Applications of Empirical Amino Acid Potential Functions

R. L. JERNIGAN,[1] L. YOUNG,[1,2] D. G. COVELL[2], and S. MIYAZAWA[1,3]
[1]*Section on Molecular Structure, Laboratory of Mathematical Biology, National Cancer Institute, NIH, Bethesda, MD 20892, U.S.A.*
[2]*PRI-Dyncorp, Biomedical Supercomputing Laboratory, Bldg. 430, Frederick, MD 21701, U.S.A.*
[3]*Gunma University, Faculty of Technology, 1-5-1- Tenjin, Kyuryu, Gunma 376, Japan*

Abstract. Many characteristics of protein folds can be described at less than atomic detail. We have been pursuing a reductionist approach to represent proteins in which residues, rather than atoms, are considered as single points. Direct averages of protein crystal structures provide empirical potential functions that describe residue–residue interactions. These functions have averaged away many details but include essential features with attractive and repulsive terms, that are overall akin to a pair-wise hydrophobicity. Most computational evaluations of protein structures are deficient in not considering a broad enough range of possible conformations; this less detailed approach makes possible a more thorough consideration of conformations. We are developing a new lattice approach to generalize and improve the efficient generation of folded protein conformations, and these empirical potentials are an important part of this approach. Applications of these potential functions presented here include the location of binding sites for peptides on proteins based on geometry and hydrophobicity, specification of similarities of sequence substitutions, and investigations on the variability of the hydrophobicity of a given type residue averaged over occurrences in individual proteins. Recent progress in these areas will be described. This new approach is leading us to the ability to consider more complete sets of protein conformations.

Key words. Protein binding, empirical potential energies, amino acid substitutions, protein fold assessment, hydrophobicity.

Introduction

One major goal of structural biology is to understand how molecules recognize and interact with one another. Comprehension of the important interactions could enhance the understanding of numerous biological processes, as well as provide a sound basis for drug design. For example, it would lead to more rational design of improved inhibitors or substrates for enzymes because it would permit better use of the remarkable tools of site-directed mutagenesis and other gene splicing techniques. Theory can play a role in providing a deeper understanding to aid in developing new approaches.

An eventual goal is the direct calculation of active, stable macromolecular conformations, from their sequences. Calculations precise enough to place many thousands of atoms at their best positions are not feasible at present. There have been numerous demonstrations that the full atomic details of even small proteins cannot be calculated directly. The development of successful higher order approaches to molecular structure is essential before we can achieve a complete understanding of all the complexities of biological macromolecules themselves, as well as their interactions with other molecules and their assembly into biological structures. Even today, it is difficult to foresee the ability to determine rapidly the

A. Pullman et al. (eds.), Modelling of Biomolecular Structures and Mechanisms, 151–166.

most stable conformation of the largest protein, by calculating interactions among all atoms. Specifying all details of protein self-assembly or their assembly into larger structures would be even more difficult.

The approach that we have been pursuing is to develop ways to treat a more complete set of conformations, but with less detail, with the intention of obtaining approximately correct overall folds. Following this it is possible, as has been demonstrated, to complete the atomic details by refinement of the overall trajectory of the protein backbone. This approach has been successfully demonstrated, and the tactic has also been taken up by other research groups. The main tool in this process is a set of empirical residue–residue potential functions derived directly from structural data. In our case [1] these potentials are derived simply by counting in crystal structures the number of close approaches of all pairs of residues of different types.

Here we are focusing on several applications of these residue–residue potential functions that had been derived previously. These concepts have been applied to locate peptide binding sites on the surfaces of proteins [2], to develop a matrix relating the ease of amino acid substitutions [3], to evaluate the quality of non-native folds [4], and to develop some other simple models of proteins to answer questions such as: Is the observed native state identical to the lowest energy conformation [5]? How easy is it to make large scale transitions in proteins [6]? Below, we will summarize and discuss in detail the peptide–protein binding site results. Successes with all of these applications are encouraging us to pursue some further improvements to these residue–residue potential functions. More detailed residue–residue mean field interaction potentials are going to be developed, to include details related to the dense packing of residues. Further important factors to include in such potential functions are near neighbor interactions within short chain segments and an overall effect caused by the constraint of compactness which has recently been investigated [5]. It was found that the longer the separation in sequence, the more favorable the interaction because of the compactness.

Previously, in treating globular protein conformations [4], the protein had been confined to a space exactly the size and shape of the known native conformation, and all conformations were generated in that space. It was remarkable that, by this enforced constraint it became possible to generate thousands of highly varied conformations, in the approximation of one point per residue on a lattice. These were evaluated with the same residue–residue interaction potentials mentioned above. By applying the interaction values to all of the compact conformations enumerated for several small proteins, the native conformation was always found among the best few per cent of all the forms, and by use of these potentials most incorrect folds could be discarded.

Interactions in known protein–ligand and protein structures have been analyzed extensively to develop a tool to aid inhibitor design. We have found that hydrophobicity is the substantial determinant of binding strength, and hence the location of the binding site, in a verification of the biochemist's intuitive view of 'sticky patches'.

I. Residue–Residue Potentials

The aim here is to discuss a range of applications that demonstrate the utility of residue–residue potentials and to outline new improvements to these potentials.

METHOD

Our previously derived residue–residue potential functions were obtained from the frequencies of non-bonded proximate residues in crystal structures, on the assumption that these placements, on average, must reflect their effective energies of interaction. (See Reference 1 for the details of the method.) The basic underlying assumption, in treating the interactions between protein residues in this way, is that the close residues are those with the strongest interactions, and that we can ignore, to good approximation, longer distance interactions. The values clearly indicate the importance of hydrophobic interactions, as well as specific preferences for charged pairs of opposite sign. They indicate that the strongest interactions take place among the hydrophobic types of residues, and that the most specific, but weaker, interactions are between polar residues. See Figure 1.

