An Empirical Energy Function for Threading Protein Sequence Through the Folding Motif

Stephen H. Bryant,[1] and Charles E. Lawrence[2]
[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20879; [2]Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, New York 12201

ABSTRACT **In this paper we present a new residue contact potential derived by statistical analysis of protein crystal structures. This gives mean hydrophobic and pairwise contact energies as a function of residue type and distance interval. To test the accuracy of this potential we generate model structures by "threading" different sequences through backbone folding motifs found in the structural data base. We find that conformational energies calculated by summing contact potentials show perfect specificity in matching the correct sequences with each globular folding motif in a 161-protein data set. They also identify correct models with the core folding motifs of hemerythrin and immunoglobulin McPC603 V_l-domain, among millions of alternatives possible when we align subsequences with α-helices and β-strands, and allow for variation in the lengths of intervening loops. We suggest that contact potentials reflect important constraints on nonbonded interaction in native proteins, and that "threading" may be useful for structure prediction by recognition of folding motif.**

Key words: protein folding, residue contacts, conformational energy

INTRODUCTION

Proteins occur in structural families which may be distinguished by the folding motifs of their polypeptide backbone.[1,2] As a method of structure prediction, one may imagine choosing among alternative folding motifs that which represents the most likely conformation for a polypeptide with a given sequence. Prediction by motif recognition will yield an alignment of sequence and folding motif, as in homology search of a data base where three-dimensional structures are known, and might similarly provide a starting point for detailed molecular modeling.[3,4] However, by directly comparing sequence and folding motif, one may recognize similarities that are unapparent by sequence alignment, due to the relatively rapid rate of sequence evolution.[5] One may also evaluate models based on novel folding motifs, those predicted theoretically, for example, on the basis of rules governing polypeptide backbone topology.[6–9]

Prediction by motif recognition requires a computational technique that can identify alignments of sequence and folding motif that are likely to represent stable conformers. This question has been addressed in computational chemistry as a problem of distinguishing correct structures from misfolded proteins.[10,11] Attempts have been based on empirical energy functions which approximate solvent and entropic contributions to conformer stability, and it has been shown that such functions can identify grossly misfolded models.[11–15] For structure prediction, however, one must distinguish correct alignments from the best that are possible for other combinations of sequence and motif, and indeed from alternative alignments of the same sequence and motif. Local complementarity of sequence and conformation will occur by chance in many alignments, and it remains to be seen whether an empirical energy function can be sufficiently accurate to distinguish correct models from these more plausible alternatives.

In this paper, we describe an empirical energy function derived from the relative frequencies of pairwise residue interactions in proteins of known three-dimensional structure. We tabulate contacts by residue type and distance interval, and define energetic parameters by comparing observed contact frequencies to those expected for other, random sequences, when assigned the folding motifs of the known structure data base. In our definition, side chains are represented by virtual C_β coordinates, so that the contact matrix characteristic of a folding motif carries little geometric "memory" of the original sequence. We consider contacts only within distance intervals where strong van der Waals or electrostatic interactions are possible, and only between

Received August 7, 1992; revision accepted December 29, 1992.

Address reprint requests to Dr. Stephen H. Bryant, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8th Floor, 8600 Rockville Pike, Bethesda, MD 20879.

residues which are nonlocal in sequence, so that the derived potential reflects features of protein tertiary structure, rather than local correlations of sequence and secondary structure. To estimate contact potentials, we employ a log-linear or Boltzmann-like statistical model. This model defines residue-specific hydrophobicities from observed differences in contact numbers between correct and random sequence assignments for each folding motif, and pair-specific potentials as departures from the log-odds ratios expected on the basis of hydrophobicity alone. The derived potential thus places residue-residue and residue-solvent interactions on a common energetic scale.

To test whether contact potentials can be used for motif recognition, we "thread" sequences though folding motifs from the known structure data base and ask whether computed conformational energies are lowest for correct models. Similarly, we ask whether sequences from this data base have lowest energy when assigned the correct, native folding motif. We consider in detail the folding motifs of hemerythrin[16] and immunoglobulin McPC603 V_l domain,[17] structures which have served as benchmarks for identification of misfolded proteins.[10,11] Sequences are "threaded" through folding motifs representing their evolutionarily conserved core structures by gapped alignment, by assigning subsequences to individual β-strands and α-helices independently, in all ways allowed by chain connectivity. With these computer experiments, we test whether one may identify correct alignments of sequence and core folding motif from a field containing millions of incorrect alignments, including the best found for any incorrect match of sequence and motif. Success in these tests would suggest that a simple empirical potential can indeed reflect important constraints on nonbonded interaction in native proteins, and that it is possible to rapidly measure complementarity of sequence and folding motif.

