Composites of local structure propensities: Evidence for local encoding of long-range structure

DAVID SHORTLE

Department of Biological Chemistry, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

(RECEIVED July 26, 2001; FINAL REVISION October 1, 2001; ACCEPTED October 4, 2001)

Abstract

To estimate how extensively the ensemble of denatured-state conformations is constrained by local side-chain–backbone interactions, propensities of each of the 20 amino acids to occur in mono- and dipeptides mapped to discrete regions of the Ramachandran map are computed from proteins of known structure. In addition, propensities are computed for the trans, gauche−, and gauche+ rotamers, with or without consideration of the values of phi and psi. These propensities are used in scoring functions for fragment threading, which estimates the energetic favorability of fragments of protein sequence to adopt the native conformation as opposed to hundreds of thousands of incorrect conformations. As finer subdivisions of the Ramachandran plot, neighboring residue phi/psi angles, and rotamers are incorporated, scoring functions become better at ranking the native conformation as the most favorable. With the best composite propensity function, the native structure can be distinguished from 300,000 incorrect structures for 71% of the 2130 arbitrary protein segments of length 40, 48% of 2247 segments of length 30, and 20% of 2368 segments of length 20. A majority of fragments of length 30–40 are estimated to be folded into the native conformation a substantial fraction of the time. These data suggest that the variations observed in amino acid frequencies in different phi/psi/chi1 environments in folded proteins reflect energetically important local side-chain–backbone interactions, interactions that may severely restrict the ensemble of conformations populated in the denatured state to a relatively small subset with nativelike structure.

Keywords: Amino acid propensities; rotamers; denatured state; side-chain–backbone interactions; Ramachandran plot; threading

The ensemble of conformations adopted by polymers in solution is strongly influenced by interactions between adjacent monomers (Flory 1969). These local interactions were analyzed in proteins by Ramachandran and colleagues (Ramachandran and Sasisekharan 1968), who mapped out many of the steric clashes that restrict the backbone angles of polypeptide chains. Their work accurately anticipated the range of phi/psi values found in native structures by X-ray crystallography. From simple-steric arguments, it is clear that special consideration must be given to the amino acids proline and glycine, whose side chains impose steric constraints quite different from those of the other 18 naturally occurring amino acids. These constraints are clearly reflected in the patterns of phi/psi angles adopted by these two residues in high-resolution structures (Richardson and Richardson 1989).

Although much less dramatic than proline and glycine, the other 18 amino acids also display nonuniform distributions in phi/psi angles, most easily seen in the three types of local backbone structure: alpha helix, beta strand, and turn. Chou and Fasman (1974) quantified these patterns by calculating propensities for the amino acids to occur in these three structures and found approximately two- to threefold variations in their values between the nonglycine, nonproline amino acids. With these propensities and a few simple

Reprint requests to: David Shortle, Department of Biological Chemistry, 725 North Wolfe Street, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; e-mail: shortle@welchlink.welch.jhu.edu; fax: (410) 955-5759.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1101/ps.31002.

rules, they were able to predict with modest accuracy the secondary structure of protein segments from amino acid sequence. More detailed examination of the distributions of phi/psi angles, either in the context of repeating secondary structure or not, has revealed variations in the propensities of different amino acid types to occupy different subdivisions within the large β-sheet region of the Ramachandran map (Munoz and Serrano 1994; Swindells et al. 1995). Presumably these variations, like those observed in the alpha, beta, and turn propensities, reflect relatively subtle interactions between the side chain (the only variable from one amino acid type to another) and the two peptide bonds that flank it.

Considerable experimental and theoretical work since Chou and Fasman has been directed toward more fully quantifying the secondary structural propensities of the 20 amino acids and explaining their physical chemical origin. Although a variety of chemical mechanisms have been emphasized over the years, there is general agreement that structural propensities arise primarily from repulsive interactions (i.e., avoidance of steric overlap [Creamer et al. 1995; Srinivasan and Rose 1999; Street and Mayo 1999), with perhaps a smaller component due to attractive interactions involving dispersion forces (Yang and Honig 1995a,b) or burial of nonpolar surface (Blaber et al. 1993). As pointed out by Creamer and Rose (1992), the backbone of a polypeptide chain will adopt values of phi/psi that maximize the configurational entropy of the side chains.

Although steric repulsion is probably the dominant interatomic force shaping the structure of molecules, calculation of the reduction in conformational entropy that arises from clashes between neighboring monomers in a polymer chain is an exceeding difficult problem (Flory 1969). Consequently, there is no clear picture as to how great a role these short-range interactions play in constraining the conformations accessible to a protein chain as it folds. The principal question addressed here is, How severely is the ensemble of denatured conformations restricted by local steric effects? Recent work from this laboratory has shown the persistence of nativelike long-range structure in a denatured protein in 8M urea (Shortle and Ackerman 2001), conditions that greatly reduce hydrophobic interactions and increase the exposure of the peptide backbone to solvent. These findings raise the possibility that long-range structure may not depend entirely on long-range interactions, but instead may arise through the cumulative effects of many local interactions between each amino acid and its immediate neighbors (Pappu et al. 2000).

Below is a preliminary attempt to utilize the nonuniform distributions of phi/psi values of the 20 amino acids observed in folded proteins to estimate the relative importance of side-chain–backbone interactions in shaping protein structure. Beginning with the Boltzmann hypothesis (Sippl 1993; Finkelstein et al. 1995) that these skewed distributions reflect the free energy of placing a side chain in a specific backbone environment, statistical potentials are generated from a large library of protein structures and used to estimate the free-energy difference between native and incorrect structures. By adding together several propensities involving only one or two adjacent amino acid residues, simple scoring functions can be constructed that, for fragments of 40 residues, can correctly identify the one native conformation out of a total of 300,000 conformations with an accuracy of up to 70%.

