PROTEINS: Structure, Function, and Genetics 23:38–48 (1995)

Rigid Domains in Proteins: An Algorithmic Approach to Their Identification

William L. Nichols,[1] George D. Rose,[2] Lynn F. Ten Eyck,[1,3] and Bruno H. Zimm[1]
[1]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093; [2]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205; [3]San Diego Supercomputer Center, San Diego, California 92186-9784

ABSTRACT **A rigid domain, defined here as a tertiary structure common to two or more different protein conformations, can be identified numerically from atomic coordinates by finding sets of residues, one in each conformation, such that the distance between any two residues within the set belonging to one conformation is the same as the distance between the two structurally equivalent residues within the set belonging to any other conformation. The distance between two residues is taken to be the distance between their respective α carbon atoms. With the methods of this paper we have found in the deoxy and oxy conformations of the human hemoglobin $\alpha_1\beta_1$ dimer a rigid domain closely related to that previously identified by Baldwin and Chothia (J. Mol. Biol. 129: 175–220, 1979). We provide two algorithms, both using the difference-distance matrix, with which to search for rigid domains directly from atomic coordinates. The first finds all rigid domains in a protein but has storage and processing demands that become prohibitively large with increasing protein size. The second, although not necessarily finding every rigid domain, is computationally tractable for proteins of any size. Because of its efficiency we are able to search protein conformations recursively for groups of non-intersecting domains. Different protein conformations, when aligned by superimposing their respective domain structures, can be examined for structural differences in regions complementing a rigid domain.**

Key words: difference-distance matrix, hemoglobin rigid core, structure search

INTRODUCTION

Structural domains in proteins have been defined in numerous ways, among the better known being visually recognizable conformational regions[1] and sets of proximate residues within difference maps.[2] More quantitative definitions include clustering,[3] use of cutting planes,[4] minimization of interfacial surface area,[5] maximization of solvent exclusion,[6] minimization of specific volume,[7] isolation of coherent regions from normal mode analysis,[8] and maximization of compactness.[9]

In this paper tertiary structures existing in different protein conformations define a rigid domain if the distance between any two residues of the rigid domain structure in one conformation is the same as the distance between the two equivalent residues of the rigid domain structure in every other conformation. The distance between two residues is defined to be the distance between their respective α carbon atoms, which can be found from atomic coordinates. The tertiary structures defining a rigid domain in different protein conformations are geometrically congruent and can be superimposed by aligning their equivalent residues. The residues of a domain (we shall refer to a rigid domain simply as a domain) do not have to be sequentially or spatially contiguous. The conformations being searched for domains must have their primary sequences at least partially aligned prior to implementing the algorithms of this paper. For this reason our methods are easily applied to the T and R states of an allosteric protein. No assertions are made about the persistence of structural rigidity of a domain along transitional pathways between conformations.

Figure 1 illustrates the concept of a domain in an eight-residue peptide with two conformations, A and B. The heavy line connecting successive α carbon atoms represents the peptide backbone. Distances between all pairs of α carbon atoms are shown by dashed lines in conformation A. Five residues form a domain within the peptide, as shown by the dashed lines in conformation B, which indicate that the distances between all pairs of residues in the domain are the same in both conformations. No one of the other three residues can be included in the domain because its distance from at least one of the five residues of the domain is not the same in conformation A as in conformation B.

We would like to have tertiary structures that are nearly but not exactly congruent to each other nev-

Received December 19, 1994; revision accepted February 9, 1995.

Address reprint requests to William L. Nichols, Department of Chemistry and Biochemistry 0317, University of California, San Diego, La Jolla, California 92093.

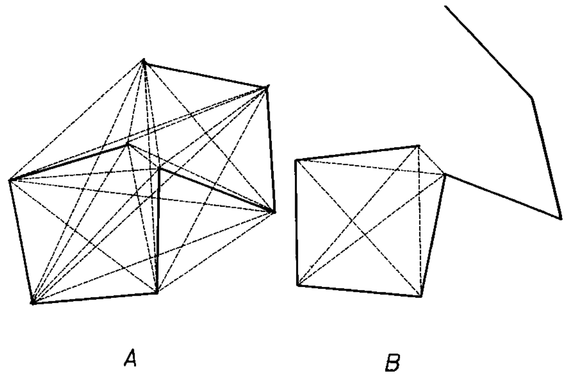

Fig. 1. A conformational change in an eight-residue peptide in which five residues form a domain. Dashed lines connect all pairs of residues in the initial structure (**A**) on the left, but only pairs from the five residues that form a domain in **B** on the right. The solid line represents the α-carbon backbone of the peptide.

ertheless to define a domain. For these structures, distances between equivalent residues will differ among conformations. Differences in distances can arise from insignificant dissimilarities between structures defining a domain or from experimental uncertainty in coordinates. To allow us to include geometrically incongruent structures as domains we generalize our definition of a rigid domain by specifying a parameter ε so that the distance between two residues of a domain in one conformation can differ from the distance between the structurally equivalent residues of the domain in another conformation by as much as ε. The number of residues included in a domain then depends on the value chosen for ε. Domains found with small values of ε reveal more detailed differences in structure between conformations, while domains found with larger values of ε identify gross similarities among conformations. When searching a group of protein conformations for domains, a good initial choice for ε is the precision measure of the atomic coordinates. The efficacy of the methods of this paper for identifying structural similarities in protein conformations is due in part to their not relying upon a least-square measure of similarity to identify domains but only upon the maximum absolute deviation in inter-residue distance as given by ε.

Two distinct domains may have residues in common or be entirely disjoint. The minimum number of residues in a domain can be as small as two and still be consistent with our definition, but the maximum number of residues in a domain is limited only by the number of residues in the protein. The hemoglobin molecule can serve to illustrate these points. Hemoglobin consists of two monomers, termed α and β. Two copies of each monomer associate to form the native tetramer. The hemoglobin structure has been solved by X-ray crystallography in both oxy and deoxy forms. Coordinates for human deoxyhemoglobin[10] and human oxyhemoglobin[11] were obtained from the Protein Data Bank[12] as entries 2HHB and 1HHO, respectively. The search of residues 1–42 of the α_1 monomer (the N-terminus, A, B, and C helices) for domains whose inter-residue distances differ by no more than 0.30 Å between deoxy and oxy conformations finds 643 domains with two residues each. The five largest domains have 21 residues each and have 18 residues in common. These results, appearing in Figure 2, will be discussed further in following sections.

