PROTEINS: Structure, Function, and Genetics 26:167–171 (1996)

Failures of Inverse Folding and Threading With Gapped Alignment

Gordon M. Crippen
College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109

ABSTRACT **To calculate the tertiary structure of a protein from its amino acid sequence, the thermodynamic approach requires a potential function of sequence and conformation that has its global minimum at the native conformation for many different proteins. Here we study the behavior of such functions for the simplest model system that still has some of the features of the protein folding problem, namely two-dimensional square lattice chain configurations involving two residue types. First we show that even the given contact potential, which by definition is used to identify the folding sequences and their unique native conformations, cannot always correctly select which sequences will fold to a given structure. Second, we demonstrate that the given contact potential is not always able to favor the native alignment of a native sequence on its own native conformation over other gapped alignments of different folding sequences onto that same conformation. Because of these shortcomings, even in this simple model system in which all conformations and all native sequences are known and determined directly by the given potential, we must reexamine our expectations for empirical potentials used for inverse folding and gapped alignment on more realistic representations of proteins.**

Key words: protein folding, Lattice models, Contact potentials, Protein structure prediction

INTRODUCTION

Now that there are many globular proteins having experimentally determined three-dimensional structures, and orders of magnitude more known sequences of biologically important proteins, there is great interest in using this information to correlate sequence and structure. One such problem is the inverse folding task: given some particular polypeptide conformation, calculate what amino acid sequence, if any, will adopt it as its native structure. A somewhat easier version of this is the sequence identification problem: given a protein of known sequence and structure, select out of a data bank of protein sequences (all of which presumably fold uniquely to some structure or another) those that fold up to the given structure. The simple form of sequence identification assumes that all the sequences have the same length as the given structure. However, it is well known that sequences with high sequence identity (after some sort of optimal alignment allowing insertion and deletions) will fold to very similar three-dimensional structures where the aligned, evolutionarily conserved residues are located mostly in the interior, and the insertions and deletions correspond to exterior loop regions.[1–3] Therefore, the more useful goal is to solve the sequence identification problem with gapped alignment, since this would at least give a general idea of much of the structure for many more sequences, given the limited set of known structures with which we have to work.

There are several methods for sequence identification that rely on empirical potentials and various search strategies, resulting in various levels of success and accuracy.[4–13] The lingering doubt is that perhaps they would perform yet better if we had better scoring functions. Conventional wisdom reasons that if the empirical potentials better approximated some features of the true energetics, results would surely improve. Here we test this idea by studying a simplified model for protein folding where we know by construction exactly what the "true" energy function is.

A favorite model for studying the statistical mechanics of heteropolymers is the self-avoiding square lattice walk. Each of the n_{res} residues is a point on the lattice and is permitted to have as a type either A or B. Let the potential function be a sum over interresidue close contacts

$$E(c,s) = \sum_{i<j-2}^{n_{res}} \begin{cases} e_{AA} & \text{for } r_{ij} = 1, \text{ types A, A} \\ e_{AB} & \text{for } r_{ij} = 1, \text{ types A, B} \\ e_{BB} & \text{for } r_{ij} = 1, \text{ types B, B} \\ 0 & \text{for } r_{ij} > 1 \end{cases} \quad (1)$$

which depends on the sequence, s, and the conformation, c, where r_{ij} is the distance in lattice units between residues i and j. Even with such a simple model, one can still observe[14] native and denatured states, folding pathways and bottlenecks, globular

Received March 19, 1996; accepted April 22, 1996.
Address reprint requests to Gordon M. Crippen, College of Pharmacy, University of Michigan, Ann Arbor, MI 48109.

structures, interior vs. exterior residues, secondary structure, entropy, free energy, and multiple local minima for E. For a given arbitrary choice of values for the contact energies, e_{AA}, e_{AB}, and e_{BB}, there are relatively few sequences, s, such that there is a unique conformation, $c(s)$, that is the global minimum of E. These correspond to real polypeptide sequences that fold to well-defined conformations, and we will therefore refer to them as "native sequences."[15]

$$E(c(s),s) < E(c,s) \; \forall c \neq c(s) \tag{2}$$

A consequence of this definition for native sequences is that the given potential always satisfies Equation 2. In other words, it solves the three-dimensional structure identification problem (3DID): given the native sequence and a large set of conformations all having that same chain length, identify which is the native conformation. The question before us now is whether we can use the given potential to perform sequence identification reliably with or without gapped alignment. One might naively reason that since the potential reflects the compatibility between sequence and structure, sequence identification and 3DID are equivalent problems that ought to be solved using the same, given potential.[16] It has been argued they are nonequivalent,[17,18] and here we reinforce that conclusion by simple counterexamples.