Results

Calculation and use of composite propensity functions

To generate scoring functions for estimating the energetics of side-chain–backbone interactions, several local structural propensities are calculated. The only structural parameters used to represent local interactions are the backbone dihedral angles phi and psi and the side-chain dihedral angle chi1. If X represents a discrete environment characterized by one or more of these parameters and B a particular amino acid type, a propensity P can be defined as

$$P_{B,X} = (\text{number of } B \text{ with value } X/\text{number of } B)/(\text{number of amino acids with value } X/\text{number of amino acids}).$$

Thus, the propensity is a measure of the likelihood that amino acid B will be found in an environment described by X. A propensity >1.0 indicates that B is favored relative to the mean of all 20 amino acids; whereas a propensity <1.0 indicates it is disfavored. Inherent in its definition, a propensity is a relative measure of preference, and thus is always normalized to an average or mean residue type. A probability, on the other hand, is a measure of the absolute likelihood that an amino acid will adopt one out of a specified set of structures. A probability is not normalized to other amino acids and can be defined as

$$p_B(x) = \text{number of } B \text{ with value of } X/\text{number of } B.$$

Note that a propensity is the ratio of two probabilities.

For phi/psi values, propensities are calculated for single-residue positions assigned to discrete regions or states of the Ramachandran plot, as shown in Figure 1. These propensities are designated by pN, where N is the number of discrete phi/psi states employed. For adjacent pairs of residue positions or dipeptides, propensity pairs are also calculated using discrete states. These propensities are designated pN_1xN_2, where N_1 designates the number of discrete phi/psi states describing residue i, and N_2 the number of states describing both i+1 and i-1.

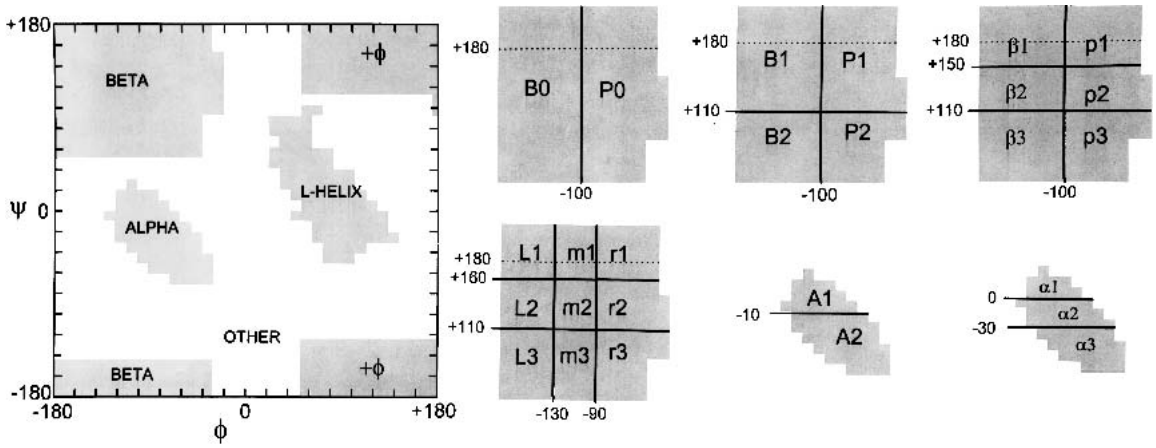

Fig. 1 .The major regions of the Ramachandran plot and the subdivisions used in this work. The p4 propensities used subdivisions alpha, beta, L-helix, and ϕ combined with other. The p6 propensities were B0, P0, alpha, L-helix, ϕ, and other. The p9 propensities were B1, B2, P1, P2, A1, A2, L-helix, ϕ, and other. The p12 propensities were β1, β2, β3, p1, p2, p3, α1, α2, α3, L-helix, ϕ, and other. The p15 propensities were L1, L2, L3, m1, m2, m3, r1, r2, r3, α1, α2, α3, L-helix, ϕ, and other.

For chi1 values, probabilities are calculated for the states trans, gauche+, and gauche−, corresponding to values −120 to +120, 0 to +120, and 0 to −120, respectively. When phi/psi angles are ignored, the resulting backbone-independent rotamer probabilities are designated *r*0 and are found to be in good agreement with those reported by Ponder and Richards (1987). When the rotamer states are analyzed for residues assigned to *N* discrete phi/psi states, backbone-dependent rotamers (Dunbrack and Karplus 1993) are designated *rN*.

The propensity of each amino acid is treated as an equilibrium constant for partitioning that side chain into the corresponding local environment. Thus, the logarithm of the propensity approximates the free energy of exchanging an average side chain with a specific one. The free energy for each residue position is assumed to be additive, so the score for a sequence fragment is the sum of the log of the propensities at each position. In some cases, composite propensities are used. Because these propensities can be viewed as representing different steric interactions, the simple assumption is made that these interactions are independent, allowing logarithms of the propensities to be added to form a composite.

The additional assumption is made that native structure of each fragment corresponds to a global minimum in free energy. Therefore, if the native structure achieves a better score than all other conformations sampled for a sequence fragment, it follows that the scoring function probably contains important components of the true energy function. The odds of identifying the native structure by chance should be proportional to the inverse of the number of conformations sampled. Thus, the initial focus of this work is to identify the composite propensity function that most consistently assigns the best score to the native conformation.

The library of protein structures used in this work consists of 1700 protein structures from the VAST nonredundant database (Madej et al. 1995). This combined set of proteins is divided at random into a training set of 1579 structures (93%) for calculation of propensities/probabilities and a test set of 121 structures (7%), listed in Materials and Methods, for generating sequence fragments for analysis. The training set provides structural information on a total of 332,768 amino acid positions.

Using all 121 proteins in the test set, four series of sequence fragments (lengths of 10, 20, 30, and 40 residues) are constructed, starting with residue 3 and ending at or before 3 residues prior to the carboxyl terminus. Successive fragments obtained from one protein overlap by 10 residues; so the first fragment begins at residue 3, the next at residue 13, then residue 23, and so on, until no more of that length can be generated. In this way, standard lists of arbitrarily chosen fragments are produced from the proteins in the test set, consisting of 2130 fragments of length 40, 2247 of length 30, 2368 of length 20, and 2489 of length 10.