There is extensive experience in the field of polymers for treating interaction energies among groups of atoms as parameters to be evaluated from experimental data on small analogous molecules or even by fitting physical measurements of large macromolecules [7]. (As an example, see Reference 8.) Such approaches were utilized in developing the energy parameters for diverse polymer repeat units with rotational isomeric state models. Part of the rationale for treating these interaction terms as adjustable parameters is that the relative populations of conformers can change, depending on the experimental conditions. For example, studies by Mizushima [9] demonstrated changes in the relative likeliness of isomeric states for different conditions; these results suggest the approach of adjusting potential functions to mimic the specific conditions. In the case of proteins, it is possible to use crystal structures directly to develop relative interaction preferences based upon the observed frequencies of occurrences of residue–residue contact pairs. The direct approach taken in extracting such functions from crystal structures is described in detail in Reference 1. Previously Tanaka and Scheraga [10] had performed a similar averaging of structural data. Subsequently, others including Gregoret and Cohen [11], Sippl [12] and Hinds and Levitt [13] have made similar analyses. Our residue–residue interaction energies resemble a pair-wise hydrophobicity index with the reference state being the residues individually in contact with water. The most important features of the interactions can be obtained only if averaging of the data is performed, to remove errors in the structures as well as to remove individual aberrant cases. Use of all of the structural data is less likely to yield ultimate success than using results after judicious averaging. We have also looked at such interactions in more general ways [5] as a uniform homogeneous term, the same for all interactions, and have probed individual positions to see where changes in interaction strengths would most easily cause changes in conformation [6].

These potential functions have numerous uses for quickly assessing conformational quality. They are useful to select sets of good conformations from large

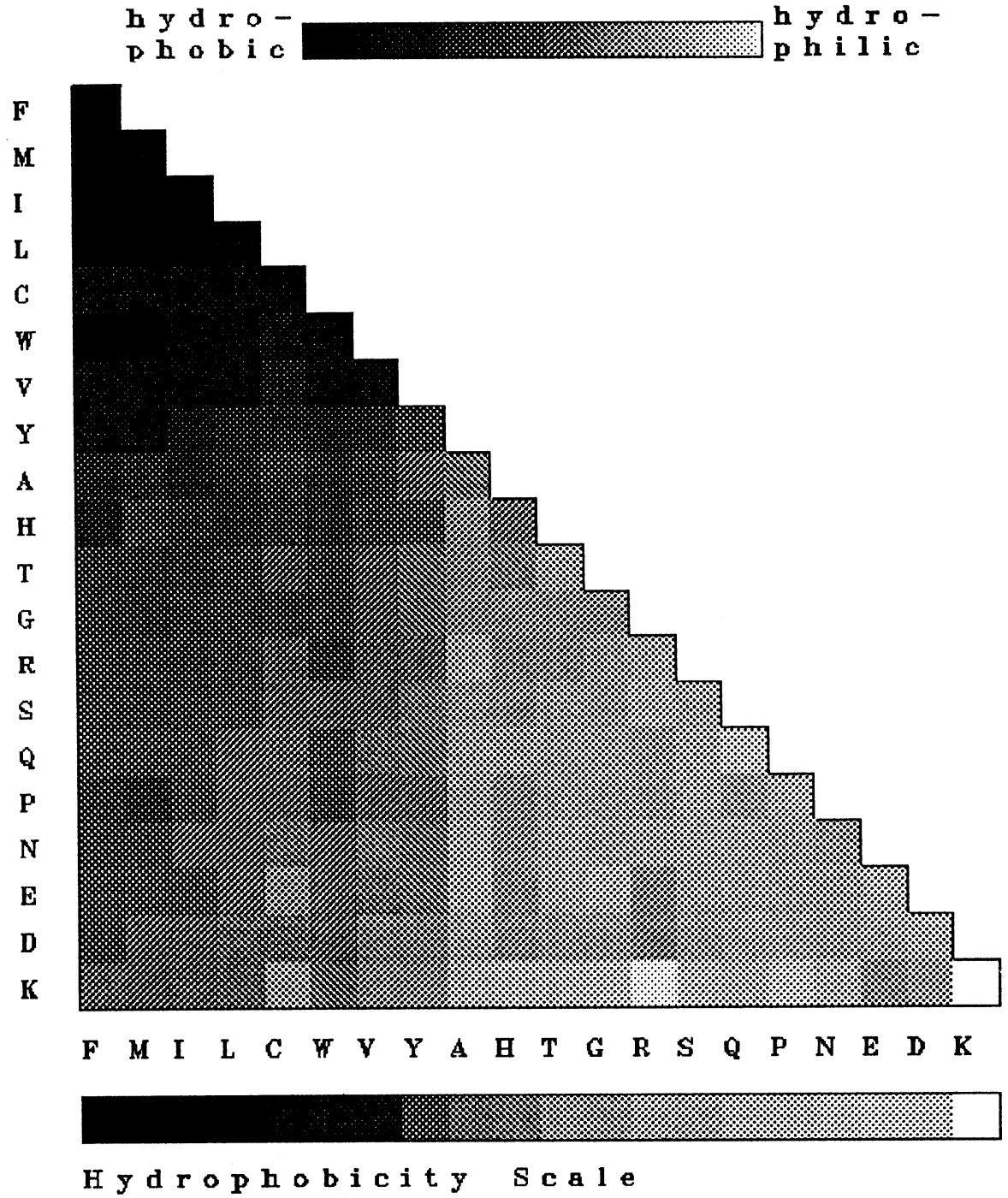

Fig. 1. The residue–residue interaction energies from Reference 1; residues are designated by one letter codes. The darker squares are the lowest interaction energies and the lighter square the highest.

sets of conformations [4], and to derive amino acid substitution matrices for use in sequence comparisons [3]. Another use of the potentials is to evaluate the effects of amino acid substitutions on stability [14]. An important further area of application of potential functions is for the evaluation of different sequences threaded onto known structures [15,16]. In a similar way, hydrophobicities can be used to locate peptide binding sites on protein surfaces (see Section II below). The successes achieved with all of these applications encourage us to pursue