METHODS

Given n_{res}, the full set of lattice conformations was generated by a straightforward recursive program, requiring only that the structures be self-avoiding and there be no duplication due to translation, rotation, or mirror inversion. There was no restriction on the diameter of any structure.

Similarly, all possible sequences were exhaustively enumerated. Throughout this study, the sequences involved the same n_{res} residues as the conformations, so there were no structure/sequence comparisons having large insertions or deletions.

Determining the set of native sequences amounted simply to evaluating the potential for every conformer, given a particular sequence, and noting whether there was exactly one structure having the global minimum of energy. If so, the sequence is a native one; otherwise it was discarded. Which sequences are natives therefore depends on the chain length and on the potential function. Only native sequences were used in the gapped and ungapped alignments, since in a "real" application, the sequences would come from a data base of experimentally determined, nonrepeating protein sequences that can all be assumed to fold to some sort of reasonably well-defined globular structure.

Ungapped comparisons were nothing more than using all the different native sequences on all possible conformations, where both have the same chain length. In other words, residue i of the sequence is assigned to position i in the conformation for $i = 1, \ldots, n_{res}$. For any two-dimensional lattice conformation, each position is designated *loop* if it has no contacts according to Equation 1, or *core* otherwise. Gapped alignments assign to each residue in the given sequence either an undetermined loop position or one of the core positions of the conformation. The assignment preserves sequence/position ordering, assigns exactly one sequence residue to each core position, and two sequentially adjacent core positions are always matched with two successive residues in the sequence. If the minimal lattice distance between the end of one core segment and the beginning of the next is k steps, then the alignment must assign at least $k - 1$ intervening residues of the sequence to loop status. All these alignment requirements are fairly standard for more realistic models of proteins,[11] but here we need to introduce one more requirement peculiar to the square lattice. If the lattice distance between two successive core segments in the conformation is k steps, then an insertion must involve $k - 1 + 2n$ residues of the sequence assigned to loop status, where $n > 0$. If the insertion involves an odd number of extra residues, then there is no lattice walk that can bridge the gap, even relaxing the self-avoiding requirement. The search for gapped alignments consisted of nothing more than an exhaustive enumeration of all alignments obeying the above restrictions.

RESULTS

Ungapped Alignment

For some given choice of contact energies in Equation 1 and for some chain length, n_{res}, most energy levels will by highly degenerate for almost all of the $2^{n_{res}}$ possible sequences, and for some choices of interaction energies, there are none at all, e.g., $e_{AA} = 1$, $e_{AB} = 1$, and $e_{BB} = 1$. We have tried various nontrivial choices and find that the phenomena listed below are not restricted to some special set of interaction energies. Unless otherwise stated, the choice is $e_{AA} = 1$, $e_{AB} = 0$, and $e_{BB} = -1$, corresponding qualitatively to A residues being hydrophilic and B residues hydrophobic, favoring a hydrophobic core in the native conformation.

For $n_{res} = 4$, 6, 7, and 8, E does work for the sequence identification problem in that $E(c(s_i),s_i) < E(c(s_i),s_j)$ for all $s_i \neq s_j$ and $c(s_i) \neq c(s_j)$. There are also many cases where different native sequences have the same native conformation and their E values are the same, much like homologous sequences folding to nearly the same protein structure with little change in melting temperature. For $n_{res} = 9$, 10, and 11 there are increasing numbers of cases where two different native sequences have different native structures, yet $E(c(s_i),s_i) = E(c(s_i),s_j)$, so that