To evaluate the energetics of each sequence fragment arranged in a large number of alternative conformations, it is threaded through simplified linear representations of all 1700 proteins in the combined set. Thus, the set of nonnative conformations sampled by each sequence fragment consists of ~300,000 structural fragments taken from known proteins. Whereas this represents a small sample of conformation space, it has the advantage of high computational efficiency and provides a very broad, albeit coarse-grained sampling over a substantial portion of conformation space. During threading, a running collection is kept of the 300 best-scoring fragments, along with their proteins of origin and their end points.

Table 1. *The p15 propensities for the 20 amino acids*

Amino acid	No.	α1	α2	α3	L1	L2	L3	m1	m2	m3	r1	r2	r3	LH	+ϕ	Other
ALA	27574	0.669	1.11	1.44	0.931	1.08	0.828	0.518	0.538	0.529	0.762	0.996	0.744	0.307	0.288	0.746
ARG	16858	1.04	1.03	1.2	0.88	1.24	0.968	0.85	0.882	0.921	0.578	0.846	0.812	0.582	0.256	0.996
ASN	15025	2.35	1.09	0.655	1.07	0.65	2.68	1.01	0.667	2	0.899	0.791	2.13	2.53	0.266	1.48
ASP	19487	2.06	1.16	0.881	0.822	0.469	2.07	0.926	0.537	2.01	1.25	1.05	2.51	1.17	0.321	1.19
CYS	5714	0.945	0.878	0.704	1.79	1.28	1.89	1.52	1.23	1.46	1.13	1.21	1.38	0.5	0.226	1.12
GLN	13183	1.07	1.15	1.26	0.64	1.06	1.12	0.876	0.84	0.847	0.609	0.783	0.778	0.655	0.176	0.9
GLU	21803	0.787	1.25	1.37	0.517	0.852	0.828	0.598	0.727	0.697	0.484	0.866	0.881	0.473	0.204	0.977
GLY	24889	0.621	0.482	0.432	1.75	0.388	0.284	1.52	0.206	0.184	1.73	0.297	0.395	7.07	10.9	1.91
HIS	7645	1.55	1.05	0.832	1.28	1.34	2.06	0.862	0.926	1.64	0.629	0.849	1.05	1.04	0.312	1.32
ILE	18776	0.511	0.561	1.09	0.487	1.21	0.747	0.747	2.19	1.59	0.237	0.873	0.795	0.049	0.0825	0.659
LEU	29746	0.885	1.07	1.29	0.299	0.673	0.619	0.781	1.29	1.03	0.521	0.945	0.879	0.198	0.124	0.712
LYS	20187	0.977	1.13	1.17	0.62	0.984	0.949	0.898	0.856	0.857	0.61	0.949	0.904	0.785	0.269	1.07
MET	6905	0.965	1.06	1.23	0.931	1.31	1.23	0.694	0.982	0.96	0.524	0.842	0.815	0.33	0.234	0.668
PHE	13109	1.2	0.772	0.952	1.65	1.42	1.14	1.06	1.44	1.5	0.465	0.829	0.841	0.465	0.143	0.841
PRO	15701	0.474	1.38	0.621	0.013	0.007	0.006	0.278	0.071	0.0997	4.38	3.82	2.4	0.006	0.0205	0.697
SER	20111	1.01	1.43	0.745	2.05	1.44	1.13	1.47	0.731	0.635	2	1.02	0.725	0.475	0.436	1.11
THR	19156	1.32	1.09	0.761	1.84	1.42	0.751	2.51	1.32	0.746	1.48	0.792	0.442	0.152	0.192	1.11
TRP	4725	1.02	1.04	1.08	1.44	1.05	1.16	0.941	1.21	1.09	0.542	1.02	0.828	0.277	0.15	0.639
TYR	11868	1.24	0.781	0.915	1.75	1.53	1.12	1.23	1.41	1.24	0.521	0.848	0.669	0.402	0.147	0.976
VAL	23464	0.421	0.572	0.955	0.74	1.57	0.789	0.895	2.28	1.37	0.275	0.887	0.73	0.065	0.0935	0.689

The column headings correspond to subdivisions of the Ramachandran plot shown in Figure 1.

Single-peptide propensities

The two-dimensional space defined by angles phi and psi (i.e., the Ramachandran plot) is partitioned into 10-degree-by-10-degree squares and grouped into allowed regions using approximately the same boundaries employed by Swindells et al. (1995). These regions are further subdivided as shown in Figure 1 to represent the entire plot by 4, 6, 9, 12, or 15 discrete states. The areas covered by these five representations are given in the Figure 1 legend.

Propensities for each of the 20 amino acids are computed from the training set. (The propensities for the 15-state representation p15 are listed in Table 1.) For use in scoring functions, propensities <0.03 are set to the value of 0.03 before taking the logarithm. All proteins in the combined set are reduced to a one-dimensional vector, with each residue position assigned a single numerical value based on the region of the discrete Ramachandran plot to which it is mapped. This value, plus the amino acid type, serve as indices to recover the logarithm of the amino acid propensity for scoring each residue position during threading.

The score of the native conformation is compared to the scores for the incorrect conformations and its relative ranking determined. The fraction of fragments of length 40 that ranked as either the best 1 or within the top 0.1% (300 out of 300,000) are shown in Figure 2 for the p4, p6, p9, p12, and p15 propensities. Each increment in the number of discrete states representing the Ramachandran plot yields a scoring function that more consistently assigns the best score to the native conformation. Although only 1.0% (21/2130) of fragments are identified as having the most favorable score with the 4-state representation, this value increases to 19.8% (421/2130) for the 15-state map. Similarly, the percentage of fragments that rank within the top 0.1% climbs from 20.8% to 61.7%.