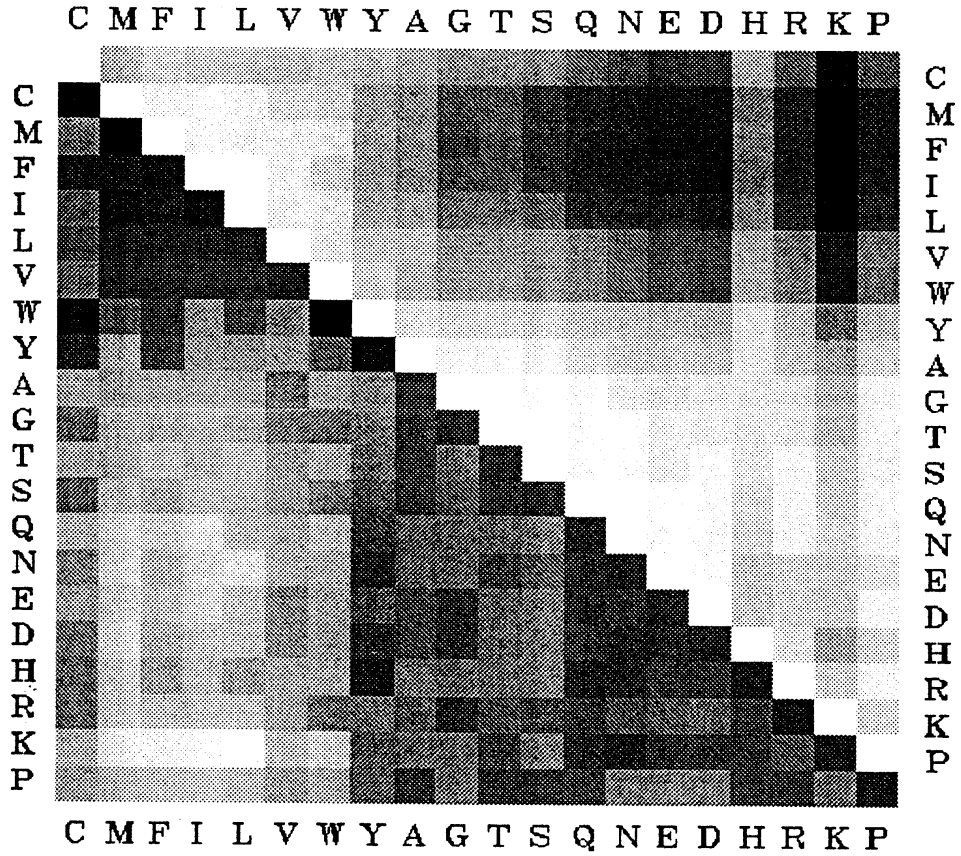

Fig. 2. The change in energy upon exchange of two types of residues is shown in the upper diagonal half. White corresponds to the most favorable and black to the most disfavored exchanges. Generally, exchanges are more favored within the same class of hydrophobic or hydrophilic residues. The lower diagonal half displays the log relatedness odds matrix corresponding most closely to the PAM 250 matrix. Most similar substitutions are black, intermediate ones gray, and least similar ones white. The pair contact energies [1] and codon frequencies were utilized in its derivation [3].

several additional improvements. To this end, we are beginning the development of a newer generation of residue–residue potential functions.

Results

Amino Acid Substitutions

An amino acid substitution matrix to indicate the relative conservativeness of a substitution was derived. The bases for this table are the set of residue–residue potentials derived from known structures and the probabilities for amino acid codon substitutions. These structurally derived results exhibit a strong correlation with the Schwartz and Dayhoff numbers based on sequences of protein families. Their results have found extensive application for finding similarities among protein sequences. Our results from Reference 3 are shown in Figure 2.

Effects of Compactness

An additional aspect that has been pursued is the development of the long range components of residue–residue potential functions arising simply from the constraints of compactness [5]. The result of compactness is that the intrinsic interactions between residues become monotonically more favorable with an increase in their sequence separation. Hence, the largest effects of compactness are found to favor interactions between the two chain ends, in agreement with observations on known protein structures [4].

CONCLUSIONS

1. The most feasible amino acid substitutions are among hydrophobic residues and among hydrophilic residues. Changes between these two classes are less acceptable.
2. Compactness causes the frequencies of interactions to increase directly with increase in sequence separation. Consequently, interactions between the chain ends are most favored.

SECOND GENERATION OF RESIDUE–RESIDUE POTENTIALS

Since our original derivation of residue–residue potentials [1], the number of diverse protein crystal structures has increased rapidly. One problem with including all known structures is that many of them are closely related to one another. For this purpose, we have devised a weighting scheme that gives greater weights to structures with the most diverse sequences. In 1985 we had simply selected 42 diverse high quality structures for our data collection; now we have used, together with this weighting scheme, 1168 separate protein structures with a total of 1661 subunit structures, each of which is longer than 49 residues and has a stated resolution of 2.5 Å or better. This includes more than 54 000 effective residues, an increase of about 6-fold over the earlier data collection. A comparison is shown in Figure 3 between some of the old and new data that has been collected.

One modification to the potential functions will be the addition of a repulsive term to the previous attractive term. Again, the repulsive part of the residue–residue potential is being derived statistically from known structures in the X-ray crystal database. This additional term will complement the previously derived attractive component that was found to be effective in determining proper or improper chain folds. Results with this previous attractive potential were often too compact [17]. This new version should improve evaluations of residue packing in proposed structures. Other modifications will include some of the short range interactions which previously were ignored. These are collected from average side chain positions that are sequence neighbors, including their virtual bond lengths, virtual bond angles and virtual torsion angles. The sequence dependence of such terms should reflect the effects of solvent and the interior environment of the protein better than have previous potentials.