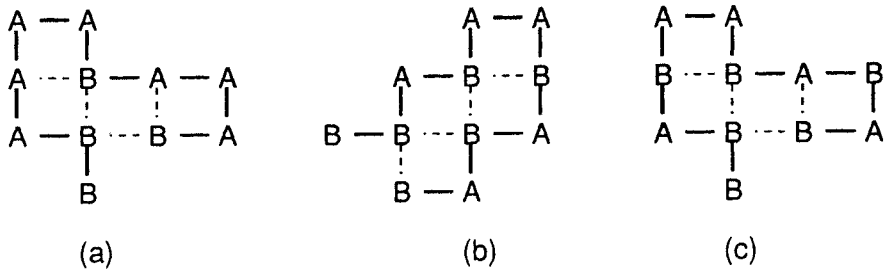

Fig. 1. An example where even E fails to correctly identify that sequence s_1 folds to conformation $c(s_1)$. Instead, it finds that s_2 is more compatible with $c(s_1)$, even though it really folds to $c(s_2)$. Dotted lines indicate the contacts that contribute to the E values. **a:** $E(c(s_1),s_1) = -2$. **b:** $E(c(s_2),s_2) = -4$. **c:** $E(c(s_1),s_2) = -3$.

even E is unable unambiguously to select only those folding sequences that are compatible with the given structure. Other sequence identification profiles and potentials employing more realistic protein models have similar difficulties,[11] but in this model system we are dealing not merely with some empirical potential, but the true, underlying one.

For $n_{res} = 11$, there are 5,513 conformations. Our hydrophobic/hydrophilic default potential function then gives rise to 44 native sequences, several of which share common native conformations, so that there are only 17 distinct $c(s)$. We find four cases where two sequences have the same native structure, but $E(c(s_i),s_i) > E(c(s_i),s_j)$, so that the true potential does not even give the compatibility of a sequence with its native structure top ranking. Figure 1, however, shows the worst outcome of all, where E indicates that s_2 is more compatible with $c(s_1)$ than s_1 is, even though s_1 actually folds to $c(s_1)$, and s_2 folds to $c(s_2)$. The message is clear: even though the true potential is by definition perfect at 3DID, it can fail at the sequence identification problem, even when no insertions or deletions are allowed in the alignment.

Gapped Alignment

The situation becomes even worse once insertions and deletions are permitted. For $n_{res} = 11$ and the same contact energies as before, there are 2 out of the 44 native sequences where $c(s_i) \neq c(s_j)$ but $E(c(s_i),a(s_j)) = E(c(s_i),s_i)$. In other words, the native alignment of the native sequence on its own native structure scores no better than some alignment of a differently folding sequence on that structure. In a similar vein, there are many cases where there is some nonnative alignment of a sequence on its own structure that scores as well as the native alignment. Another type of error is found for five sequences s_i where $c(s_i) = c(s_j)$ but $E(c(s_i),a(s_j)) < E(c(s_i),s_i)$. This means that both sequences actually fold to the same native structure, although a nonnative alignment of one of them scores better than the native alignment for the other. Finally, we find seven sequences where $c(s_i) \neq c(s_j)$ and $E(c(s_i),a(s_j)) < E(c(s_i),s_i)$. One of these examples is shown in Figure 2.

The performance of the potential at gapped sequence identification varies with the choice of contact energies. As long as BB contacts are energetically favorable while other contacts are either neutral or unfavorable, the native sequences average about 50% A residues and 50% B, with most of the loop residues having type A, and most of the core having type B. However, the residue type segregation between loop and core is never complete. For $n_{res} = 11$, $e_{AA} = 1$, $e_{AB} = 2$, and $e_{BB} = -1$, there are numerous examples of ambiguities where the native alignment of the native sequence onto its native conformation scores no better than various alignments of other sequences onto that conformation. However, there are no errors of any worse type. On the other hand, for $n_{res} = 11$, $e_{AA} = 0$, $e_{AB} = 0$, and $e_{BB} = -1$, there are 31 native sequences in all, of which 3 exhibit the worst kind of error, where a differently folding sequence scores distinctly better than the native on its native conformation.

DISCUSSION

These lattice studies are convenient because for short chains, one can exhaustively enumerate all conformations, all native sequences (given a particular set of contact energies), and all alignments of native sequences on native conformations. The search problem is settled, and we know by construction the true interresidue potential function. Then we have shown by counterexample that even the given potential fails in many cases to solve the sequence identification problem. In many more cases, the native sequence best matches its own conformation, but that score is no better than for numerous other sequences aligned onto the same conformation. Generally, the error rate for gapped alignment is worse than for ungapped.