A similar analysis for fragments of length 10, 20, and 30 residues is presented in Figure 2 and clearly shows that all scoring functions become less discriminating with shorter fragments. Presumably, scores for shorter fragments contain

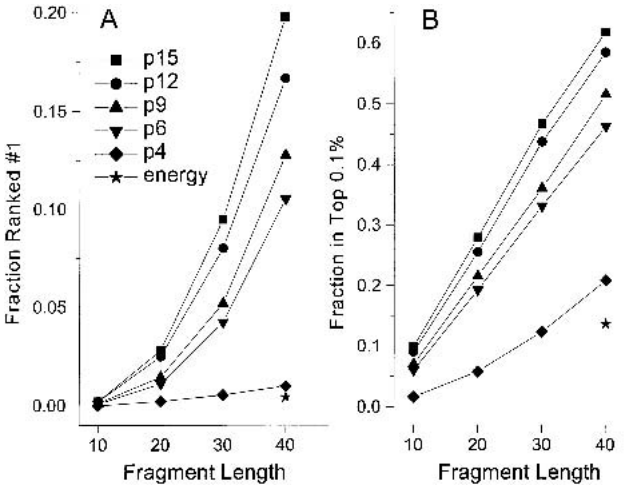

Fig. 2. Fraction of 2140 sequence fragments of length 40 for which the native conformation ranked (*A*) No. 1 (i.e., with the best score) or (*B*) in the top 0.1% of conformations (300 per 299,779). The score consisted of the sum of the logarithm of the single-peptide propensities, pN, where $n = 4$, 6, 9, 12, 15, or the Bryant and Lawrence (1993) empirical pair potential score (energy).

more statistical noise; whereas longer fragments, with the summation of more terms, will more effectively average out the noise. However, since the number of possible conformations is smaller for short fragments, threading through a fixed set of proteins will generate a more complete sampling for these sequences.

Although there is no rigorous way to quantify the contribution of long-range, as opposed to short-range, interactions, empirical potentials for pairs of amino acids separated by varying distances have been shown to be effective in fold recognition (Moult 1997; Vajda et al. 1997) and are used extensively for ab initio folding (Bonneau and Baker 2001). When the same series of 40-residue fragments are evaluated using the distance-dependent potentials of Bryant and Lawrence (1993), the scores of the native conformations are seldom the best (0.42%), although there is a modest ability to assign them to the top 0.1% (13.6%).

Dipeptide propensities

Propensities for single-residue positions can only capture side-chain interactions with the two nearest peptide groups. To include interactions with the two next-nearest groups, which belong to the preceding and following residues, propensities are calculated for dipeptide pairs, using the amino acid type and discrete Ramachandran plots for residue *i*. For residues *i*+1 and *i*-1 residue, the amino acid type is not considered, only its phi/psi values. Rather than deal with tripeptides and the small number of examples available in the structural library, the assumption is made that the *i*-1 to *i* and *i* to *i*+1 interactions are independent. Therefore, separate propensity tables are calculated for each of these two dipeptides and residue positions in the protein library are now described by two environments. On scoring a sequence fragment, an individual residue is assigned the average of the logarithms of these two propensities.

There is no requirement that the number of discrete phi/psi states be the same for residue *i* as for its neighbors; so several compatible combinations are calculated. These pN_1xN_2 composite propensities, where N_1 is the number of states for residue *i* and N_2 the number for both residues *i*-1 and *i*+1, are calculated from the set of training structures and used in scoring functions. The results of threading the test set of length 40 fragments for 10 different dipeptide propensities are given in Figure 3A, which shows the increase in fraction of fragments for which the native conformation receives the best score, as a function of the number of states N_2.

For each of the five representations of residue *i*, there is a modest improvement when the single-peptide propensity *pN* score is changed to a dipeptide form, suggesting that side-chain interactions with the two next-nearest peptide groups do contribute to the overall energetics. Although performance of the scoring function initially improves with increasing N_2, this improvement appears to plateau around $N_2 = 6$.

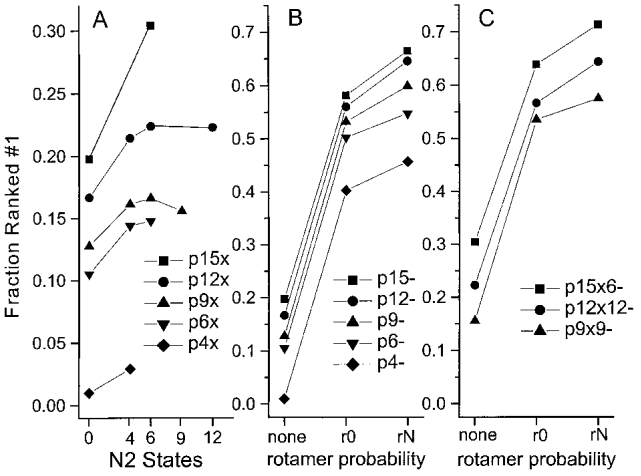

Fig. 3. Fraction of 2140 sequence fragments of length 40 for which the native conformation had the best score (i.e., ranked No. 1). (*A*) As a function of the number of N_2 states used to represent residues i−1 and i+1 in dipeptide propensities. (*B*) As a function of rotamer probability used in combination with single peptide propensities. (*C*) As a function of rotamer probability used in combination with dipeptide propensities.

Rotamer probabilities

The single and dipeptide propensities described above ignore the chi1 angles of the side chains. Although the trans, gauche+, and gauche− rotamers could be used to further subdivide the N_1xN_2 phi/psi states for calculation of propensities, this would triple the number of bins for distributing the data, leading to increased noise due to small sample size. Instead, the assumption is made that rotation about each side chain's chi1 angle is an independent variable that can be restored to the description of the averaged structure described by the phi/psi propensity by simply adding the logarithm of its probability. Although this assumption is unlikely to be strictly true, it is made as a first approximation.