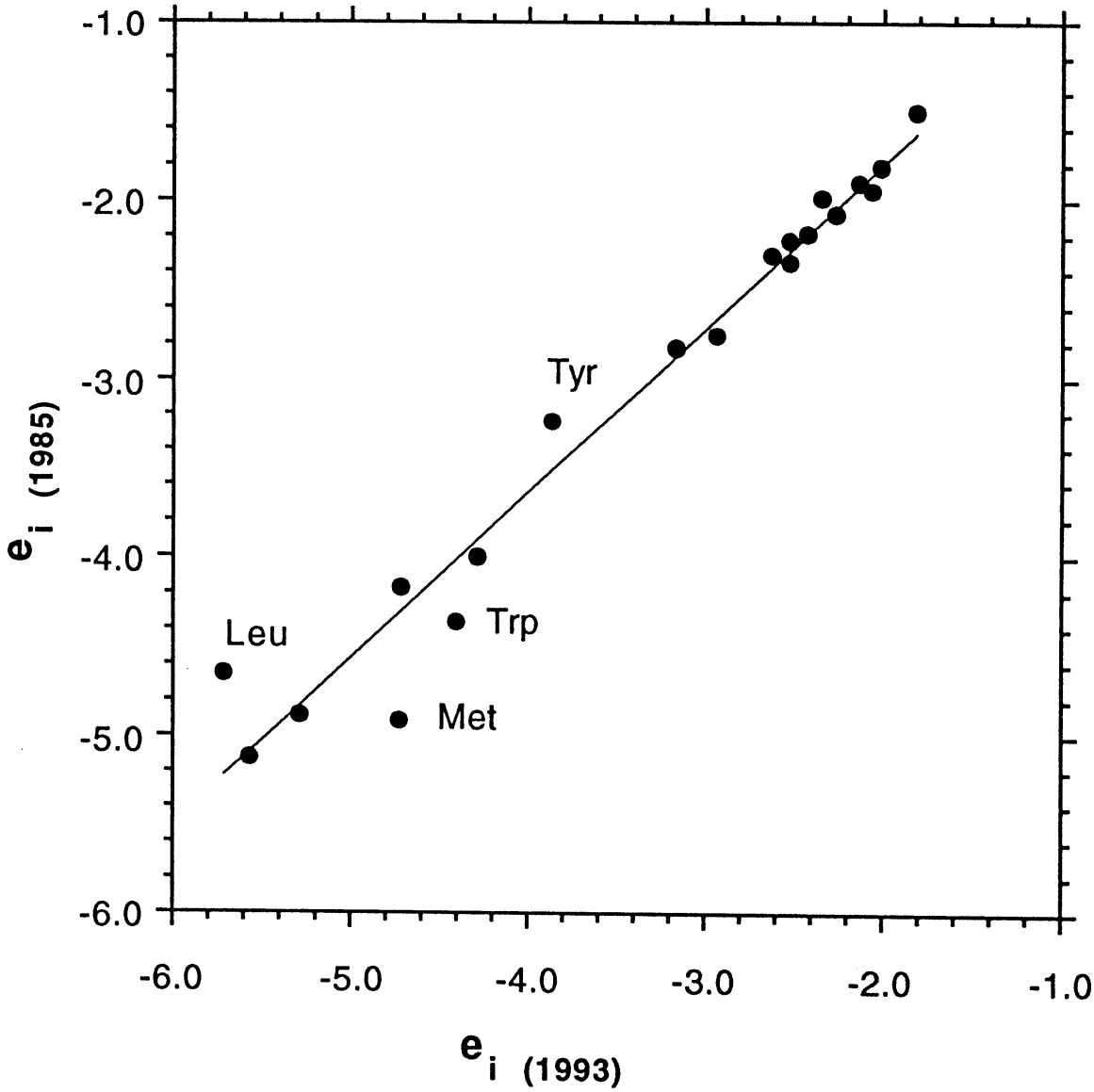

Fig. 3. Comparison between the hydrophobicity scale derived from 42 protein structures ('85) and the new ones derived from 1661 different protein subunits, by using sequence similarity weights ('93). This scale is calculated by Equation (1) from the pairwise contact energies. There are some expected changes; for instance, the most infrequent residues Trp and Met have substantial changes, presumably because of the increased amount of data. Substantial changes are also observed in the values for Tyr and Leu. The latter is especially surprising since all of the other high frequency residues exhibit only small changes.

DISCUSSION

One of the questions to investigate is the extent of burial or exposure of hydrophobic and polar residue on a local basis. This could be investigated initially, for instance, by determining whether or not polar residue pairs close in sequence are physically closer than polar-hydrophobic pairs. Biased conformations derived from such an investigation could be used statistically as conformational tiles in an overall conformation generation. These considerations of short range interactions differ

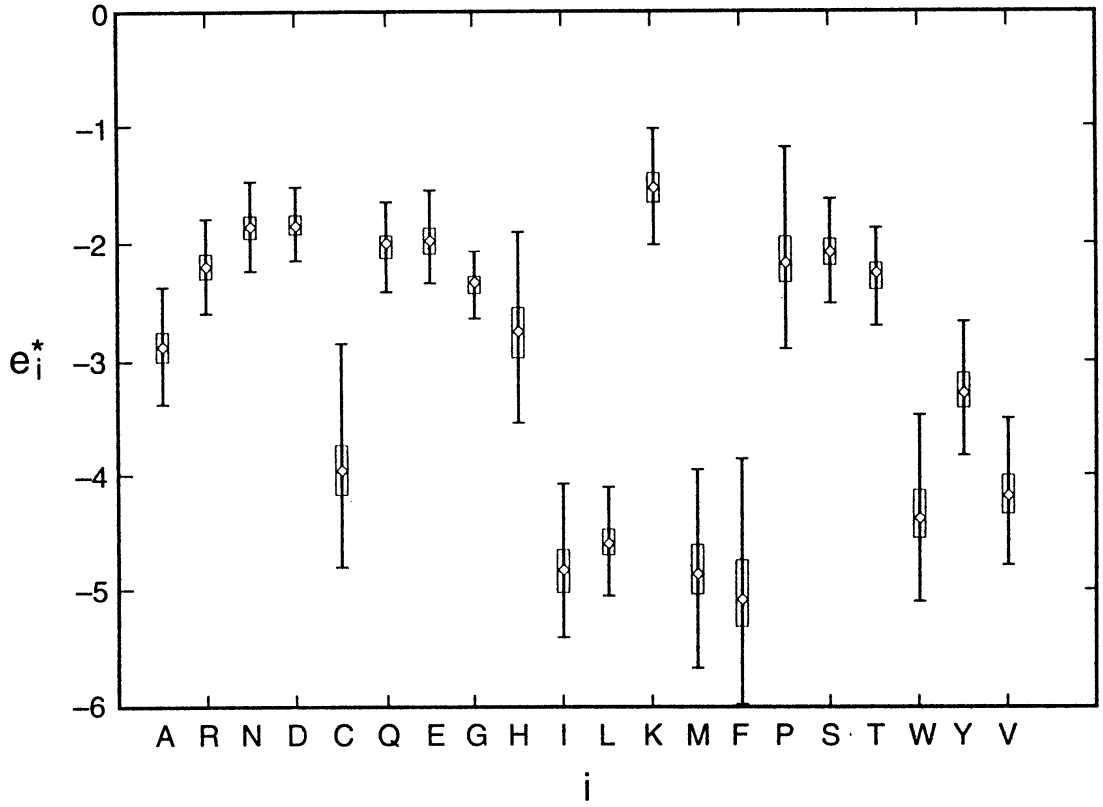

Fig. 4. Variability of environmental interaction energies. These were calculated from the e_{ij} values in Reference 1 by summing over all neighboring residues within 6.5 Å in 709 protein subunit crystal structures. Means of the values for each protein are given as central points; half of the occurrences are within the boxes and 99.8% of all occurrences are within the range of the bar.

substantially from previous studies where short range interactions were usually calculated only in vacuo between nearest neighbor residues.

The regularity of the positions of residues around a central residue is striking [1]. The coordination number is not so variable, and also the hard core radii do not vary so widely. And, as we will see, the environment of a given residue type is remarkably constant.

HOW CONSTANT ARE ENVIRONMENTS OF A GIVEN RESIDUE TYPE?

In an attempt to investigate this question, we undertook the following data collection: For each of 709 crystal subunits, we have calculated the average interaction energy per interacting pair. In Figure 4 is shown the distribution of these values for the set of proteins for each type of residue.

$$e_i^* \text{ (each protein)} = \sum_j n_{ij} e_{ij} \Big/ \sum_j n_{ij}$$

where n_{ij} is the number of i-type, j-type pairs having the mean of their side chain atoms within spheres of 6.5 Å radius for one protein. The nearest sequence neighbors have been excluded in the counts.