Figure 3 schematically illustrates the situation we have, supposing we can represent the multidimensional conformation space as the vertical axis

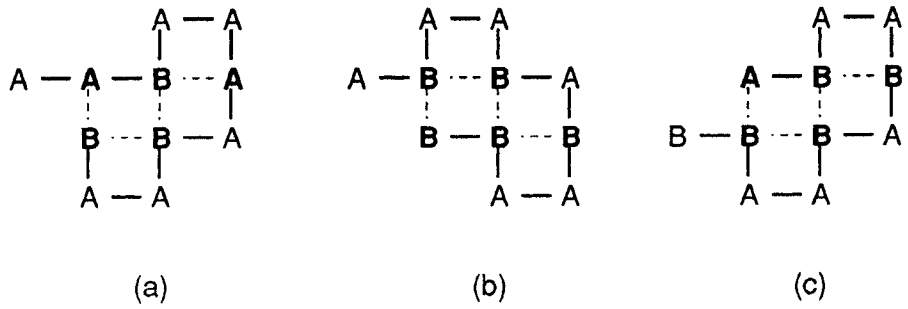

Fig. 2. A failure of E in sequence identification with gapped alignment. Dotted lines indicate the contacts, and boldface letters occupy the core positions in each conformation. **a:** $E(c(s_1),s_1) = -2$. **b:** $E(c(s_2),s_2) = -4$. **c:** $E(c(s_1),a(s_2)) = -3$, where the alignment in this case is just a shift of one position with no internal insertions or deletions.

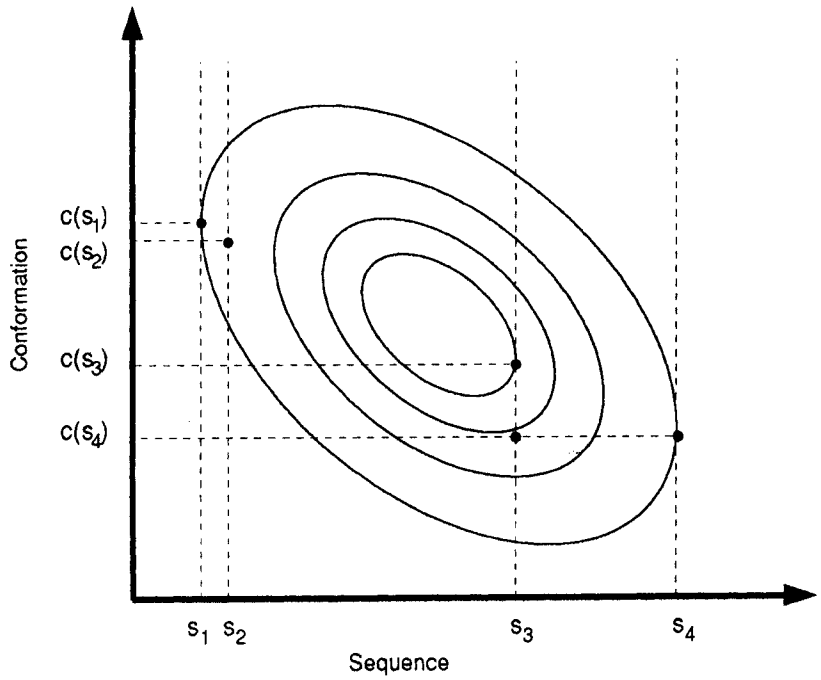

Fig. 3. Schematic representation of a very simple potential function, $E(c,s)$, that has only a single minimum, as indicated by the elliptical contour lines.

and the multidimensional sequence space as the horizontal one. Even if $E(c,s)$ is so simple as to have a single minimum, as indicated by the contour lines, the picture represents the behavior of inverse folding. For any sequence, s, the vertical slice through the E surface has its global minimum with respect to conformation at the corresponding native structure, $c(s)$, as in 3DID. Just as with real proteins, similar sequences, such as s_1 and s_2, have similar native conformations, $c(s_1)$ and $c(s_2)$. Going in the sequence identification direction, E correctly ranks s_3 as a better match to $c(s_3)$ than s_4 is. However, s_3 scores better with $c(s_4)$ than s_4 does, even though s_3 really folds up to $c(s_3)$.