The side-chain–backbone interactions that arise through the discrete values of chi1 (trans, gauche−, and gauche+) at position *i* are converted to either backbone-independent (*r*0) or backbone-dependent (*rN*) probabilities, where *N* is the number of discrete states representing the Ramachandran plot. In both cases, the frequency of occurrence of each rotamer is normalized so the sum of all probabilities equal one. To give equal weight to the rotamer state in a composite propensity function, the probabilities for each of the three rotamers are multiplied by 3.0 before taking the logarithm. Positions either in the protein structure or in the sequence being threaded involving alanine and glycine (which have no value of chi1) are given no score. These

composite propensities are designated by appending the rotamer probability name to the backbone propensity name.

As shown in Figure 3B, the addition of a rotamer probability term to each of the single-peptide propensities significantly improves the scoring function. Similarly, inclusion of the rotamer probability score improves the performance of three different dipeptide propensities: p9x9, p12x12, and p15x6 (Fig. 3C). Although for p9x9 there is very little difference between *r*0 and *r*9, both p12x12 and p15x6 score higher with the backbone-dependent rotamer probabilities. p15x6-r15 gives the best score of any composite propensity reported here: 71.2% of all fragments of length 40 give the best score with the native conformation and 95.6% of fragments score in the top 0.1%. The values with this propensity function for shorter fragments are as follows: length 30–48.2% and 89.3%, length 20–20.4% and 70.9%, and length 10–2.9% and 31.3%.

Estimation of conformational stability

To estimate the energetic significance of the local side-chain–backbone interactions that underlie the propensity-based scores presented above, the scores for individual sequence fragments can be converted into the probability that the native conformation will be populated out of all conformations in the ensemble. From statistical mechanics, this probability *p*(native) is

$$p(\text{native}) = e^{-\Delta G\text{native}/RT}/\Sigma e^{-\Delta G\text{conf}/RT},$$

where the denominator is the sum overall conformations (i.e., the partition function). Because

$$-\Delta G_{conf}/RT = \ln K_{eq},$$

the probability can be rewritten as

$$p(\text{native}) = K_{eq}(\text{native})/\Sigma K_{eq}(\text{conf})$$

where the denominator is again the sum over all allowed conformations.

At least four assumptions must be made before this calculation can be justified. (1) The frequencies of protein structural features describe an equilibrium ensemble at a temperature close to physiological. The validity of this assumption has been extensively debated in the literature (Sippl et al. 1996; Thomas and Dill 1996). Simply for the sake of evaluating its consequences, this assumption is made. (2) Each conformation is treated as a representative example of many closely related conformations. Just as the single native conformation represents a large number of slightly altered conformations accessible by thermal motions, each nonnative conformation acts as an average surrogate for a small family of closely related conformations. Then the assumption must be made that the 300,000 conformations are a relatively complete coarse-grained sample of conformation space. Although there is no good way to assess the reasonableness of this assumption, it becomes a better approximation of reality when rotamer states are ignored or when sequence fragments are shorter. (3) The log of the probability for a conformation of a particular fragment can be approximated by the sum of the logarithms of the propensities at every residue position. (4) For a composite propensity function, the logarithms of the individual propensities and probabilities describing a single residue position can be added. It should be noted that for each function, only one side-chain substitution event is allowed . Thus, for the dipeptide propensities, the average of two events is taken; whereas with rotamers, the use of probabilities instead of propensities implies the side chain has been specified previously.

The probability that the native conformation is populated is then calculated by

$$p(\text{native}) = \text{antilog}(\text{native fragment score})/\Sigma\text{antilog}(\text{all fragment scores}).$$

In Figure 4, histograms display the calculated range of probabilities for occupying the native conformation using four different propensity functions, two that do not include rotamers, p15 and p15x6, plus these same two combined with the r15 rotamer probabilities. The results are surprising. With backbone information alone, the p15 and p15x6 results suggest that 35%–50% of all sequence fragments of length 40 will occupy the native conformation with a probability >0.01 and 5%–11% of all fragments will occupy the native conformation >80% of the time.

The estimated stabilities of native conformations become much greater when rotamer states are included, with the p15x6-r15 function suggesting that >50% of fragments will occupy the native conformation >95% of the time. The fraction of sequence fragments that are mostly native declines with decreasing fragment length, but the p15x6-r15 function estimates that ~50% of 20mers will be native >1% of the time.

These percentages should be considered as very crude estimates based on patently optimistic assumptions about the completeness of sampling and the additivity of the probabilities involved. The exact number of conformations sampled by fragment threading are quite small: 297,779 for length 40; 314,070 for length 30; 330,997 for length 20; and 347,987 for length 10. For a fragment of length 40, there could be as many as 3^{40} or 10^{19} different rotamer combinations. Therefore, the only safe conclusion that can be drawn is a simple one: These propensities must reflect underlying physical interactions that encode structure at distances beyond the one or two residues involved in the in-

