Equilibrium Folding–Unfolding Pathways of Model Proteins: Effect of Myoglobin–Heme Contacts

ROBERT L. JERNIGAN and SANZO MIYAZAWA,* *Laboratory of Mathematical Biology, DCBD, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20205*

Synopsis

The equilibrium of protein folding–unfolding has been investigated with a simple two-state, native and random-coil, model; we have termed this the globule–coil model. Energies are calculated by favoring native long-range contact pairs. Most-probable native domains are obtained at all stages of the transition; plausible folding pathways are constructed by connecting these domains by assuming simple growth. Even though native heme–protein contacts represent less than 6% of the total number of native contact pairs, their inclusion appears to change the folding pathway of apomyoglobin from the growth and merging of two native domains to the growth of a single domain. This indicates that pathways derived with this method may be critically sensitive to the details of the contact map and physical constraints during the folding process.

INTRODUCTION

A simple model based on the growth–merge mechanism was proposed by Wako and Saito[1] and later termed a noninteracting local structure model by Gō and Abe.[2] This is similar to the helix–coil model except that native domains of the protein correspond to the ordered state rather than helix regions. The principal features of this noninteracting globule–coil model are (1) each residue can be in either the native or random-coil state, and (2) energies of native regions are calculated by favoring the native long-range interactions observed between residues within a single native fragment. It is our intention to apply the model to realistic proteins and demonstrate its practicality. Previously, some detailed Monte Carlo generations[3] of trypsin inhibitor conformations have indicated that some classes of long-range interactions omitted in this simple model, such as those between separate native regions, were improbable. This provides some justification and motivation for further applications of this model.

In that calculation for trypsin inhibitor, detailed calculations of equilibrium folding pathways[3] were performed by means of Monte Carlo generations of conformations. The protein was regarded as a chain of hard-sphere C^{α} and C^{β} atoms. The backbone dihedral angles were permitted to take discrete values at every 10°. Approximate conformational free

* Address correspondence to the above address, Building 10, Room 4B 56.

Biopolymers, Vol. 22, 79–85 (1983)

CCC 0006-3525/83/010079-06$01.60

energies were formulated principally in terms of the number of long-range nativelike contacts. A total of 230,000 molecular conformations, ranging from the native state to the denatured state, were generated randomly by changing the sampling bias. The folding–unfolding coordinate was taken to be the conformational energy, E; conformational entropy, $S(E)$ was estimated for each value of E from the number of samples. Calculated free energies indicated the folding–unfolding transition for this idealized model to be an "all-or-none" type. Interresidue contact probabilities, averaged over samples at various stages of folding, served to characterize folding intermediates.

By assuming growth, most probable equilibrium pathways for the folding–unfolding transition were constructed by connecting conformationally similar intermediates. In such an equilibrium calculation, it is possible for unlike conformations to appear at adjacent points along the folding coordinate. In such cases, the assumption of growth indicates paths between nonadjacent conformations. Such a pathway construction often yields multiple folding pathways. The specific details of the folding pathway obtained for trypsin inhibitor were the following. (1) Folding began with the appearance of nativelike short-range contacts at a β-turn and at the α-helix. (2) These grew to include the native pair of interacting β-strands. This state included intact regular secondary conformations, as well as the interstrand sheet contacts, and corresponded to an activated state with the highest free energy on the pathway. (3) Additional native long-range contacts were completely formed either toward the amino terminus or toward the carboxyl terminus. These results were quite similar to those obtained in an application of the noninteracting globule–coil model, except that in step (3) there was growth only in the direction toward the carboxyl terminus. This similarity indicates that this model, in spite of its simplicity, may be useful for treating folding–unfolding transitions. Subsequently, we applied this model to the equilibrium folding of several proteins,[4,5] including apomyoglobin. Here we want to explore in a brief way the sensitivity of the results to the specific arrangement of contacts on the native contact map.

NONINTERACTING GLOBULE–COIL MODEL

Partially folded protein conformations are represented as native globules separated from other native globules by random-coil regions. No interactions are considered other than those within each separate native globule, and these are taken to correspond specifically to those within the native conformation. A partition function is formulated to include all numbers, sizes, and positions of native globules.[4] The premise of the conformational energies is that the native conformation is the lowest energy form. Conformational energies are taken to have two parts: (1) intraresidue contributions evaluated from the frequency distributions in (ϕ,ψ) compiled from 20 protein crystal structures by Némethy and Scheraga[6] and (2) int-

erresidue interactions based on the number of long-range native contacts. The free energy of the native region comprising residues i–j is given by[4]