But what does this mean for more realistic protein models? Even when the search over alignments is performed perfectly,[11] and the scoring function captures all physically important factors in the free energy function of a real protein, sequence identification can be expected to be ambiguous and have distinct failures. This is because nature works only in the other direction, as in 3DID. Real protein chains fold up from a denatured state to their native conformation under thermodynamic and/or kinetic guidance, rejecting all nonnative conformations. They are not attracted to a greater or lesser extent by some conformational template presented to them. Bidirectional "sequence/structure compatibility" is a bogus concept.

Then what should we do, since sequence identification is nevertheless a useful aid to suggesting function for novel sequences? One approach is to ac-

cept sequence identification as an imperfect screening procedure to be verified afterward by the opposite calculation. That is, first find all sequences in a data base that score relatively well with a proposed conformational template. Then for each of those sequences, calculate its native conformation. (This is of course an unsolved problem; we are relatively successful at 3DID,[19,20] but this is harder.) Only those candidate sequences that actually fold up to something resembling the template are true matches, but this is at least easier than predicting the fold for all sequences in the data base. Godzik comes to essentially this same conclusion by different means.[21]

Another intriguing direction for future study is to devise nonphysical scoring functions. Even in the square lattice model where we know that the true folding potential fails at sequence identification, it is not a foregone conclusion that there is no function of conformation and sequence that will succeed. All we can say at this point is that if such functions exist, they are not likely to be the same as those that solve 3DID.

REFERENCES

1. Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A., Sauer, R.T. Deciphering the message in protein sequences: Tolerance to amino acid substitutions. Science 247:1306–1310, 1990.
2. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
3. Orengo, C.A., Jones, D.T., Thornton, J.M. Protein superfamilies and domain superfolds. Nature 372:631–634, 1994.
4. Bowie, J.U., Clarke, N.D., Pabo, C.O., Sauer, R.T. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. Proteins 7:257–264, 1990.
5. 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.
6. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through the folding motif. Proteins 16:92–112, 1993.
7. Bryant, S.H., Altschul, S.F. Statistics of sequence-structure threading. Curr. Opin. Struct. Biol. 5:236–244, 1995.
8. Fetrow, J.S., Bryant, S.H. New programs for protein tertiary structure prediction. Biotechnology 11:479–484, 1993.
9. Lüthy, R., Bowie, J.U., Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature 356: 83–85, 1992.
10. 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.
11. Lathrop, R.H., Smith, T.F. Global optimum protein threading with gapped alignment and empirical pair score functions. J. Mol. Biol. 255:641–665, 1996.
12. Sippl, M.J., 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 13:258–271, 1992.
13. Wilmanns, M., Eisenberg, D. Inverse protein folding by the residue pair preference profile method: Estimating the correctness of alignments of structurally compatible sequences. Protein Eng. 8:627–639, 1995.
14. Chan, H.S., Dill, K.A. Transition states and folding dynamics of proteins and heteropolymers. J. Chem. Phys. 100:9238–9257, 1994.
15. Thomas, P.D., Dill, K.A. Statistical potentials extracted from protein structures: How accurate are they? J. Mol. Biol. 257:457–469, 1996.
16. Rooman, M.J., Wodak, S.J. Are database-derived potentials valid for scoring both forward and inverted protein folding? Protein Eng. 8:849–858, 1995.
17. Sippl, M.J. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J. Comput. Aided Mol. Design 7:473–501, 1993.
18. Crippen, G.M., Maiorov, V.N. Proposed rules of the protein folding game. In: "Protein Folds. A Distance Based Approach." Bohr, H., Brunak, S. (eds). New York: CRC Press, 1995:189–201.
19. Maiorov, V.N., Crippen, G.M. Contact potential that recognizes the correct folding of globular proteins. J. Mol. Biol. 227:876–888, 1992.
20. Maiorov, V.N., Crippen, G.M. Learning about protein folding via potential functions. Proteins 20:167–173, 1994.
21. Godzik, A. In search of the ideal protein sequence. Protein Eng. 8:409–416, 1995.