89] is applied to obtain self-consistent values of the water-accessible nonpolar surface areas of the native and completely denatured states of the entire set of globular proteins for which both crystal structures and ΔC_p° of folding have been determined and for the set of liquid and liquefiable hydrocarbons for which ΔC_p° of transfer are known. Both processes (hydrocarbon transfer and protein folding) exhibit the same direct proportionality between ΔC_p° and ΔA_{np}. We conclude that the large negative heat capacity changes observed in protein folding and other self-assembly processes involving proteins provide a quantitative measure of the reduction in the water-accessible nonpolar surface area and of the contribution of the hydrophobic effect to the stability of the native state and to protein assembly.

Noncovalent assembly processes involving proteins, such as folding, oligomerization, and ligand binding, are typically accompanied by large reductions in water-accessible nonpolar surface area. Kauzmann (1959) proposed that the removal of nonpolar amino acid side chains from water (the "hydrophobic effect") should provide a large driving force (ΔG_{hyd}°) for assembly or association processes involving proteins. To quantify the contribution of ΔG_{hyd}° to the observed standard free-energy change (ΔG_{obs}°) for the assembly or association process, most work has focused on analyzing the free energy of transfer (ΔG_{tr}°) of amino acids or their analogues from water to an organic solvent (Cohn & Edsall, 1943; Nozaki & Tanford, 1971; Fauchère & Pliska, 1983) or to the gas phase (Wolfenden et al., 1981). Various hydrophobicity scales have been proposed that rank amino acids according to either their experimental transfer behavior [cf. Cornette et al. (1987)] or their observed distribution in protein crystal structures between the exterior and interior of the folded form (Janin, 1979; Rose et al., 1985a; Miller et al., 1987a). However, a comparison of hydrophobicity scales reveals that, in general, values of ΔG_{tr}° from different scales do not correlate well with each other and that even the relative ranking of amino acids varies from scale to scale (Rose et al., 1985b). Alternatively, the relationship between ΔG_{tr}° of amino acids and the *total* water-accessible surface has been examined (Chothia, 1974) as well as the relationship between ΔG_{tr}° and surface area at the functional group level (Eisenberg & MacLachlan, 1986; Ooi et al., 1987). The use of these relationships to estimate values of ΔG_{hyd}° is complicated by the same problem

† This work was supported in part by NIH Grant GM23467.
* Correspondence should be addressed to this author.
‡ Department of Biochemistry.
§ Department of Chemistry.

0006-2960/91/0430-4237$02.50/0