$$F_{ij} = \sum_{k=i}^{j} f_k + n_{ij}^{c} f^{c} + (j - i + 1)\alpha$$

where f_k is the difference in free energy between the native and random-coil conformations of residue k from the Némethy-Scheraga frequencies. The number of long-range contact pairs, n^c is taken from the C^β–C^β contact map defined by distances $\leq$6.5 Å. Contact maps based on other distances, with changes principally in the density of contacts, did not appear to significantly affect the calculated folding pathways. The energy per contact pair, f^c, is the principal parameter, with values usually chosen between -1 and -2 kcal/mol. Note that it is taken here to be identical for all pairs. Some justification for this range of values is provided by the range of -1.3 to -1.5 kcal/mol estimated for hydrophobic contact energies by Janin and Chothia.[7] The parameter α is selected by choosing the melting point; this is achieved by adjustment of α until the native and denatured state free-energy minima are approximately equal.

The number of native residues, n, is employed directly as a folding–unfolding coordinate. A partition function is formulated for each number of native residues[4]; then, most-probable conformations are calculated at each point along this folding coordinate. By considering only conformations included in the partition function for each point along the transition coordinate, we can obtain detailed descriptions of the conformational characteristics of even relatively rare intermediates on the pathway. Equilibrium folding–unfolding pathways have been constructed by connecting similar conformations along the folding coordinate.

For all molecules studied, the later half of folding can be represented as nucleation from only a few sites, followed by growth. Typically, helixes and turns appear prior to formation of β-sheets; β-sheet formation usually corresponds to a large maximum in the free energy because of the attendant large loss of conformational entropy. Previously, it has been determined that free energies depend strongly on the value of the contact energy, but that, for the range of values from -1 to -2 kcal/mol, the most-probable conformations are relatively unaffected. More favorable contact energies produce sharper transitions. Proteins studied included concanavalin A, lysozyme, apomyoglobin, ribonuclease A, and trypsin inhibitor.

Here, we will consider the case of folding–unfolding pathways for myoglobin and apomyoglobin, i.e., with and without the heme group. The heme is regarded as part of residue 93. In the crystal structure of sperm whale metmyoglobin (1MBN from the Brookhaven Protein Data Bank[8]), there are 287 close contacts ($\leq$6.5 Å) between C^β atoms, excluding sequentially nearest-neighbors. In addition, there are 17 close contacts between heme atoms and C^β atoms, excluding residues 92–94. This includes the eight explicitly enumerated by the crystallographers. Parameters previously used[4] for apomyoglobin were $RT = 0.695$ kcal/mol, $f^c = -1.0$ kcal/mol, and

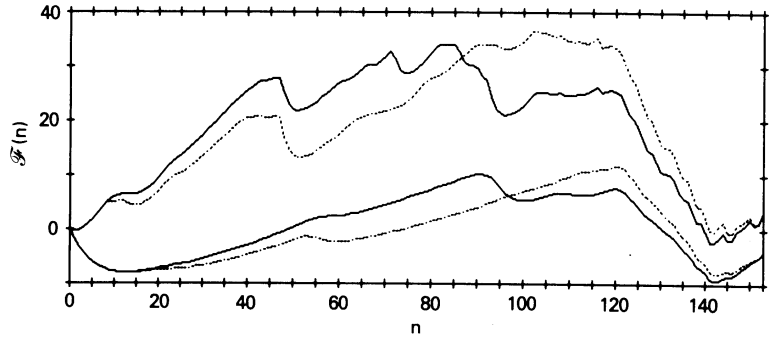

Fig. 1. Free energies (in kcal/mol) for myoglobin (solid lines) and apomyoglobin (dotted curves) at different stages of folding indicated by the number of native residues, n, along the abscissa. The top and bottom solid curves are for $f^c = -1.7$ and -0.944, respectively. The top and bottom dotted curves are for $f^c = -1.8$ and -1.0, respectively.

$\alpha = -1.55$ kcal/mol. We have adjusted the contact energy for holomyoglobin to maintain the total contact energy of apomyoglobin for the native structure; this yields $f^c = -0.944$ kcal/mol. This also corresponds closely to the melting condition for holomyoglobin.

RESULTS

The free energies calculated for holo- and apomyoglobin with two sets of contact energy parameters are shown in Fig. 1. The major effect of including the heme is to increase the relative number of long-range contacts. Previously, we concluded that a protein with relatively more long-range interactions evidences a free-energy maximum closer to the denatured side of the folding coordinate.[4] The heme–protein contacts are relatively long range in character and cause the maximum in the curves to shift toward the denatured side of the folding axis.