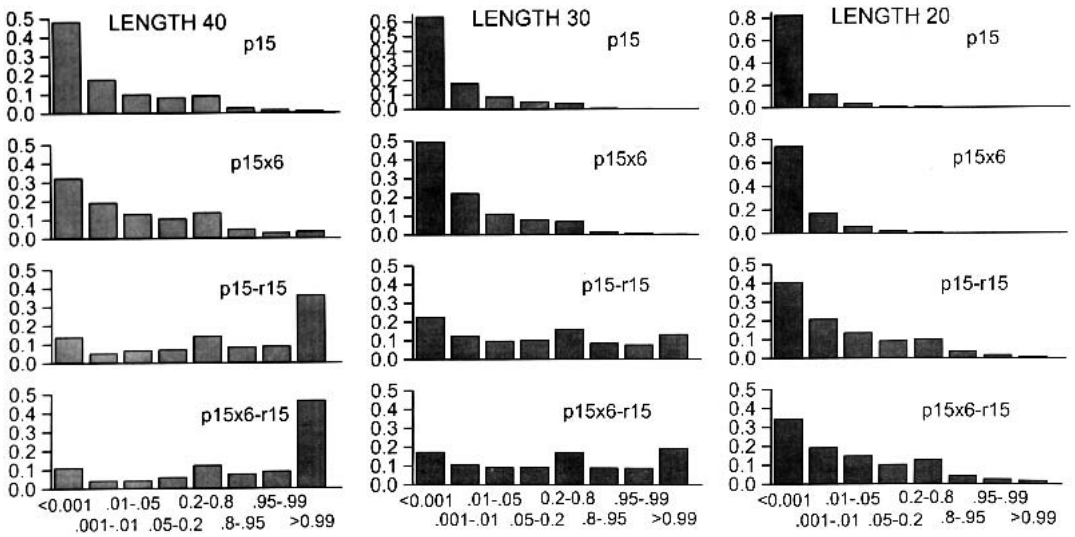

Fig. 4. Histograms of estimated probabilities that sequence fragments occupy the native conformation, as opposed to 300,000 incorrect conformations, as a function of length and composite propensity function. Fragment length and propensity function used in threading are given in the panel. The ranges of probabilities are listed below each column of the histograms.

teraction, perhaps a dominant role, perhaps only a modest role.

A totally independent line of empirical evidence supporting this inference would be welcome. If these local interactions are important, it would be expected that for some proteins, a greater level of stability to unfolding might be attained through optimization of these interactions, a situation reflected in higher propensity scores. Therefore, 23 homologous protein pairs (Kannan and Vishveshwara 2000), one from a thermophilic organism and the other from a mesophile, are evaluated with the p15x6-r15 function. As shown in Figure 5, the results are somewhat suggestive that thermophilic proteins, on average, may have higher scores. Of the 23 pairs, in only 5 cases does the mesophilic protein score higher than its thermophilic homolog, and in all but 1 instance (No. 4) by <2 log units per 100 residues. On the other hand, 18/23 pairs have the thermophilic protein scoring higher, in 10 instances by more than 4 units per residue. Although not compelling, these data suggest that there may be a significant bias toward more favorable side-chain–backbone interactions in proteins from thermophilic organisms, a feature that could contribute to thermostability by enhancing nativelike structure in the denatured state (Wrabl and Shortle 1996).

Discussion

Several patterns in the data are consistent with a physical basis for the side-chain–backbone propensities reported above. (1) As more discrete states are used to describe phi/psi angles, single-peptide propensities calculated from a library of structures become increasingly more accurate in identifying the native conformation of long sequence fragments. (2) Dipeptide propensities, which are expected to capture the steric effects of the second nearest pair of peptide groups, give rise to small but consistent improvements in scores. However, use of more than six states to describe the phi/psi angles of the neighboring residue appears to have

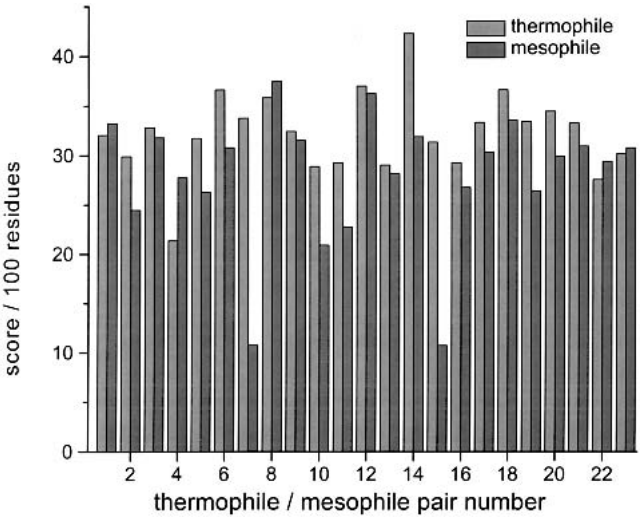

Fig. 5. Composite propensity function scores (p15x6-r15) per 100 residues for 23 homologous pairs of proteins, one from a thermophilic organism and the other from a mesophilic organism. Their Protein Data Bank designations are as follows: 1:1thl/1npc, 2:1ldn/1ldm, 3:3pfk/2pfk, 4: 1ril/2rn2, 5:1bmd/4mdh, 6:2prd/1ino, 7:1php/3pgk, 8:1thm/1st3, 9:1ebd/1lvl, 10: 1btm/1tim, 11:2fxb/1dur, 12:1yna/1xyn, 13:1xyz/2exo, 14:1caa/6rxn, 15: 1gd1/1gdp, 16:1tib/1lgy, 17:1zip/1ak2, 18:2prd/1obw, 19:1ais/1vol, 20: 1ffh/1fts, 21:1pcz/1vok, 22:1obr/2ctc, and 23:1phn/1cpc. This set of proteins is from Kannan and Vishveshwara (2000).

little effect. (3) Inclusion of the rotameric state of the side chain leads to substantially better scoring functions, with the backbone-dependent rotamer probabilities being consistently better than probabilities that ignore the residue's phi/psi angles.

The principle assumption on which this work is based is the Boltzmann hypothesis, by which the biases observed in the distribution of local structure features in folded proteins can be treated as equivalent to a partitioning reaction at equilibrium. The frequency of occurrence of an amino acid within a local structure reflects the free energy of exchanging it for an average side chain. This hypothesis (Pohl 1971; Finkelstein et al. 1995) has been used for estimating the free energy of a number of structural features observed in proteins: side-chain hydrophobicity (Rose et al. 1985; Miller et al. 1987), energies of mean pair-wise interactions among side chains (Miyazawa and Jernigan 1985; Sippl 1993), internal cavities (Rashin et al. 1997), *cis*/*trans* isomers of proline (MacArthur and Thornton 1991), and so on. When the free energies estimated from propensities have been compared to data measured by conventional physical methods, often surprisingly good agreement has been found. As has been argued by several investigators (Jones and Thornton 1996; Moult 1997), one need make no assumptions to justify the summation of logarithms of propensities as a scoring function. Instead, construction of such functions can be viewed as a practical exercise, justified by the utility of the results. Still, the Boltzmann hypothesis provides a concrete conceptual framework for interpretation of propensities and thus it has been used here.