$\varepsilon = 0.30$Å

J	N(J)	$C(R,J)$
2	643	861
3	5341	11480
4	29121	111930
5	114643	850668
6	343572	5245786
7	808298	26978328
8	1520258	26978328
9	2311635	445891810
10	2861660	1471442973
11	2895704	4280561376
12	2398436	11058116888
13	1623937	25518731280
14	894886	52860229080
15	398055	98672427616
16	141005	166509721602
17	38937	254661927156
18	8098	353697121050
19	1196	446775310800
20	112	513791607420
21	5	538257874440

Largest rigid domains:

{7 8 10 20 21 23 24 25 26 27 28 29 30 33 35 36 37 38 39 41 42} RMS = 0.205Å
{7 8 10 20 21 23 24 25 26 27 28 29 30 33 36 37 38 39 40 41 42} RMS = 0.209Å
{7 8 20 21 23 24 25 26 27 28 29 30 31 33 35 36 37 38 39 41 42} RMS = 0.208Å
{7 8 20 21 23 24 25 26 27 28 29 30 31 33 36 37 38 39 40 41 42} RMS = 0.213Å
{7 8 20 21 23 24 25 27 28 29 30 31 32 33 36 37 38 39 40 41 42} RMS = 0.213Å

Residues common to all of the largest rigid domains:

{7 8 20 21 23 24 25 27 28 29 30 33 36 37 38 39 41 42}

Fig. 2. An exhaustive determination of domains within the N-terminus, A, B, and C helices of the α_1 monomer of human hemoglobin with $\varepsilon = 0.30$ Å. The number of residues in a domain is J. The number of domains with J residues is $N(J)$. The number of possible sets with J residues that can be found from all the R = 42 residues is $C(R,J)$. For the peptide of this example, the largest rigid domains have 21 residues. The residues in each of the five largest domains are listed at the bottom of the figure along with the 18 residues common to all. Root-mean-square values for the fit of the oxy onto the deoxy domains are given to the right of each.

Different protein conformations can be aligned by superimposing a common domain. A measure of how well domain structures align is the root-mean square (RMS) fit of superimposed domain residues:

$$\mathrm{RMS} = \sqrt{\sum_i \frac{\Delta x_i^2 + \Delta y_i^2 + \Delta z_i^2}{R}}.$$

$\Delta x^2 + \Delta y^2 + \Delta z^2$ is the squared distance between corresponding residues in two different superimposed domain structures of R residues each. RMS comparisons of entire conformations tend to conceal small differences in structure, differences that are readily apparent when domain structures found with a sufficiently small ε are superimposed. Many others have addressed aspects of structural alignment, among them Vriend and Sander[13] and Holm and Sander.[14]

Within the hemoglobin tetramer, the identical $\alpha_1\beta_1$ and $\alpha_2\beta_2$ dimers undergo a quarternary reorientation relative to one another with the change from the deoxy to the oxy conformation.[15] Baldwin and Chothia[16] found a group of residues within the $\alpha_1\beta_1$ dimer interface whose α carbon atoms remain rigid during this allosteric transition. With our methods we find such a set of α carbon atoms as well and refer to it as the rigid core of the dimer. The dimer core can be used to align the oxy and deoxy structures to reveal conformational changes with ligand binding.

METHODOLOGY

Calculating the Difference-Distance Matrix

The initial computational step for finding domains from atomic data is to construct distance matrices and to use them to find the difference–distance matrix. For a given conformation, elements of a distance matrix D_{ij} are the distances between α carbon atoms i and j. With $D_{ij}^{(1)}$ the distance matrix for one conformation and $D_{ij}^{(2)}$ the distance matrix for another conformation, the difference–distance matrix[17–19] is given by the absolute value of the matrix difference,

$$\Delta_{ij} = | \mathrm{D}_{ij}^{(1)} - \mathrm{D}_{ij}^{(2)} |. \tag{1}$$

For computational efficiency, a matrix δ_{ij} is defined such that if

$$\begin{aligned} \Delta_{ij} &> \varepsilon, \ \delta_{ij} = 0, \\ \Delta_{ij} &\leq \varepsilon, \ \delta_{ij} = 1. \end{aligned} \tag{2}$$

Residue pairs i, j with a change of inter-residue distance (between α carbon atoms) of no more than ε Å have matrix elements $\delta_{ij} = 1$; otherwise $\delta_{ij} = 0$. The value chosen for ε will depend on the purpose of the calculation, being small if we seek subtle differences between conformations or large if we are searching for gross similarities among conformations.

Exhaustive Search For Domains

An exhaustive search for domains within a polypeptide constructs all rigid residue pairs and from this set finds rigid triples, quadruples, and so forth, until all combinatorial possibilities have been considered. That is, an exhaustive search begins with the set of all pairs of residues i, j for which $\delta_{ij} = 1$ and from this set finds the set of all distinct rigid triplets of residues i, j, k. A triplet is rigid if $\delta_{ij} = 1$ for i and j any of the residues in the triplet. The search is iteratively enlarged by finding every domain with $J + 1$ residues from each domain with J residues until all possible combinations of residues have been exhausted. Figure 2 illustrates this method. The number of residues in a domain is J while $N(J)$ is the number of distinct domains with J residues. The binomial coefficient $C(R,J)$ is the number of possible subsets of J residues, rigid or otherwise, that can be found in a set of R residues. The number of domains $N(J)$ for each J can be fairly large, although it is usually much smaller than the number of possible subsets $C(R,J)$. From Figure 2, for example, 4,280,561,376 subsets of 11 residues exist within a set of 42 residues, but only 2,895,704 domains of 11 residues each can be identified when $\varepsilon = 0.30$ Å in the first 42 residues (the N-terminus, A, B, and C helices) of the α_1 monomer of human hemoglobin. Figure 2 is further examined in the Discussion section. Searching for domains in this way is computationally demanding because the number of subsets in a set of R residues increases rapidly with R. We need to look for a faster way to find domains.

An Incomplete but Fast Search for Domains

We now introduce an alternative method for finding domains that is computationally feasible for a polypeptide of arbitrary length. With this method, only a single domain is identified for J residues [rather than the $N(J)$ domains required for an exhaustive search], with J near the maximal domain size. The complement of this domain is then searched exhaustively to identify all larger domains that include it as a subdomain. The method saves considerable computational effort and will still find domains suitable for conformational comparison.