The values reflect the hydrophobicities; the polar residue types have weaker

interactions and usually lie in a band between a value between −1.5 and −2.5 for e_i^*. The non-polar types have lower values, typically between −4 and −5. Several residue types fall between these two ranges, notably alanine, cysteine, histidine, proline and tyrosine.

The fascinating part of this result is that, even over the wide range of proteins that are included in the sample, there are clearly separable regions of total hydrophobicity of the environments for all occurrences of a residue type. Polar residue types have weaker total interactions but usually lie within a relatively narrow band of values. The non-polar types have lower values, but larger variabilities. This smaller variability of the hydrophobic environment for polar residues indicates that they would be particularly good residues for specifying their most probable location on the protein globule surface.

Another important class of interactions is between ions and proteins. Understanding their specific roles in stabilizing conformations and the sensitivities of conformation to the coordination with different ions is important. We have begun work on this problem by investigating the relative stabilities of K^+ and Na^+ ions in four stranded DNA structures. The stabilizing effects of Zn^{2+} in some protein structures and the regulating effect of Ca^{2+} on protein function are exciting problems to pursue [18,19]. Tools to place such ions, based on their coordination geometries, could be developed and included in lattice simulations. Treating the relative stabilities of protein or nucleic acid ion coordinated structures may, however, require better treatments of solvent and electric calculations. A better understanding of specific ion effects in protein structures ought to make possible the addition of ion coordination complexes as important construction elements in protein design.

II. Using Hydrophobicities to Locate Peptide Binding Sites on Proteins

The aim here is to develop a method to scan rapidly the surface of a protein of known structure to locate probable peptide or protein binding sites. The hypothesis is that it ought to be possible to select target binding sites on a macromolecule on the basis of its surface geometry and potential interactions, especially those of a hydrophobic nature.

METHOD

A new method to locate the most probable peptide binding sites on protein surfaces is presented. This simple method permits rapid screening of molecular surfaces for sites that would interact strongly with any hypothetical peptide. Favorable binding positions exterior to the surface of the target protein are derived on the basis of the total hydrophobicity of neighboring protein residues. When tested on 23 available peptide-protein X-ray structures, the observed sites always fall among the best 0.7% of all possible binding sites, with one exception. *These results strongly support the view that the hydrophobic interactions are responsible for the strength of protein–protein association and that the hydrophobic surface clusters can be used to choose small sets of surface loci for ligand attachment.*

Previous methods to locate important surface targets that include both active and non-active sites have typically focused on identifying regions with high densities of

hydrogen bond sites [20,21], regions with desirable electrostatic properties [22], regions with appropriate features of surface curvature [23], and regions of high surface complementarity [24].

We fit the C^α positions of the target protein to a regular lattice and use a set of lattice points immediately exterior to the protein to define the binding space. Covell and Jernigan [4] found that the lattice providing the best fits of virtual bond and torsion angle geometry in the C^α backbone was the face-centered cubic lattice with unit edge 3.8 Å, corresponding to the fixed virtual bond distance between neighboring $C^{\alpha'}$s. With this lattice, an accurate model for the C^α positions of the target molecule is obtained, with overall fits of about 1 Å RMS deviation from the crystallographic protein structure. We simply extend the same lattice used to define the C^α backbone into the region surrounding the protein [25]. These lattice points define a region exterior to the protein surface (between 6.1 and 9.0 Å) that is appropriate for placement of an inhibitor peptide. The scheme in Figure 5 describes the geometry and shows how the clusters of points are developed. The size of the cluster of points reflects the concavity of the surface.

The hydrophobicity of a given protein cluster is taken to be simply the sum of the hydrophobicities of its constituent residues. The hydrophobicity scale used is based upon the contact energies calculated by Miyazawa and Jernigan [1] given at the bottom of Figure 1. This statistical study of non-bonded residue contact frequencies in protein structures in the Brookhaven Protein Data Bank (PDB) [26,27] investigated the local neighborhoods of the 20 residue types in 42 high resolution crystal structures and used a quasi-chemical model of the pairwise interactions between the different types of amino acids and solvent to calculate a set of residue–residue contact energies e_{ij}. These pairwise contact energies were derived for a model using a single point representation of the amino acids in which the point is placed at the center of side chain atom positions and all interactions within a distance of 6.5 Å, are counted [1].

Averages of these pairwise contact energies are used as the hydrophobicities for each residue type, i:

$$e_i = \frac{\sum_{j=1}^{20} e_{ij} n_{ij}}{\sum_{j=1}^{20} n_{ij}} \tag{1}$$

where the contact energy e_{ij} is the Miyazawa and Jernigan energy difference between an ij amino acid pair. i and j are each one of the standard 20 types and solvent. n_{ij} is the number of ij contacts in the set of structures used to calculate e_{ij}. These contact energies can be understood in terms of the hydrophobic-hydrophilic designations of amino acids and the pairings that contribute most to protein stability:

strongest	hydrophobic–hydrophobic
intermediate	hydrophobic–hydrophilic
weakest	hydrophilic–hydrophilic

This hydrophobicity scale in Equation (1) shows a strong correlation with the Nozaki–Tanford scale [28] and others, as shown by Cornette *et al.* [29]. These hydrophobicities are used to define the energy for a protein cluster which is then