Most-probable native residues at all points along the folding coordinate are indicated by the dots in Fig. 2. For $1 < n \lesssim 30$ and $120 \lesssim n < 153$, the results for most-probable residues are almost identical for myoglobin and apomyoglobin. The probable native residues are similar from the denatured side up until about $n = 75$, except for the longer persistence of a small helix that includes the heme residue. The largest apparent differences between the two molecules occur in the range $90 \lesssim n \lesssim 120$; apomyoglobin has residues 100–139 in native form, but residues 77–99 are random. Myoglobin is the reverse. For the minimum at $n = 101$, all heme contacts are formed; this native heme pocket consisting of residues 21–115 includes complete helixes B, C, D, E, F, and G. Until the heme–protein contacts appear, the free energies are higher for myoglobin because of the larger contact energy parameter; subsequently, the appearance of heme–protein contact pairs serves to lower the free energy relative to apomyoglobin.

In Fig. 3 the growth-folding pathways inferred from the dots in Fig. 2 are shown. Stable conformations at both ends of the pathway are the same.

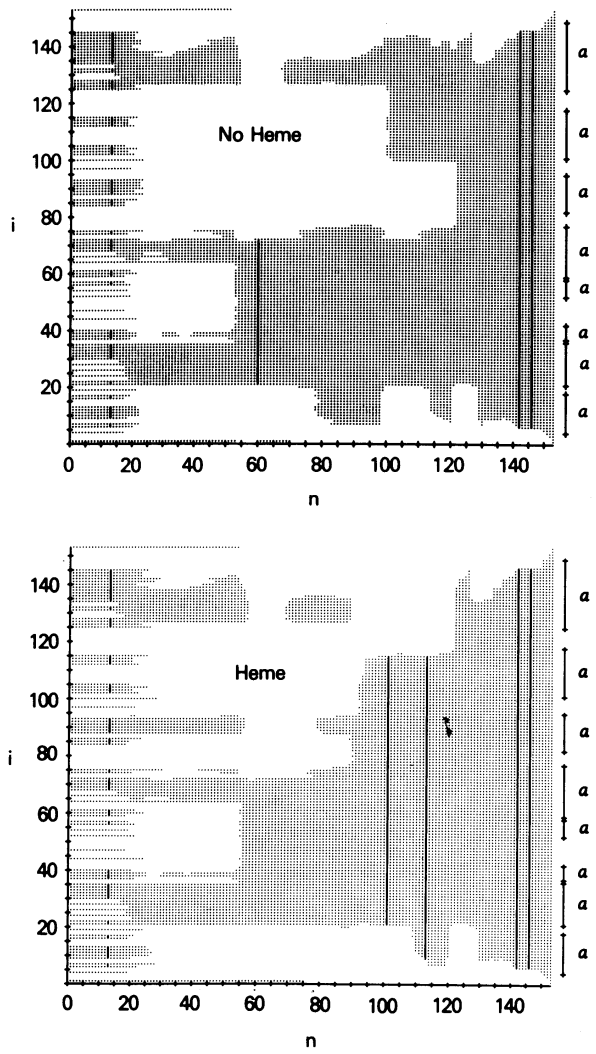

Fig. 2. Locations of most-probable native residues for apomyoglobin (top) and holomyoglobin (bottom) for all numbers of native residues, n. A dot indicates that the probability of residue i being native is greater than or equal to n/N, where N is the number of residues in the protein. Dots correspond to the most-probable native residues. Locations of free-energy minima in Fig. 1 are shown as solid lines. Regular α-helix regions are indicated by the vertical bars at the right side of the figure. Results are for $f^c = -0.944$ for myoglobin; $f^c = -1.0$, apomyoglobin.

The intermediates differ, however, with apomyoglobin showing a two nucleus growth pattern and holomyoglobin a single nucleus pattern. In particular, the apomyoglobin pathways indicate two growing native domains on either side of the heme at 93, which grow together only near n = 120. For myoglobin, the same stable nucleus at residues 21–72 next adds the heme to the stable nucleus about n = 90 and continues its growth in a regular way toward the carboxyl terminus.