We assert that a residue i differing by a small amount in relative position from one conformation to another will have many of its $\delta_{ij} = 1$, while a residue differing by a large amount in relative position will have many of its $\delta_{ij} = 0$. To quantify the changes in residue position, sums S_i of δ_{ij} are evaluated for each residue i over all other residues j in the polypeptide:

$$S_i = \sum_j \delta_{ij}. \tag{3}$$

The residues for which position differs the least between conformations tend to have the largest S_i while those for which position differs the most between conformations tend to have the smallest S_i.

The search for a domain is initiated by choosing an integer N_s and finding all those residues i for which $S_i \geq N_s$. The set of residues for which this condition is true is defined as $U_\varepsilon(N_s)$. $U_\varepsilon(N_s)$ will be the entire protein when N_s is zero and will have only the more rigid residues as N_s approaches the number of residues in the protein. $U_\varepsilon(N_s)$ is usually not itself a domain, because distance matrix elements between residues in $U_\varepsilon(N_s)$ for one conformation can differ by more than ε from those for other conformations.

However, the following strategy will find at least one domain in $U_\varepsilon(N_s)$, if any exist.

For each residue i in $U_\varepsilon(N_s)$ find all other residues j in $U_\varepsilon(N_s)$ for which $\delta_{ij} = 0$, that is, such that the residue pair i, j is not rigid. The residue i that has the largest number N_M of residues j for which $\delta_{ij} = 0$ is removed from $U_\varepsilon(N_s)$, leaving a subset of $U_\varepsilon(N_s)$ with one less residue. [The subsets of $U_\varepsilon(N_s)$ will be affected by the order in which residues with the same value for N_M are deleted.] The reduction is repeated until $N_M \leq 1$. This subset of $U_\varepsilon(N_s)$ is a domain except for possible non-rigid pairs of residues. A domain $D_\varepsilon(N_s)$ can then be constructed from this subset by removing all non-rigid pairs.

$D_\varepsilon(N_s)$ can be enlarged by searching its complement exhaustively to find all domains that preserve $D_\varepsilon(N_s)$ as a subdomain. Among these will be domains found by adding back some residues that were previously removed as non-rigid pairs in $U_\varepsilon(N_s)$, but other domains are often discovered as well.

Constructing a domain of J residues using the method described in this section is much faster than finding one by exhaustive enumeration of all $N(K)$ possibilities, as K grows from 2 to J. By choosing N_s appropriately, the domains found by enlarging $D_\varepsilon(N_s)$ will generally be maximal or, if not, will have residues in common with the maximal domains. The algorithm to find $D_\varepsilon(N_s)$ is most efficient for values of N_s near R, the number of residues in the protein, because construction of the domain $D_\varepsilon(N_s)$ is computationally fast, and the exhaustive search through the complement of $D_\varepsilon(N_s)$ will not have to find many larger domains.

A domain can be used to align protein conformations by translating the centroid of the domain for each conformation to the coordinate origin and rotating the domain of one conformation onto that of the others with methods originally described by Kabsch.[20,21] The resulting transformed coordinates give a least-square fit between the domains of the different conformations. The entire protein can now be visually or numerically investigated for conformational differences in other regions.

A summary of the above algorithm follows.

```
I. Read the coordinates of all residues $i$ for each
   conformation.
II. Construct the distance and difference-distance
    matrices.
    A. Choose $\varepsilon$. [See the remarks about choosing $\varepsilon$
       after Eq. (2) above.]
    B. Calculate the difference-distance matrix $\Delta_{ij}$
       for all pairs of residues $i, j$.
    C. If $\Delta_{ij} > \varepsilon$ then $\delta_{ij} = 0$; otherwise $\delta_{ij} = 1$.
III. Find a domain (not necessarily the largest).
    A. Choose $N_s$.
    B. Calculate $S_i$ for each residue $i$.
    C. For each $i$, if $S_i \geq N_s$ then include residue $i$ in
       the set $U_\varepsilon(N_s)$.
    D. For each $i$ in the set $U_\varepsilon(N_s)$, find all residues
       $j$ also in $U_\varepsilon(N_s)$ for which $\delta_{ij} = 0$.
    E. Remove from $U_\varepsilon(N_s)$ that residue $i$ that has
       the most other residues $j$ for which $\delta_{ij} = 0$.
    F. When for every residue $i$ remaining in $U_\varepsilon(N_s)$
       at most only one other residue $j$ can be found
       for which $\delta_{ij} = 0$, remove both $i$ and $j$ from
       $U_\varepsilon(N_s)$ to give $D_\varepsilon(N_s)$. Otherwise repeat III.D.
       and III.E.
IV. Search for larger domains.
    A. Examine each residue $j$ in the complement of
       $D_\varepsilon(N_s)$ to see if $\delta_{ij} = 1$ for all residues $i$ in
       $D_\varepsilon(N_s)$.
    B. For each $j$ for which all $\delta_{ij} = 1$ in IV.A., in-
       clude $j$ in $D_\varepsilon(N_s)$ to form a domain one residue
       larger.
    C. Repeat IV.A. and IV.B. with each such do-
       main until no larger domains can be found.
```

DISCUSSION

We illustrate the methods defined above by searching for domains in the first $R = 42$ residues of the α_1 monomer of human hemoglobin (the N-terminus, A, B, and C helices). (The reader will please note that we are using hemoglobin as a convenient example for the application of these methods; we make no pretense to a thorough study of this protein in this paper.) As previously, J is the number of residues in a domain, $N(J)$ is the number of domains with J residues, and $C(R,J)$ is the number of possible subsets of J residues in a set of R residues. Figure 2 outlines an exhaustive search for domains within this peptide when $\varepsilon = 0.30$ Å. The number of rigid residue pairs is 643 while the number of possible pairs of residues is 861. Similarly, the number of possible triplets is 11,480, but only 5,341 of them are rigid. The number of possible sets $C(R,J)$ grows combinatorially with J, while the number of domains eventually converges to 5 when J is 21. The 18 residues common to all five largest domains are listed at the bottom of the figure. The rigidity of these five largest domains is assessed by superimposing the deoxy and oxy structures,[20,21] with RMS fits as shown in the Figure. For conformation alignment all the largest domains are effectively the same, as can be seen in Figure 3.