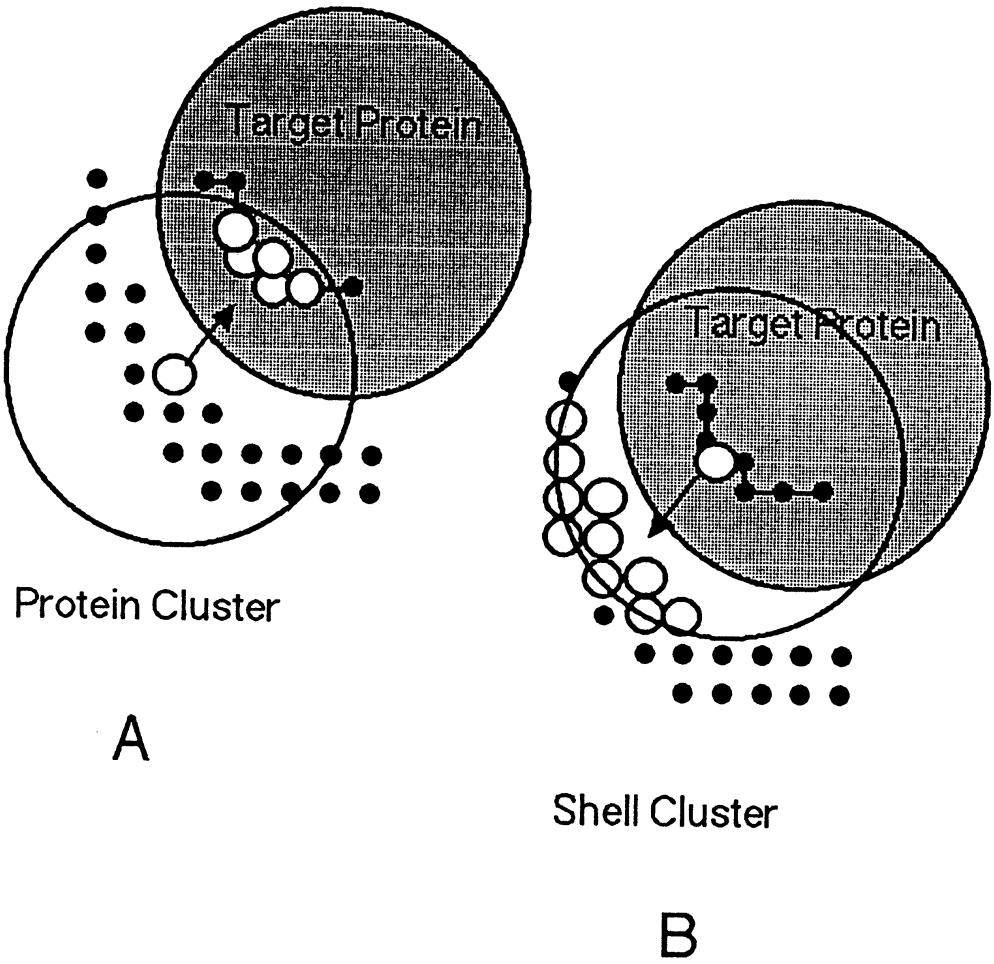

Fig. 5. Two stage method for defining clusters of protein residues associated with a single exterior binding point (A) and the associated cluster of binding points centered about this point (B). The connected dots represent a portion of the protein under investigation and the unconnected dots the external potential binding points. The open circle with attached arrow in A is an exterior point which defines a cluster of protein residues, the small open circles located within the large circle centered upon the exterior point. In B each of the designated protein points is used together to define the cluster of exterior points shown here for one protein point.

$$E_{\text{cluster}} = \sum_{k=1}^{m} e_k$$

where m is the number of residues in the cluster.

PREDICTION OF BINDING SITE

The lattice points exterior to the protein are limited to a shell of points. Protein clusters are defined by centering a sphere of 9 Å radius on each of these shell points (Figure 5A); the number of lattice points composing the shell is usually a few thousand. The protein clusters are then ranked according to their cluster hydrophobicity as explained above. Clusters of binding points exterior to each protein cluster are then identified (Figure 5B).

Accessible surface areas are used to compare the results of the calculation with

TABLE I
Examples of best calculated surface protein clusters

	% rank	Overlaps	
		M_1 with respect to	M_2
		X-ray	Calculated
Protein/ligand			
Trypsin serine protease			
Anhydro-trypsin/pancreatic trypsin inhibitor	0.40	74.7	65.5
β-trypsin/pancreatic trypsin inhibitor	0.40	70.8	68.4
Trypsinogen/porcine pancreatic trypsin inhibitor	0.12	49.8	64.0
Trypsinogen/pancreatic trypsin inhibitor	0.55	72.2	68.1
Serine proteinase/potato inhibitor	0.13	83.7	55.9
Kallikrein A/bovine pancreatic trypsin inhibitor	0.17	60.6	44.6
Chymotrypsin serine protease			
α-chymotrypsin/turkey ovomucoid third domain	0.06	53.7	33.3

experiment. Actual and predicted binding sites are defined as the accessible surface area of the protein buried by the ligand (A_{actual}) and binding cluster ($A_{\text{predicted}}$), respectively. A comparison of the two should account not only of the extent of coincidence, but also for their total size, to account for any overprediction.

The best overlap between these two surface areas is used to identify which of the binding clusters corresponds most closely to the actual binding site. Ideally, the calculation would yield a predicted binding site exactly the same size as the actual binding site, with the two in perfect register. Two measures M_1 and M_2 of percentage overlaps have been used:

$$M_1 = \frac{A_{\text{predict}} \cap A_{\text{actual}} \times 100}{A_{\text{actual}}}$$

$$M_2 = \frac{A_{\text{predict}} \cap A_{\text{actual}} \times 100}{A_{\text{predict}}}$$

The set theoretic operator $\cap$ indicates the overlap between the two areas on each side of it. The larger the measure M_1, the better is the coverage of the actual binding site by the calculated cluster. These measures of percentage overlap for the two surfaces – actual and predicted – together indicate the relative sizes of the predicted and actual binding sites. The ideal is 100% for both M_1 and M_2, corresponding to the situation where the exterior cluster buries precisely the same protein surface as the ligand in the known structure.

RESULTS

Table I shows for some examples the percentile ranking of the surface cluster most similar to the actual structure as well as the measures M_1 and M_2. For 23 protein–peptide complexes in the protein data bank, the binding sites of 22 com-

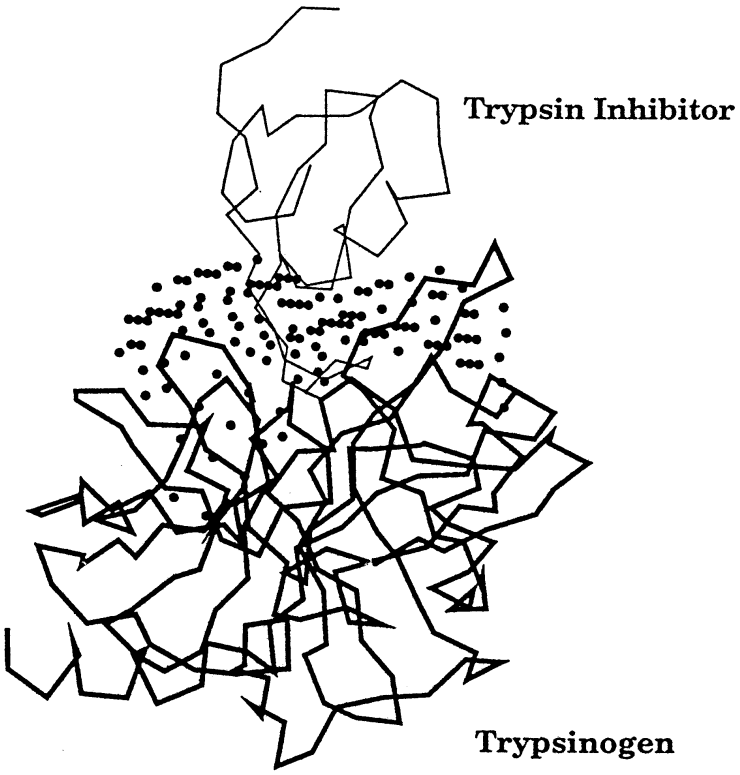

Fig. 6. Example of a strongly hydrophobic protein binding site for the trypsinogen/trypsin inhibitor complex. The trypsinogen is shown in thick lines and the inhibitor in thin lines. The dots indicate the second-most hydrophobic cluster. Note its good coincidence with the actual binding surface.