Previous results[4,5] indicated relative insensitivity of folding pathways

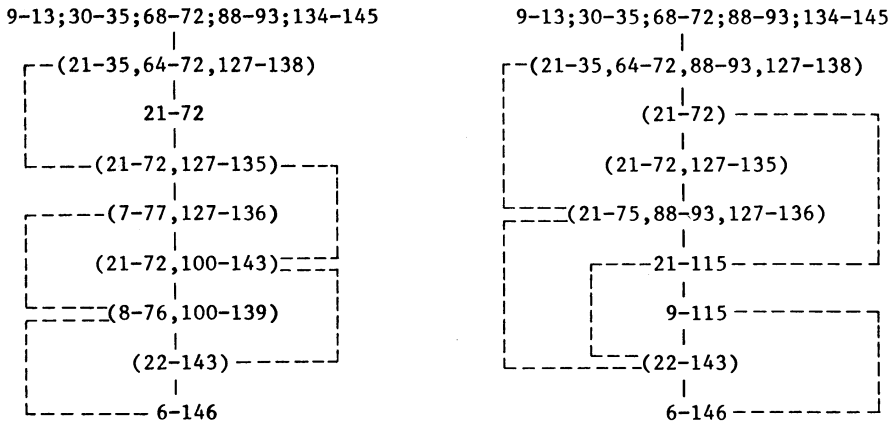

Fig. 3. Equilibrium folding–unfolding pathways for apomyoglobin (left) and myoglobin (right). Numbers designate native regions. Parentheses indicate a conformation that is not at a local free-energy minimum. Semicolons are placed between native regions that can appear only in separate molecules. Dashed lines are hidden pathways chosen to connect conformations by growth that are not adjacent to one another along the folding coordinate, n.

to contact energies when applied uniformly to all contact pairs, as well as to the distance used to define the contact map. In the latter case, additional contacts were included but represented only increased density of contacts, unchanged in character. The present result, in which relatively few additional characteristically different contacts have been included, indicates a strong effect on the folding pathways.

If a formulation of conformational energies in which some specific contacts were favored over others were appropriate, then one might expect large and significant effects on the folding pathway. Also, physical constraints during folding—such as attachment to the biosynthetic apparatus—might have large effects on folding pathways.

The present results are simple and indicate preferred folding pathways. The free energies in Fig. 1 should be interpreted only qualitatively; these values depend strongly on the poorly characterized contact energy parameter. Also, if some contacts were significantly more favorable than others, then the relative stabilities of intermediate conformations could be drastically affected. In particular, if the heme–protein contacts were substantially more favorable than others, then the conformation corresponding to the minimum at $n = 101$ could become more favorable as an intermediate. Such a substantial stabilizing role for the heme is suggested by the results of Leutzinger and Beychok,[9] who report that removal of the heme of hemoglobin results in a decrease of helicity from 75 to 30%. Certainly the heme has a large effect on the conformations of the globins. Also, the detailed study of Lesk and Chothia[10] indicates the importance of heme–protein contacts during evolution.

Some evidence has been offered that heme–protein contacts form early

in the folding process[9] and persist until late on the unfolding pathway.[11] The present intermediate given by the solid line at $n = 101$ includes most of the heme contacts. It is not likely, however, to correspond to an early stage of folding, since it represents a native nucleus composed of about two-thirds of the protein. However, Craik et al.[12] reported the strong binding of human β-globin fragment 31–104 (aligned[10] sequence number P43–P123) to heme. This provides some evidence to support the calculated stability of the intermediate of $n = 101$ composed of residues 21–115 (aligned sequence number P32–P133), since it consists principally of these same residues.

We are grateful to a referee for informing us about the experimental studies.

References

1. Wako, H. & Saito, N. (1978) *J. Phys. Soc. Jpn.* **44,** 1939–1945.
2. Go, N. & Abe, H. (1981) *Biopolymers* **20,** 991–1011.
3. Miyazawa, S. & Jernigan, R. L. (1982) *Biopolymers* **21,** 1333–1363.
4. Miyazawa, S. & Jernigan, R. L. (1982) *Biochemistry* **21,** 5203–5213.
5. Miyazawa, S. & Jernigan, R. L. (1983) *J. Stat. Phys.* (in press).
6. Némethy, G. & Scheraga, H. A. (1977) *Q. Rev. Biophys.* **10,** 239–352.
7. Janin, J. & Chothia, C. (1979) in *Protein: Structure, Function and Industrial Applications*, Pergamon Press, Oxford, UK, pp. 227–238.
8. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542 (The Brookhaven Protein Data Bank).
9. Leutzinger, Y. & Beychok, S. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 780–784.
10. Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol.* **136,** 225–270.
11. Kihara, H., Takahashi, E., Yamamura, K. & Tabushi, I. (1980) *Biochem. Biophys. Res. Commun.* **95,** 1687–1694.
12. Craik, C. S., Buchman, S. R. & Beychok, S. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 1384–1388.

Received June 20, 1982
Accepted September 2, 1982