The computational demands of an exhaustive search for domains are apparent in Figure 2. The number of sets with J or fewer residues that could be rigid is the sum of all the binomial coefficients $C(R,J)$ from 2 through J, a number that grows exponentially with J. The fast search avoids such an encumbrance by finding only one of the $N(J)$ domains, the domain $D_{0.30}(N_s)$, and exhaustively enlarging only this one domain. With a value $N_s = 25$, the search of the difference-distance matrix results in a set $U_{0.30}(25)$ of 35 residues:

$$U_{0.30}(25) = \{1\,2\,3\,6\,7\,8\,9\,10\,13\,14\,18\,19\,20\,21\,22\,23\,24\,25\,26\,27\,28\,29\,30\,31\,32\,33\,34\,35\,36\,37\,38\,39\,40\,41\,42\}.$$

The residue belonging to the most non-rigid residue pairs is residue 14, which is not rigid with $N_M = 12$ other residues in $U_{0.30}(25)$ (see table below). After deleting residue 14 from $U_{0.30}(25)$, the subsequent search for the residue with the largest sum N_M finds residue 22, with ten non-rigid pairs. Iteration until N_M is not larger than 1 results in the removal of 13 residues from $U_{0.30}(25)$, which leaves a set of 22 residues, the last one removed being residue 40. The following summarizes the deletion of residues from $U_{0.30}(25)$:

J	EXCLUDED RESIDUE	N_M
35	14	12
34	22	10
33	34	8
32	26	7
31	23	6
30	10	6
29	21	5
28	7	5
27	25	4
26	29	3
25	8	3
24	42	2
23	40	2

A set of 22 residues with two non-rigid residue pairs, (19,41) and (32,35), remains after the removal of residue 40. Deleting both non-rigid residue pairs leaves a domain $D_{0.30}(25)$ with 18 residues:

$D_{0.30}(25) = \{1\ 2\ 3\ 6\ 9\ 13\ 18\ 20\ 24\ 27\ 28\ 30\ 31\ 33\ 36\ 37\ 38\ 39\}$.

$D_{0.30}(25)$ is only one of the 8,098 domains with 18 residues found exhaustively in Figure 2. An exhaustive search through the complement of $D_{0.30}(25)$ finds four domains with 20 residues each, the two additional residues being one from each of the non-rigid residue pairs (19,41) and (32,35):

{1 2 3 6 9 13 18 19 20 24 27 28 30 31 32 33 36 37 38 39} RMS = 0.191 Å
{1 2 3 6 9 13 18 19 20 24 27 28 30 31 33 35 36 37 38 39} RMS = 0.186 Å
{1 2 3 6 9 13 18 20 24 27 28 30 31 32 33 36 37 38 39 41} RMS = 0.213 Å
{1 2 3 6 9 13 18 20 24 27 28 30 31 33 35 36 37 38 39 41} RMS = 0.205 Å

The RMS value for the superposition of each oxy domain upon its deoxy counterpart is listed after each domain. No other residues could be found in the complement of $D_{0.30}(25)$ that would fit rigidly in any of the four domains listed above.

We now compare the above with the exhaustive search of Figure 2, which revealed 112 domains of 20 residues each but only 5 domains of 21 residues each. The first two domains above are actually more rigid, in the sense of smaller RMS, than any of the five domains with 21 residues found exhaustively in Figure 2. The four domains found above also intersect extensively with the 5 of 21. Of the 18 residues common to the five domains with 21 residues, 10 also occur in the above four domains of 20 residues. Thus the four domains found above are close approximations to the five largest domains found exhaustively in the structure, which themselves, because of their extensive overlap, represent what is essentially one domain. As a matter of interest, the RMS fit for the entire peptide using the centroid of all 42 residues is 0.324 Å.

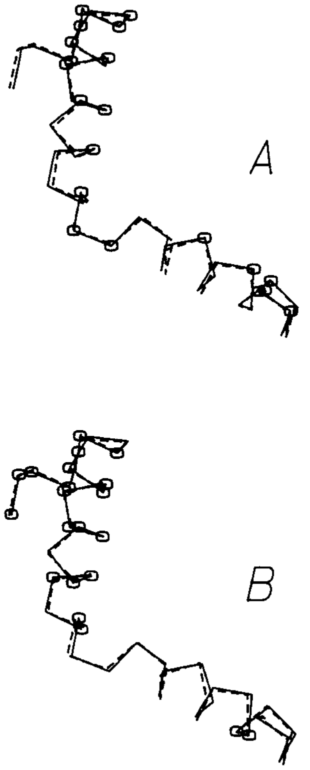

Fig. 3. The superposition of the oxy (dashed) upon the deoxy conformation of the A through C helices of the α_1 monomer of hemoglobin. **A:** Superposition using a domain found from $D_{0.30}(25)$ with an RMS fit of 0.186 Å. **B:** Superposition using a domain found from $D_{0.30}(36)$ with an RMS fit of 0.213 Å. The N-terminus is at the lower right and C-terminus of the C helix is at the upper left for both A and B views of the superimposed peptide. The graphical superpositions illustrate the differences between these two domains but show their equivalence for structural comparison.

We resume our search of the first 42 residues of the α_1 monomer of human hemoglobin to see the dependence of both $U_\varepsilon(N_s)$ and $D_\varepsilon(N_s)$ upon N_s. Repeated trials show that 36 is the smallest value for N_s that leads to a set $U_{0.30}(N_s)$ of residues common to all the largest domains of the exhaustive search of Figure 2. Because $U_{0.30}(36)$ is a domain and no residues need to be removed from it, $U_{0.30}(36)$ and $D_{0.30}(36)$ are identical.

$U_{0.30}(36) = D_{0.30}(36) = \{20\ 27\ 29\ 30\ 33\ 41\}$.

The complement of $D_{0.30}(36)$ is now searched for residues rigid with those already in $D_{0.30}(36)$, adding such residues one at a time to $D_{0.30}(36)$, just as in Figure 2, but now with many fewer combinations to examine. This gives the following table, in which the first column is the size of a domain and the second is the number of domains of that size for which $D_{0.30}(36)$ is a subdomain:

J	N(J)
6	1
7	25
8	245
9	1306
10	4478
11	10816
12	19290
13	26008
14	26827
15	21278
16	12953
17	5981
18	2038
19	485
20	72
21	5

No larger domains exist to 0.30 Å. The five domains of 21 residues each in the last line of the table are identical to those revealed by the exhaustive search of Figure 2.