METHODS

Protein Data

Atomic coordinate data are from the Protein Data Bank,[18] a distribution dated October, 1991. A set of 161 proteins with complete atomic coordinates and distinct structures are included in the analysis:

1ABP 1ACX 1AK3 1ALC 1BMV 1BP2 1CA2 1CD4 1CMS 1CSE 1CTF 1DHF 1ECD 1ETU 1F19 1FC1 1FCB 1FDL 1FX1 1GCR 1GD1 1GOX 1GP1 1GPD 1HNE 1I1B 1LAP 1LDB 1LLC 1LRD 1LZ1 1MCP 1MCW 1OMD 1P2P 1PFC 1PFK 1PHH 1PP2 1PSG 1PYP 1R1A 1R69 1RBP 1REI 1RHD 1RNH 1SGT 1SNC 1TEC 1TGN 1TIM 1TNF 1TON 1TRM 1UBQ 1UTG 1WSY 1YPI 2AAT 2ACT 2ALP 2APR 2BP2 2CAB 2CDV 2CGA 2CI2 2CPP 2CRO 2CTS 2CYP 2FB4 2FBJ 2FXB 2GBP 2GCR 2GN5 2HFL 2HLA 2HMG 2LBP 2LDX 2LIV 2LTN 2LZ2 2LZT 2MEV 2PAB 2PKA 2PLV 2PRK 2PTN 2RHE 2RSP 2SGA 2SNI 2SSI 2STV 2TAA 2TBV 2TIM 2TMV 2TRX 2TS1 2TSC 2WRP 3ADK 3APP 3BLM 3CLA 3CLN 3CNA 3DFR 3EST 3FAB 3GAP 3GBP 3GPD 3GRS 3HFM 3ICB 3LZM 3MCG 3PFK 3PGK 3PGM 3RN3 3RNT 3RP2 3SGB 3TLN 4APE 4DFR 4FAB 4FXN 4HVP 4MDH 4PEP 4RHV 4SBV 4TNC 4XIA 5ACN 5CHA 5CPA 5CPV 5CTS 5LDH 5RSA 5RUB 5TNC 6LDH 6XIA 7API 7CAT 8ADH 8ATC 8DFR 9PAP 9WGA.

This set excludes small proteins with large prosthetic groups and/or many disulfide bonds, where cross-links may play a role in specific folding. Hemerythrin (1HMQ) and homologs are excluded by this criterion. The potential used in tests with McPC603 V_l domain (1MCP) is parameterized from a subset that excludes proteins with immunoglobulin V-domains:

1CD4 1F19 1FDL 1MCP 1MCW 1REI 2FB4 2FBJ 2HFL 2RHE 3FAB 3HFM 3MCG 4FAB.

Use of this subset introduces a small shift in energetic parameters (median $\Delta\mu_{rsd} = 0.028$, see below), and does not affect correct identification of sequences or motifs for 1MCP, 1HMQ, or other test proteins. Results below are also little affected by use of further subsets which exclude other homolog groups and/or test proteins (not shown).

For estimation of energetic parameters, each protein is reduced to a simplified representation of two summary coordinates per residue. Peptide groups of the protein backbone are designated by a point midway between sequentially adjacent C_α atoms, and side chains by a point 2.4 Å from C_α, in the direction C_β-C_α. In the case of glycine, the expected position of C_β is derived from backbone coordinates, assuming standard geometry. This side chain coordinate corresponds to the mean projection of side chain centroids onto C_β-C_α, taken across all nonglycine residues in the current data set. It is effectively a function of backbone coordinates, and as such contains no explicit information on the residue types observed at each position in native structures.

For each protein, pairwise contacts identified from summary coordinates are counted according to chemical type and distance interval. The 20 side chain types plus the peptide group generate 231 distinct pair types, corresponding to the lower triangle and diagonal of a 21 × 21 array. Distance is classified by intervals 0–5; >5–6; >6–7; >7–8; >8–9; and >9–10 Å. Use of finer intervals leads to only small changes in likelihood ratio statistics g^2 (see below), suggesting that the present choice is adequate. Data for the entire protein set take the form of a four-dimensional contingency table, with cell (Arg, Glu, >5–6, 3FAB), for example, containing the number of Arg-Glu side chain interactions, within interval >5–6 Å, for protein 3FAB. The total number of interaction categories is (231 × 6) = 1,386, and the total number of observed contacts in the 161-structure set is 1,302,669. Each parameter is thus determined by approximately 1,000 contact observations.

Only nonlocal pairwise interactions are included in the contact tabulation, with contacts between residues whose sequence indices differ by less than 5 being omitted. Correct alignment of sequences and motifs for 1HMQ, 1MCP, and other test proteins is unaffected as this criterion is varied between 3 and 9 (not shown). The ratio of local to nonlocal contacts is in each case small, and it appears that the local contacts have little influence on the derived potential. Contacts to prosthetic groups are also omitted from the contact tabulation, as are residue contacts mediated by disulfide bonds or covalent linkage to a common prosthetic group. These omissions reflect our intention to derive an empirical potential for nonlocal, nonbonded interaction of groups common in proteins. Disulfide bonding and interactions with prosthetic groups may be addressed separately, as additional energetic constraints on alignment of sequence and folding motif.