plexes that are most similar to the actual sites are found within the top 0.7% of the predicted sites. For example, examination of the protein clusters for 1TPA, the first on the list in Table I, indicates that the seventh ranked site corresponds to the actual site of the ligand. For α-chymotrypsin/turkey ovomucoid third domain 1CHO, the actual site is the top ranking predicted site. These results represent a 2 to 3 order of magnitude reduction in the numbers of surface locations necessary to be examined before finding the native binding target. Usually the known binding site is found among the best ten or so calculated sites. Figure 6 shows one example of the correspondence between most favorable binding sites on the receptor and ligand and the actual position of the inhibitor. The number of lattice points contained in the binding sites defined by the clusters ranked for the 22 structures, ranges from 40 to 179, with an average of 96.

The two rightmost columns of Table I present the overlaps of the actual and predicted binding sites. In all cases, the predicted binding sites substantially overlap the actual binding sites. More complete comparisons show that the portion of the actual binding site included in the overlap ranges from 45.6 to 99.7%, with an average of 67.9%. This indicates that around two-thirds of the actual ligand binding site are found in the predicted site. The portions of the predicted binding sites included in the actual site range from 3.7 to 72.3% with an average of 44.5%. These results indicate that the predicted sites tend to cover a larger surface than the actual site. The extent of this overprediction does represent, however, only a

small fraction of the protein's total accessible surface. Calculating the size of the binding sites as a percentage of the total protein surface area yields an average of only 9.2% for the predicted binding sites, the largest being 19.0% and the smallest, 4.6%. Similar results are obtained for the actual binding sites, the average size being 5.8% of the receptor surface area with values ranging from 0.92 to 14.0%. There is only one anomalous case: the result for one of the lysozyme/antibody structures HyHEL-5 (2HFL) is quite poor apparently because of the strong electrostatic character of its interactions [30].

OTHER STRUCTURES

It is also possible to apply the method to cases of a known protein structure but unknown ligand–protein structure if there are experiments that can indicate which are the binding residues. Useful data could consist of binding studies of proteins with modified sequences studied with the intention of mapping out the ligand binding site.

We have considered two cases where proteins are known to form complexes with peptides or other proteins, the human histocompatibility antigen HLA-2 (3HLA) (class I encoded in the MHC complex) and the fragment CD4 (2CD4) receptor. Both of these molecules are of considerable therapeutic interest as potential immunological targets. The analysis of HLA considered the α_1 and α_2 domains (residues 1 to 182), which are the two alpha helices above a beta sheet that form a cleft for the expected binding site [31]. The analysis of CD4 considered the N-terminal fragment comprising two domains (V1, V2) of the four predicted immunoglobulin-like extracellular domains [32,33]. This portion of CD4 is reported to be a receptor for the HIV gp120 fragment of he HIV coat protein.

For HLA, 1773 shell points surround the lattice model of the antigen binding domains referred to as $\alpha_1\alpha_2$. Based on the results for crystal complexes, the best 0.7% of the ranked $\alpha_1\alpha_2$ clusters are examined. These twelve clusters as ranked by their hydrophobicities bury 38.6% of the available surface area of $\alpha_1\alpha_2$. The strongest clusters are located in the cleft formed by $\alpha_1\alpha_2$ and bear a resemblance to the extra electron density found in the crystal structure [31]. The cleft itself is convexly curved, and fewer template points are found near its center. Hence there appear to be anchoring residues at the ends of peptides of different lengths that bind to the cleft with a varying bulge in the center of the peptide [34,35]. The three strongest $\alpha_1\alpha_2$ clusters are located in the binding site where the structures for two different viral peptides [36] have been solved. Considerable agreement is found between the list of residues defining the three strongest clusters and residues with atoms observed in van der Waals contact with these peptides. Another somewhat less strongly interacting region is at the interface of the $\alpha_1\alpha_2$ and $\beta_2 m$ domains. Choosing the best 0.7% of the 1773 points requires examining twelve clusters, and six of these are found to have substantial overlap with other clusters. It is useful to eliminate the weaker cluster if it has more than 50% of its points in common with another.

A total number of 1728 shell points surrounds the lattice model of CD4, and examining the best 0.7% of the associated clusters corresponds to examining twelve clusters. Three of these can be eliminated because of overlap. The remaining ones

bury 64.4% of the total accessible surface of the CD4 two-domain fragment, nearly twice the total found for the HLA domains. This unusually large fraction of its surface area predicted to be involved in binding is consistent with studies of interactions between CD4 and gp120, monoclonal antibodies and class II MHC molecules [32,33,37–39]. These findings indicate an unusually large fraction of the surface of CD4 to be potentially interactive with other molecules.

DISCUSSION

Locating the most likely binding sites on a selected protein is an important initial step in designing a new ligand. For 22 of 23 enzyme-inhibitor complexes found in the Brookhaven Protein Data Bank, simple measures of hydrophobicity applied to an alpha-carbon model can be used to locate the correct receptor binding site within the top 0.7% of all possible sites. The procedure requires no prior knowledge of the ligands; only the structure of the target enzyme is required. The success of this procedure to locate surface targets lends additional support to the notion that surface hydrophobic interactions participate strongly in molecular recognition, confirming the biochemist's often stated intuition about 'sticky patches'. Furthermore, the present success provides evidence that peptides interact with similar strengths, whether intramolecularly in the interior of a folded protein or intermolecularly on its surface. And in both cases the largest stabilization comes from burial of hydrophobic residues.

Are these results simply a function of geometry so that one need only find the lattice point which defines the cluster with the largest number of residues, that is, a concave site on the protein? Even though the most favorable clusters often have a larger number of protein residues than most other clusters, selecting targets on the basis of the number of residues alone usually yields several clusters with similar numbers of residues. Geometric considerations alone do not provide a basis for choosing among these and would require that a larger number of clusters be examined as possible binding sites. The present approach of combining a lattice model with calculation of hydrophobicities to predict the binding site, includes such geometric effects implicitly, and is more restrictive [2].