Thus we have obtained with much less computational effort the results of the exhaustive search and have also found other slightly smaller domains which are actually more rigid. A sample of the results is shown in Figure 3, where a selected two of the domains have been used to align the first 42 residues of α_1 human hemoglobin. The peptide is aligned in (A) by superimposing the oxy and deoxy forms of the 20 residue domain with an RMS fit of 0.186 Å derived from $D_{0.30}(25)$, while the alignment in (B) results from similarly superimposing the last of the five domains listed at the bottom of Figure 2, which has an RMS fit of 0.213 Å.

The various domains that we have found can be thought of as constituting a family of closely related domains that represent one rigid object with minor variations. As seen in Figure 3, two domains of this family are practically equivalent for alignment purposes.

A Method for Finding Rigid Domains From a Subdomain of $D_\varepsilon(N_s)$

Larger domains can be found from the domain $D_\varepsilon(N_s)$ by including with $D_\varepsilon(N_s)$ all combinations of residues within the complement of $D_\varepsilon(N_s)$ that are rigid with $D_\varepsilon(N_s)$. All these larger domains contain $D_\varepsilon(N_s)$ as a subset. In place of $D_\varepsilon(N_s)$, however, any subset of $D_\varepsilon(N_s)$ can be used as the domain common to all subsequently larger domains, since any subset of $D_\varepsilon(N_s)$ is also a domain. By doing so we can find domains that have specific characteristics we may wish to retain. For example, residues that are either spatially or sequentially separate from the rest of the residues in $D_\varepsilon(N_s)$ can be eliminated. The subsequent search through the complement of this subdomain of $D_\varepsilon(N_s)$ will then lead to larger domains within the protein that retain the desired residues of the subdomain.

We show how this works with the example of the previous section, the first 42 residues of the α_1 monomer of human hemoglobin. $D_{0.30}(25)$ includes residues 1, 2, 3, 6, 9, 13, and 18, all of which are part of the A helix or N-terminus. Removing these from $D_{0.30}(25)$ leaves a subdomain lying only within the B and C helices:

$$D_{0.30}(25)_{mod} = \{20\ 24\ 27\ 28\ 30\ 31\ 33\ 36\ 37\ 38\ 39\}.$$

A search through the complement of this subdomain of $D_{0.30}(25)$ finds three of the five largest domains found with the exhaustive search of Figure 2. These domains have more B helix (residues 20–35) and less A helix (residues 3–18) than the four domains of 20 residues each first found above. The search is as follows:

J	N(J)
11	1
12	21
13	162
14	626
15	1366
16	1780
17	1424
18	709
19	218
20	39
21	3

A modest computational effort can sample a family of domains in a polypeptide and thereby escape the exponentially increasing demands for processor time and storage space required by an exhaustive search.

RESULTS

Rigid Core of the $\alpha_1\beta_1$ Dimer of Human Hemoglobin

The non-exhaustive method for finding domains is applied here to the entire $\alpha_1\beta_1$ human hemoglobin dimer. Exhaustively finding domains in a hemoglobin dimer would be unacceptably slow. The two conformations of the dimer in which we shall look for domains are the deoxy conformation 2HHB[10] and the oxy conformation 1HHO,[11] both from the Protein Data Bank.[12] A hemoglobin dimer is constructed from two monomers α_1 and β_1 with 141 and 146 residues, respectively.

We first search for domains in the $\alpha_1\beta_1$ hemoglobin dimer when $\varepsilon = 0.50$ Å. After the difference matrices for both the deoxy and oxy dimers have been constructed, the difference-distance matrix is computed, and a trial value of 200 is chosen for N_s. This is shown in Figure 4, in which the 104 residues in the set $U_{0.50}(200)$ are listed at the top. The element with the largest number N_M of non-rigid residues is residue $\alpha_1 11$, a lysine residue in the A helix of the α_1 monomer. This residue is removed from $U_{0.50}(200)$ and the number of non-rigid residues for each residue in the remaining subset is recalculated. The next least rigid residue is $\alpha_1 72$, a histidine residue in the beginning of the α_1 E-F corner. This is then removed. Residues are successively deleted in this manner until a subset of residues remains for which each member is not rigid with at most one other residue in the subset. This subset of 86 residues has six non-rigid residue pairs: ($\alpha_1 3$,$\alpha_1 8$), ($\alpha_1 5$,$\alpha_1 109$), ($\alpha_1 9$,$\beta_1 51$), ($\alpha_1 10$,$\beta_1 115$), ($\alpha_1 103$,$\alpha_1 120$), and ($\beta_1 22$,$\beta_1 130$). Removing all the non-rigid residue pairs gives the set $D_{0.50}(200)$, which has 74 residues. The search through the complement of $D_{0.50}(200)$ gives 64 domains, each with 88 residues. These are the largest domains found with $\varepsilon = 0.50$ Å. The eight residues in these domains of 88, in addition to residues in the above-listed non-rigid pairs, are residues $\alpha_1 24$, $\alpha_1 27$, $\alpha_1 30$, $\alpha_1 33$, $\alpha_1 34$, $\alpha_1 111$, $\beta_1 119$, and $\beta_1 124$. The residues of one of the largest domains are listed at the bottom of Figure 4 along with the RMS fit of the oxy onto the deoxy domain.

We show a closely related domain in Figure 5, which was found with $\varepsilon = 0.75$ Å and $N_s = 210$. $U_{0.75}(210)$ has 196 residues and leads to a domain $D_{0.75}(210)$ with 131 residues. An exhaustive search through the complement of $D_{0.75}(210)$ finds 16 domains with 135 residues each, the four residues in addition to those of $D_{0.75}(210)$ being one from each of four non-rigid residue pairs removed from $U_{0.75}(210)$ when finding $D_{0.75}(210)$. One of these 16 domains is labeled by circles in Figure 5. The 135 residue domains include most of the center of the dimer, with 81 residues from the α_1 monomer and 54 from the β_1 monomer. We call the family of 16 domains, of which this is one representative, the rigid core of the dimer to 0.75 Å.