Although the estimated stabilities of fragments of length 30 and 40 are not reliable, it seems highly probable that the propensity scores on which they are based represent an underestimate of their free-energy contributions. First of all, the binning of data for large regions of the Ramachandran plot, like any type of averaging over an interval, replaces local maxima and minima across the interval with a mean value that lies between the extrema. Secondly, the choice of intervals used here was arbitrary, and the data suggest that additional subdivisions of the beta region may yield still better scoring functions. Finally, only the chi1 angle is treated in this work. In effect, the side chain beyond the CG atom has been ignored, even though its steric consequences are likely to be large.

Several features of composite propensity functions may make them well-suited for sequence-structure compatibility comparisons. Because they are normalized to an average amino acid as a reference state, their values correspond to differences, not absolute values normalized to an external standard. Individual component propensities contribute to the final score in proportion to the magnitude of the variation in frequency of occurrence of a structural feature. In effect, insignificant variations, on average, will contribute little to the final score, and to the extent the Boltzmann hypothesis applies, logarithms of propensities correspond to the thermodynamic potential that governs the details of structure—free energy, not internal energy. Both the entropic and energetic components are included in these potentials of mean force.

At least two previous studies of phi/psi propensities for scoring sequence-structure compatibility reported findings less impressive than those above. Matsuo and Nishikawa (1993) employed a five-state representation of the Ramachandran plot and used their propensities as one of four energy components; details were not given. Bahar et al. (1997) employed a two-state representation, alpha and beta, and showed by ungapped protein threading that their propensity-derived terms could recognize a significant majority of full-length proteins. It would appear from the data reported above that some of the information in phi/psi propensities may have been missed in earlier analyses because partitioning of the Ramachandran plot into a number of subregions is required to reveal the full extent of the nonrandom phi/psi distributions of the 20 amino acids. As seen in Table 1, serine and threonine display a two- to threefold preference for the upper reaches of the beta region, and aspartate, asparagine, and histidine have a similar preference for the lower reaches. Alanine and arginine, on the other hand, have a slight preference for the middle. Similar patterns can be found for the three subdivisions of the alpha helical region. Presumably, a combination of steric clashes and attractive interactions—dispersion forces plus hydrogen bonds/electrostatic interactions—account for these variations. Although modest in size, the cumulative effect of these local side-chain–backbone interactions may severely restrict the conformations accessible to a polypeptide chain.

The remarkable ability of composite propensity functions to identify the native conformation for fragments of length 40 suggests they will prove useful in scoring functions for fold recognition. Because identification of the native conformation improves with chain length, they may outperform the more commonly used empirical pair potentials. A more practical application might be in optimizing the alignment of distant sequence homologs with proteins of known structure, although success will depend on how extensively local backbone geometry is conserved among homologs. For ab initio structure prediction methods that use simplified representations of protein chains (Bonneau and Baker 2001), composite propensity functions may provide a convenient way to estimate the steric plausibility of a conformation when physically important atoms have been omitted. Perhaps reasonable models of the denatured state can be generated by splicing together several overlapping fragments of length 10–30 residues that satisfy local steric restraints. In subsequent steps, by essentially recapitulating the folding process, conformations with few long-range steric clashes could be selected and further refined for compactness and hydrophobic burial in a search for the native state.

Materials and methods

Protein structural library

VAST database was obtained from http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html in December 2000. Of these 1926 structures with BLAST *p* values of less than 10e-7, only 1700 could be read. The randomly selected subset of 121 used for testing the propensity functions consisted of members from each of the SCOP classification of fold classes. All alpha were 1afr, 1akh, 1an2, 1ax8, 1b0n, 1ba5, 1bsm, 1c3d, 1cd3, 1ctj, 1ddf, 1dn1, 1ebm, 1eh2, 1ery, 1gah, 1lre, 1qa6, 1qgk, 1qj2, 1qla, 1qq8, 1rep, 1ryt, 1tf4, 1vls, 1xpa, 1zym, 2lef, 2occ, 2prg, 2spc, and 3ygs. All beta were 1a1x, 1ahj, 1as7, 1b35, 1bdo, 1bhe, 1bmv, 1bpv, 1bw3, 1cfb, 1cn3, 1d5r, 1dab, 1ewi, 1fmt, 1gpc, 1hsq, 1icm, 1ndh, 1pdk, 1pse, 1qun, 1rmg, 1shc, 1sox, 1tsr, 1ubp, 1vcb, 1wpo, 2bbv, 2ncm, and 3msi. Alpha/beta included 1a7a, 1a9n, 1auz, 1b4v, 1cz3, 1d2r, 1dfm, 1din, 1dpg, 1hjr, 1iow, 1jfr, 1kas, 1mee, 1mug, 1ofg, 1poy, 1tkb, 1yts, 2cmd, 2ebn, 3chy, 5p21, and 8atc. Alpha plus beta were 1e01, 1a5r, 1aor, 1b87, 1bkc, 1byl, 1cjw, 1drm, 1el6, 1fug, 1gyf, 1kp6, 1mol, 1nmt, 1otf, 1plq, 1qs2, 1t1d, and 2gls. Others were 1ad2, 1dkx, 1lbe, 1bgk, 1qdp, 1cq0, 1fdm, 1c94, 1dvo, 1eej, 1ezw, 1f5x, and 1fjg. Coordinate files were obtained from the Protein Data Bank.

Computer software

All programs were written by the author in C++ using Microsoft Visual C++ version 6, plus the RogueWave Class/Template libraries Tools.h++ and Math.h++. Programs were executed on a workstation with two Pentium III Xeon 1GHz processors running under the Windows NT 4.0 operating system.