This method of exhaustive lattice searching to explore the enzyme surface is novel in its thoroughness and in its use of surface hydrophobicity to identify possible interaction sites. Success in the use of surface hydrophobicity implies that interaction energies in the interior of a protein behave the same as those between peptide and protein surface residues.

The present analysis has focused on locating probable binding sites on a target protein and represents an initial step in a new approach for the design of peptide inhibitors. After a binding site has been chosen, peptide size and the composition of the prospective inhibitor can be addressed. A new approach attempts to combine the results described here, with previous efforts to identify the proper folds of simplified models of globular proteins. A preliminary analysis indicates that, for lattice models of peptides up to seven residues in length, the complete set of conformations can be enumerated [25,40] and examined for interactions with a target site. Once candidate peptides are identified, more detailed methods can be used to extend the single point representation to include all peptide atoms.

References

1. S. Miyazawa and R. L. Jernigan: *Macromolecules* **18**, 534 (1985).
2. L. Young, R. L. Jernigan, and D. G. Covell: *Prot. Sci.* **3**, 717 (1994).
3. S. Miyazawa and R. L. Jernigan: *Prot. Eng.* **6**, 267 (1993).
4. D. G. Covell and R. L. Jernigan: *Biochemistry* **29**, 3287 (1990).
5. I. Bahar and R. L. Jernigan: *Biophys. J.* **66**, 454 (1994).
6. I. Bahar and R. L. Jernigan: *Biophys. J.* **66**, 467 (1994).
7. P. J. Flory: *Statistical Mechanics of Chain Molecules*, Hanser Publ., Munich (1969).
8. R. S. Bohacek, U. P. Strauss, and R. L. Jernigan: *Macromolecules* **24**, 731 (1991).
9. S.-I. Mizushima: *Structure of Molecules and Internal Rotation*, Academic Press, N.Y. (1954).
10. S. Tanaka and H. A. Scheraga: *Macromolecules* **9**, 945 (1976).
11. L. M. Gregoret and F. E. Cohen: *J. Mol. Biol.* **211**, 959 (1990).
12. M. J. Sippl: *J. Mol. Biol.* **213**, 859 (1990).
13. D. A. Hinds and M. Levitt: *Proc. Natl. Acad. Sci. USA* **89**, 2536 (1992).
14. S. Miyazawa and R. L. Jernigan: *Prot. Eng.*, in press (1994).
15. J. U. Bowie, R. Lüthy, and D. Eisenberg: *Science* **253**, 164 (1991).
16. R. L. Jernigan: *Curr. Opin. Str. Biol.* **2**, 248 (1992).
17. D. G. Covell: *Proteins* **14**, 192 (1992).
18. R. L. Jernigan, G. Raghunathan, and I. Bahar: *Curr. Opin. Str. Biol.* **4**, 256 (1994).
19. L. Regan: *Annu. Rev. Biophys. Biomol. Struct.* **22**, 257 (1993).
20. D. J. Danziger and P. M. Dean: *Proc. Roy. Soc. Lond.* **B236**, 101 (1989).
21. D. J. Danziger and P. M. Dean: *Proc. Roy. Soc. Lond.* **B236**, 115 (1989).
22. P. Goodford: *J. Med. Chem.* **28**, 849 (1985).
23. A. Nicholls, K. A. Sharp, and B. Honig: *Proteins* **11**, 281 (1991).
24. B. K. Shoichet, D. L. Bodian, and I. D. Kuntz: *J. Comput. Chem.* **13**, 1 (1992).
25. R. L. Jernigan, H. Margalit, and D. G. Covell: 'Coarse graining conformations: A peptide binding example', in: D.L. Beveridge and R. Lavery (Eds.), *Theoretical Biochemistry and Molecular Biophysics*, Vol. 2, Adenine Press, Schenectady, New York, p. 69 (1991).
26. E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng: in: F. H. Allen, G. Bergerhoff, and R. Sievers (Eds.), *Crystallographic Databases–Information Content, Software Systems. Scientific Applications*, Data Commission of the International Union of Crystallography Bonn and Cambridge and Chester, p. 107 (1987).
27. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi: *J. Mol. Biol.* **112**, 535 (1977).
28. Y. Nozaki and C. Tanford: *J. Biol. Chem.* **246**, 2211 (1971).
29. J. L. Cornette, K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky, and C. DeLisi: *J. Mol. Biol.* **195**, 659 (1987).
30. S. Sheriff, E. W. Silverton, E. A. Padlan, G. H. Cohen, S. J. Smith-Gill, B. C. Finzel, and D. R. Davies: *Proc. Natl. Acad. Sci. USA* **84**, 8075 (1987).
31. M. A. Saper, P. J. Bjorkman, and D. C. Wiley: *J. Mol. Biol.* **219**, 277 (1991).
32. J. Wang, Y. Yan, T. P. J. Garrett, J. Liu, D. W. Rodgers, R. L. Garlick, G. E. Tarr, Y. Husain, E. L. Reinherz, and S. C. Harrison: *Science* **348**, 411 (1990).
33. S. Ryu, P. D. Kwong, A. Truneh, T. G. Porter, J. Arthos, M. Rosenberg, X. Dai, N. Xuong, R. Axel, R. W. Sweet, and W. A. Hendrickson: *Science* **348**, 419 (1990).
34. P. Parham: *Nature* **360**, 300 (1992).
35. H. Guo, T. S. Jardetzky, T. P. J. Garrett, W. S. Lane, J. L. Strominger, and D. C. Wiley: *Nature* **360**, 364 (1992).
36. D. H. Fremont, M. Matsumura, E. A. Stura, P. A. Peterson, and I. A. Wilson: *Science* **257**, 919 (1992).
37. L. J. Clayton, N. Sieh, D. A. Pious, and R. L. Reinherz: *Nature* **339**, 548 (1989).
38. D. J. Capon and R. H. R. Ward: *Annu. Rev. Immunol.* **9**, 649 (1991).
39. G. J. Szabo, P. S. Pine, J. L. Weaver, P. E. Rao, and A. Aszalos: *J. Immunol.* **149**, 3596 (1992).
40. R. L. Jernigan: 'Generating general shapes and conformations with regular lattices, for compact proteins', in: R.H. Sarma and M.H. Sarma (Eds.), *Structure and Function: Proceedings of Seventh Conversation in Biomolecular Stereodynamics*, Vol. 2, Adenine Press, Schnectady, NY, p. 169 (1991).