The deoxy residues of the chosen representative of the rigid core shown in Figure 5 superpose on the oxy residues with an RMS value of 0.333 Å. To avoid cluttering Figure 5, only the deoxy hemes have been drawn. The change of conformation of both heme pockets relative to the rigid core is very apparent in Figure 5. The domain we have called the rigid core and its 15 relatives, which differ by only several residues, form a family of intersecting domains that represent minor variations of the same structure.

Effect of Changing ε

Non-zero difference-distance matrix elements Δ_{ij} defined by Eq. (1) owe their magnitude to either actual differences γ_{ij} in tertiary structure between conformations, some of which might be from correlated differences in domain orientation, or to uncorrelated differences χ_{ij} attributable to random error in measurement. If the distance between residues i and j in conformation (1) is $D_{ij}^{(1)}$ and the distance between the same residues in conformation (2) is $D_{ij}^{(2)}$, then

$$D_{ij}^{(2)} = D_{ij}^{(1)} + \gamma_{ij} + \chi_{ij}\,. \tag{4}$$

Matrix elements Δ_{ij} are then

$$\Delta_{ij} = |\gamma_{ij} + \chi_{ij}|. \tag{5}$$

We will assume no correlation between the Δ_{ij} and will consider only those differences χ_{ij} from experimental error. Certaintly, $\Delta_{ij} = \chi_{ij}$ if the peptide being considered is perfectly rigid. The following question arises: is the identification of a family of rigid domains critically dependent on the value chosen for ε? With the rigid core of hemoglobin as an example, the circles in Figure 6 show how the number of residues included in the domain increases with ε. The behavior of these points is the expected consequence of the limited precision of the data. When ε is chosen to be small, few residues meet the criterion, but this number increases rapidly as ε is increased. This thought can be put in quantitative terms by the following argument.

Suppose that we have a set of N residues in one conformation of a protein and we have found all the connecting distances r_{ij}. Because of limited experimental precision the positions of the residues contain small errors that are assumed to be Gaussianly distributed. We also have another conformation of the protein with the same set of N residues with different errors drawn from the same Gaussian distribution. As a result, both sets of r_{ij} contain Gaussianly distributed random errors. From Eq. (1) the probability distribution of the elements of the difference–distance matrix for this set of N residues is now

$$W(\Delta_{ij}) = \frac{2}{(2\pi)^{1/2}\sigma}\exp\left(-\frac{\Delta_{ij}^2}{2\sigma^2}\right) \tag{6}$$

where σ is the standard deviation of the errors in the Δ_{ij}. A subset of the N residues rigid to an ε similar in value to σ will have M residues, with M less than or equal to N. Because the Δ_{ij} are assumed to be uncorrelated, the probability that any subset containing M elements of the rigid set will meet the ε criterion is

$$\left(\int_0^{\varepsilon} W(\Delta)d\Delta\right)^{M(M-1)/2} \tag{7}$$

$U_{0.50}(200)$
α_1 residues:
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 19 20 31 35
36 38 39 41 42 43 60 70 71 72 102 103 104 105 106 108 109 110 113 115
116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133
β_1 residues:
15 16 17 20 22 23 32 33 34 35 37 38 50 51 52 53 55 82 107 108
109 110 111 112 113 114 115 116 117 118 120 121 122 123 125 126 127 128 129 130
131 132 133 134 135 136

J	EXCLUDED RESIDUE	N_M
104	$\alpha_1 11$	24
103	$\alpha_1 72$	20
102	$\alpha_1 4$	19
101	$\beta_1 20$	16
100	$\alpha_1 16$	11
99	$\beta_1 55$	8
98	$\alpha_1 12$	8
97	$\alpha_1 15$	7
96	$\beta_1 108$	6
95	$\alpha_1 115$	5
94	$\alpha_1 106$	5
93	$\alpha_1 71$	5
92	$\alpha_1 14$	5
91	$\beta_1 38$	4
90	$\alpha_1 105$	4
89	$\beta_1 133$	3
88	$\alpha_1 118$	3
87	$\beta_1 82$	2

Remaining non-rigid pairs: ($\alpha_1 3$,$\alpha_1 8$) ($\alpha_1 5$,$\alpha_1 109$) ($\alpha_1 9$,$\beta_1 51$) ($\alpha_1 10$,$\beta_1 115$) ($\alpha_1 103$,$\alpha_1 120$) ($\beta_1 22$,$\beta_1 130$)

$D_{0.50}(200)$
α_1 residues:
1 2 6 7 13 19 20 31 35 36 38 39 41 42 43 60 70 102 104 108
110 113 116 117 119 121 122 123 124 125 126 127 128 129 130 131 132 133
β_1 residues:
15 16 17 23 32 33 34 35 37 50 52 53 107 109 110 111 112 113 114 116
117 118 120 121 122 123 125 126 127 128 129 131 132 134 135 136

J	N(J)
75	20
76	184
77	1032
78	3942
79	10848
80	22180
81	34232
82	40081
83	35436
84	23292
85	11040
86	3568
87	704
88	64

An 88 residue domain with RMS = 0.240 Å
α_1 residues:
1 2 3 5 6 7 9 10 13 19 20 24 27 30 31 33 34 35 36 38
39 41 42 43 60 70 102 103 104 108 110 111 113 116 117 119 121 122 123 124
125 126 127 128 129 130 131 132 133
β_1 residues:
15 16 17 22 23 32 33 34 35 37 50 52 53 107 109 110 111 112 113 114
116 117 118 119 120 121 122 123 124 125 126 127 128 129 131 132 134 135 136

Fig. 4. The rigid core of the human hemoglobin $\alpha_1\beta_1$ dimer as found with ε = 0.50 Å and N_s = 200. One hundred and four residues have a sum S_i of at least 200. These residues form the set $U_{0.50}(200)$. Only 86 of the residues belonging to $U_{0.50}(200)$ are non-rigid with at most one other residue in $U_{0.50}(200)$. $D_{0.50}(200)$, a domain of 74 residues, is left after all 12 residues belonging to non-rigid pairs have been removed. Looking through the complement of $D_{0.50}(200)$, we find eight residues in addition to those of non-rigid pairs that are rigid with $D_{0.50}(200)$. Only one of the largest domains is shown at the bottom of the figure along with the RMS value for its oxy-deoxy superposition.