Propensities

Most of the tables of propensities are too large to include in this paper. An ASCII table of the p15x6 and the r15 propensities are available from the author via e-mail request.

Acknowledgments

This work was supported by NIH grant GM34171.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Bahar, I., Kaplan, M., and Jernigan, R.L. 1997. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. *Proteins* **29:** 292–308.

Blaber, M., Zhang, X.J., and Matthews, B.W. 1993. Structural basis of amino acid alpha helix propensity. *Science* **260:** 1637–1640.

Bonneau, R. and Baker, D. 2001. Ab initio protein structure prediction: Progress and prospects. *Ann. Rev. Biophys. Biomol. Struct.* **30:** 173–189.

Bryant, S.H. and Lawrence, C.E. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins* **16:** 92–112.

Chou, P.Y. and Fasman, G.D. 1974. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry* **13:** 211–222.

Creamer T.P. and Rose, G.D. 1992. Side chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. *Proc. Natl. Acad. Sci.* **89:** 5937–5941.

Creamer, T.P., Srinivasan, R., and Rose, G.D. 1995. Modeling unfolded states of peptides and proteins. *Biochemistry* **34:** 16245–16250.

Dunbrack Jr., R.L. and Karplus, M. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230:** 543–574.

Finkelstein, A.V., Badretdinov, A.Y., and Gutin, A.M. 1995. Why do protein architectures have Boltzmann-like statistics? *Proteins* **23:** 142–150.

Flory, P.J. 1969. *Statistical mechanics of chain molecules.* Wiley, New York.

Jones, D.T. and Thornton, J.M. 1996. Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **6:** 210–216.

Kannan, N. and Vishveshwara, S. 2000. Aromatic clusters: A determinant of thermal stability of thermophilic proteins. *Protein Eng.* **13:** 753–761.

MacArthur, M.W. and Thornton, J.M. 1991. Influence of proline residues on protein conformation. *J. Mol. Biol.* **218:** 397–412.

Madej, T., Gibrat, J.-F., and Bryant, S.H. 1995. Threading a database of protein cores. *Proteins* **23:** 356–369.

Matsuo, Y and Nishikawa, K. 1993. Protein structural similarities predicted by a sequence-structure compatibility method. *Protein Sci.* **3:** 2055–2062.

Miller, S., Janin, J., Lesk, A.M., and Chothia, C. 1987. Interior and surface of monomeric proteins. *J. Mol. Biol.* **196:** 641–656.

Miyazawa, S. and Jernigan, R.L. 1985. Estimation of effective inter-residue contact energies from crystal structures. Quasi-chemical approximation. *Macromolecules* **18:** 534–552.

Moult, J. 1997. Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.* **7:** 1994–1999.

Munoz, V. and Serrano, L. 1994. Intrinsic secondary structural propensities of the amino acids using statistical phi-psi matrices: Comparison with experimental scales. *Proteins* **20:** 301–311.

Pappu, R.V., Srinivasan, R., and Rose, G.D. 2000. The Flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding. *Proc. Natl. Acad. Sci.* **97:** 12565–12570.

Pohl, F.M. 1971. Empirical protein energy maps. *Nat. New Biol.* **234:** 277–279.

Ponder, J.W. and Richards, F.M. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193:** 775–791.

Ramachandran, G.N. and Sasisekharan, V. 1968. Conformation of polypeptides and proteins. *Adv. Prot. Chem.* **23:** 283–438.

Rashin, A.A., Rashin, B.H., Rashin, A., and Abagyan, R. 1997. Evaluating the energetics of empty cavities and internal mutations in proteins. *Protein Sci.* **6:** 2143–2158.

Richardson, J.S. and Richardson, D.C. 1989. Principles and patterns of protein conformation. In *Prediction of protein structure and the principles of protein conformation* (ed. G.D. Fasman), pp. 1–98. Plenum Press, New York.

Rose, G.D., Geselowitz, A.R., Lesser, G. J., Lee, R.H., and Zehfus, M.H. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science* **229:** 834–838.

Shortle, D. and Ackerman, M.S. 2001. Persistence of native-like topology in a denatured protein in 8M urea. *Science* **293:** 487–489.

Sippl, M.J. 1993. Boltzmann's principle, knowledge-based mean fields, and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Design* **7:** 473–501.

Sippl, M.J., Ortner, M., Jaritz, M., Lackner, P., and Flockner, H. 1996. Helmholtz free energies of atom pair interactions in proteins. *Fold Design* **1:** 289–298.

Srinivasan, R. and Rose, G.D. 1999. A physical basis for protein secondary structure. *Proc. Natl. Acad. Sci.* **96:** 14258–14263.

Street, A.G. and Mayo, S.L. 1999. Intrinsic beta-sheet propensities result from van der Waals interactions between the side chains and the local backbone. *Proc. Natl. Acad. Sci.* **96:** 9074–9076.

Swindells, M.B., MacArthur, M.W., and Thornton, J.M. 1995. Intrinsic phi, psi propensities of amino acids derived from the coil regions of known structures. *Nature Struct. Biol.* **2:** 596–603.

Thomas, P.D. and Dill, K.A. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* **257:** 457–469.

Vajda, S., Sippl, M., and Novotny, J. 1997. Empirical potentials and functions for protein folding and binding. *Curr. Opin. Struct. Biol.* **7:** 222–228.

Wrabl, J.O. and Shortle, D. 1996. Perturbations of the denatured state ensemble: modeling their effects on protein stability and folding kinetics. *Protein Sci.* **5:** 2343–2352.

Yang, A.-S. and Honig, B. 1995a. Free energy determinants of secondary structure formation: I. Alpha helices. *J. Mol. Biol.* **252:** 351–365.

———. 1995b. Free energy determinants of secondary structure formation: II. Antiparallel beta-sheets. *J. Mol. Biol.* **252:** 366–376.