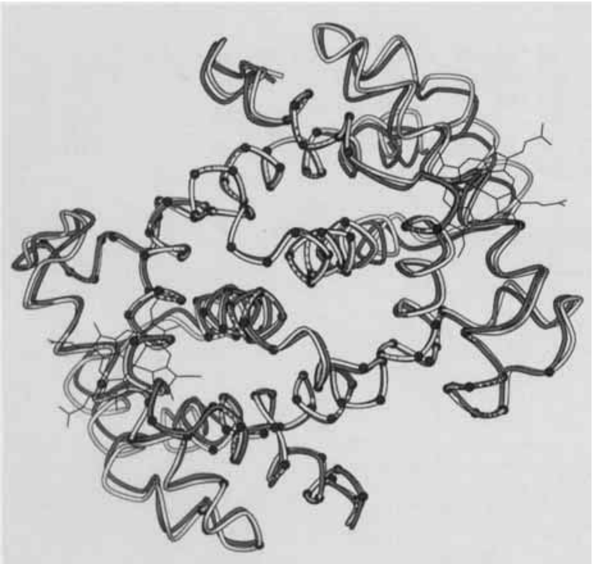

Fig. 5. Superposition of the deoxy and oxy $\alpha_1\beta_1$ dimers of human hemoglobin by aligning the residues of one of the rigid cores found with ε = 0.75 Å. The view is down the x-axis toward the dimer-dimer interface. The α_1 monomer is to the lower ledt, and the β_1 is to the upper right. Blue and green, the α carbon backbone of the deoxy conformation; red, the oxy conformation. About half of the same rigid core was found with visual methods by Baldwin and Chothia.[16] The RMS value for the oxy-deoxy superposition of this rigid core is 0.333 Å. Only deoxy hemes, colored brown, have been drawn in the figure. Both heme pockets clearly undergo considerable conformational change relative to the rigid core.

since there are $M(M-1)/2$ pairs of the M residues, all of which must have Δ less than ε. A total of $C(N,M)$ subsets of M residues exists in the set of N residues, so the expected number of subsets of M residues that define a domain is $C(N,M)$ times Eq. (7). We seek the largest value of M for a given ε for which we find at least one domain within the original set of N residues. Thus the largest M is the integer closest to the solution of the equation

$$C(N,M)\left(\int_0^{\varepsilon} W(\Delta)d\Delta\right)^{M(M-1)/2} = 1. \qquad (8)$$

Actually, for the hemoglobin dimer, we find that the data are fit much better if we assume that there are two disjoint subsets, one with m_1 points described by $W_1(\Delta)$ with standard deviation of σ_1 and the other with m_2 points described by $W_2(\Delta)$ with σ_2. Eq. (8) then generalizes to

$$C(N,M)\left(\int_0^{\varepsilon} W_1(\Delta)d\Delta\right)^{m_1(m_1-1)/2}$$

$$\left(\int_0^{\varepsilon} W_2(\Delta)d\Delta\right)^{m_2(m_2-1)/2} = 1 \qquad (9)$$

with $M = m_1 + m_2$.

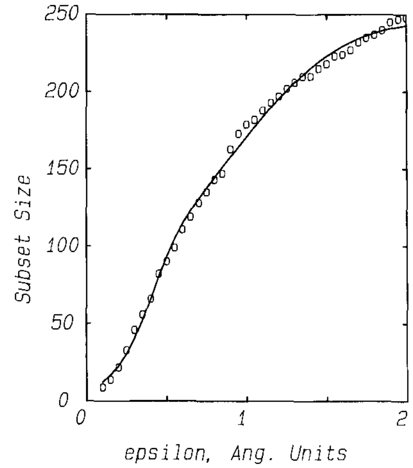

Fig. 6. The dependence of hemoglobin core size on ε. Circles mark the number of residues in a core domain for each ε as found from PDB atomic coordinates. The solid line is a best-fit of Eq. (9) with σ_1 = 0.20 Å and σ_2 = 0.86 Å to the measured points. We have assumed in this calculation that difference-distance matrix elements Δ_{ij} have a Gaussian distribution.

A best-fit curve of M versus ε obtained from the solution of Eq. (9) is shown in Figure 6. This curve was obtained by varying the parameters σ_1, σ_2, and m_1 until the sum of the squares of the deviations from the points was minimized. The values found for σ_1 and σ_2 were 0.20 Å and 0.86 Å, respectively, in the neighborhood of the experimental precision of about 0.5 Å. (The standard deviation of this curve from the points is 4.0, whereas the standard deviation found when attempting to fit with only one Gaussian subset was 25; the one-Gaussian fit was not satisfactory.) Thus even this crude theory of the dependence of the number of residues M in a domain on ε gives a reasonably good description of the observations.

In Figure 7 we show the rigid domains found with ε values of 0.25 Å (asterisks) and 0.50 Å (circles), while Figure 5 shows the 0.75 Å domain. The core structure appears to be well marked by the 0.50 Å circles. Increasing ε from 0.50 Å to 0.75 Å mainly picks up more residues in the same structure while extending the structure only slightly. Apparently the principal difference between the 0.50 and 0.75 cores is that the latter is more tolerant of errors in the data. This is in accord with Baldwin and Chothia's[16] estimate that differences between coordinates in their data were not significant unless they exceeded about 0.50 Å because of experimental uncertainty in the coordinates. This domain is definitely though sparsely marked in Figure 7 even by the 0.25 Å asterisks. Thus the identification of the gross structure of a rigid domain is not very sensitive to the value of ε for sufficiently large ε.

CONCLUSIONS

Proceeding from the premise that if rigid domains exist they should be important components of pro-

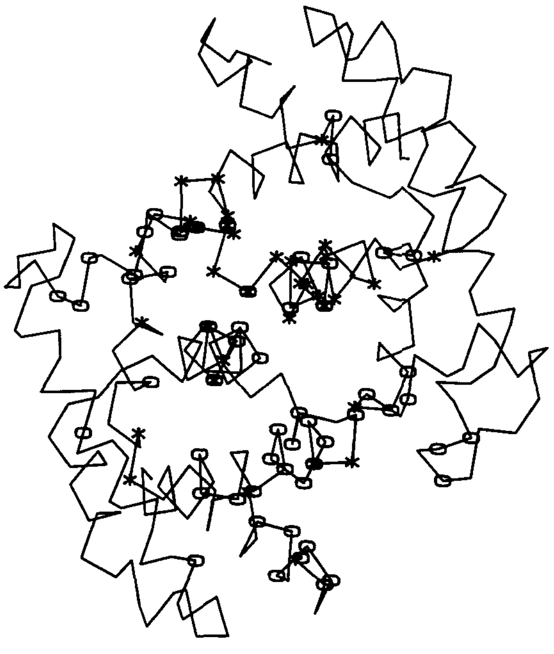

Fig. 7. A view down the x-axis of the hemoglobin dimer with residues in a rigid core for $\varepsilon = 0.25$ Å marked with asterisks and additional rigid core residues found with $\varepsilon = 0.50$ Å marked with circles. Apparently, the effect of increasing ε is to fill in secondary structures already defined by smaller values of ε.

tein structure, we have devised two methods for finding such domains, and have tried out these methods using subunits of hemoglobin. We have found that rigid domains occur in families, the members of which overlap extensively, that differ by only a few residues. The concept of a family of overlapping domains is an important generalization of the rigid-domain concept itself.

Nearly all the residues belonging to the hemoglobin dimer rigid core of Figure 5 are found within the A, B, C, G, or H helices of the α_1 and β_1 monomers. Similar structures have been noted before. Baldwin and Chothia[16] identified 68 residues that form an invariant set along the $\alpha_1\beta_1$ interface of the hemoglobin dimer. These residues are mostly the parts of the α and β B, G, and H helices and were used as a frame of reference by Baldwin and Chothia[16] from which to observe the tertiary and quaternary changes in hemoglobin. Except for residues $\beta_1 30$ and $\beta_1 31$, which are within the interior of the β_1 B helix, and residue $\beta_1 54$, a valine residue in the β_1 D helix, all are included in our family of 16 rigid core domains with $\varepsilon = 0.75$ Å. Baldwin and Chothia[16] noted as well that the α B, α C, α G, and α H helices and the β B, β D, β G, and β H helices together, except for the first few residues of the G helices and the last few residues of the H helices, remain fairly invariant between the T and R states of hemoglobin. For larger values of ε we find rigid core domains that include most of these helices but also many residues in both the α and β A helices and in the C helix of β as well. That the A, G, and H helices form a protected folding unit in apo-myoglobin has been noted by Hughson et al.[22]

The rigid core is not the only domain that can be found in the hemoglobin dimer. By removing the rigid core residues from the dimer structure and searching the remainder we can find several other smaller, independent domains associated with the heme molecules. We expect to describe these in another paper in preparation.

The primary contribution of this paper is a method to determine conserved spatial relationships. As such, it is directly applicable to analysis of complex conformational changes in proteins. Allostery is one such case; there are others, such as the calcium-triggered change in calmodulin, or the rearrangement of the hemagglutinin of influenza virus.

We have thought about the application to finding conserved cores in homologous proteins, However, that application requires substantial further development. We can calculate conserved structure given a sequence alignment, but finding the best sequence alignment for identifying conservation of structure is a another problem. The discussion of sequence alignment would take us far outside the scope of this paper.

ACKNOWLEDGMENTS

Support to the authors includes NIH 1 PO1 HL48018 to L. Ten Eyck et al., GM29458 to G. Rose, and GM11916 to B. Zimm, who also acknowledges a grant from the Wyatt Technology Corporation. This work used facilities of the San Diego Supercomputer Center under grant ASC-8902827 from the National Science Foundation. We thank Gary Ackers for helpful discussions.

REFERENCES

1. Wetlaufer, D.B. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Natl. Acad. Sci. USA 70:697–701, 1973.
2. Rossman, M.G., Liljas, A. Recognition of structural domains in globular proteins. J. Mol. Biol. 85:177–181, 1974.
3. Crippen, G.M. The tree structural organization of proteins. J. Mol. Biol. 126:315–332, 1978.
4. Rose, G.D. Hierarchic organization of domains in globular proteins. J. Mol. Biol. 134:447–470, 1979.
5. Wodak, S.J., Janin, J. Location of structural domains in proteins. Biochemistry 20:6544–6552, 1981.
6. Rashin, A.A. Location of domains in globular proteins. Nature 291:85–87, 1981.
7. Lesk, A.M., Rose, G.D. Folding units in globular proteins. Proc. Natl. Acad. Sci. USA 78:4303–4308, 1981.
8. Levitt, M., Sander, C., Stern, P.S. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. 181:423–447, 1985.
9. Zehfus, M.H., Rose, G.D. Compact units in *proteins. Biochemistry* 25:5759–5765, 1986.
10. Fermi, G., Perutz, M.F., Shaanan, B., Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. J. Mol. Biol. 175:159–174, 1984.
11. Shaanan, B. Structure of human oxyhaemoglobin at 2.1 Å resolution. J. Mol. Biol. 171:31–59, 1983.

12. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.

13. Vriend, G., Sander, C. Detection of common three-dimensional substructures in proteins. Proteins 11:52–58, 1991.

14. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233:123–138, 1993.

15. Dickerson, R.E., Geis, I. "Hemoglobin: Structure, Function, Evolution, and Pathology." Menlo Park, CA: Benjamin/Cummings Publishing Company, 1983.

16. Baldwin, J., Chothia, C. Haemoglobin: The structural changes related to ligand binding and its allosteric mechanism. J. Mol. Biol. 129:175–220, 1979.

17. Frauenfelder, H., Hartmann, H., Karplus, M., Kuntz, I.D., Jr., Kuriyan, J., Parak, F., Petsko, G.A., Ringe, D., Tilton, R.F., Jr., Connolly, M.L., Max, N. Thermal expansion of a protein. Biochemistry 26:254–261, 1987.

18. Richards, F.M., Kundrot, C.E. Identification of structural motifs from protein coordinate data: Secondary structure and first-level super secondary structure. Proteins 3:71–84, 1988.

19. Bystroff, C., Kraut, J. Crystal structure of unliganded *Escherichia coli* dihydrofolate reductase. Ligand-induced conformational changes and cooperativity in binding. Biochemistry 30:2227–2239, 1991.

20. Kabsch, W. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A32:922–923, 1976.

21. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A34: 827–828, 1978.

22. Hughson, F.M., Wright, P.E., Baldwin, R.L. Structural characterization of a partly folded apomyoglobin intermediate. Science 249:1544–1548, 1990.