Contacts of protein groups with solvent molecules cannot be tabulated directly, due to variability in protein crystal environments and solvent modeling. The functional groups we consider have approximately equal volumes, however, 4 atoms for peptide groups and 4.1 atoms for side chains, on average. One may assume that they have similar coordination geometry with respect to nonbonded interactions, and hence that any statistical tendency towards fewer contacts indicates preferred interaction with surrounding solvent.[13,19,20] Below, we estimate "hydrophobicity" parameters, which summarize this tendency. One may alternatively assume fixed coordination numbers, impute solvent contacts, and include these in the contingency table as distinct interaction categories.[19] An analysis carried out in this fashion leads to similar specificity in alignment of sequence and folding motif for several test proteins (not shown).

Contact matrices were prepared using the PKB database and function library,[21] operated on a Silicon Graphics 4D-35 workstation. PKB is available from the authors.

Estimation of Energetic Parameters

We compare observed contact frequencies for each protein to those that would be expected for random sequences of the same length and composition, when assigned the folding motif of that protein. By conditioning on amino acid composition, we define contact preferences as departures from the frequencies expected by mass action.[6,13,19,22–24] We also assume that the structures in the data set best represent stable, globular proteins when considered in their reported oligomerization states. Pairwise contacts are those of the complete, native structure, and the random sequence reference state is defined in terms of the overall folding motif and composition of each molecule, regardless of how many chains are present.

To derive energetic parameters, we assume that the relative probabilities of pairwise contact follow a Boltzmann-like law:[25,26]

$$\frac{p(R_i^p = r, R_j^p = s|\overrightarrow{X}_{ij}^p)}{p^o\,(R_i^p = r, R_j^p = s|\overrightarrow{X}_{ij}^p)} = \exp\,(-\mu\,(r, s, \overrightarrow{X}_{ij}^p)). \quad (1)$$

Here $p\ (R_i^p = r, R_j^p = s|\overrightarrow{X}_{ij}^p$ and $p^\circ\ (R_i^p = r, R_j^p = s|\overrightarrow{X}_{ij}^p$ are probabilities that sites i and j will be occupied by side chain types r and s, in correct versus random sequences for the folding motif of protein p. $\overrightarrow{X}_{ij}^p$ are relative coordinates of sites i and j. Log odds are given by energy function μ (r, s, $\overrightarrow{X}_{ij}^p$), a potential of mean force in dimensionless kT units. Here we consider only the distance between residue sites, categorized by interval, and relative coordinates $\overrightarrow{X}_{ij}^p$ reduce to a set of contact matrices X_{ijd}^p. Their elements are either one or zero, indicating that sites i and j do or do not fall within distance interval d, and are not local to one another in the polypeptide chain. For this case (1) may be abbreviated:

$$\frac{p(r(i), s(j)|d, p)}{p^0\,(r(i), s(j)|d, p)} = \exp\,(-\,X_{ijd}^p\,\mu_{rsd}). \quad (2)$$

μ_{rsd} is now a contact potential, an average of μ $(r, s, \overrightarrow{X}_{ij}^p)$ across distance interval d. The log odds of observing residue type r and s at sites i and j are given by μ_{rsd} if they are in contact, and are zero otherwise. This set of contact potentials is assumed to be universal, the same for all proteins p, and the same for all residue sites i and j.

We also assume that the likelihood of a sequence assignment may be written as a product over the probabilities of individual pairwise interactions, that pair potentials μ_{rsd} are additive. Since contact matrices X_{ijd}^p depend only on backbone summary coordinates, the log likelihood of the native sequence of a protein relative to that expected for random sequences may be written:

$$\log \frac{\mathrm{p}(\mathrm{R^p}|\mathrm{X}_{ijd}^p)}{\mathrm{p}^0\,(\mathrm{R^p}|\mathrm{X}_{ijd}^p)} = -\sum_{\mathrm{i}}\sum_{\mathrm{j}}\sum_{\mathrm{d}} \mathrm{X}_{ijd}^p\,\mu_{r(i)s(j)d} =$$

$$-\sum_{\text{contacts } i,j,d} \mu_{r(i)s(j)d}. \quad (3)$$

R^p denotes the native sequence, a vector of side chain types $r(i)$ or $s(j)$ assigned to each site i or j of the folding motif. Selection of sequence R^p given folding motif X_{ijd}^p is assumed to follow a Boltzmann-like probability law with respect to the total energy of pairwise interaction.

Each structure in the available data set may be viewed as a sample from the hypothetical selection process described by (3). Each provides information on the relative probabilities of pairwise interaction $p\ (r,s|d,p)/p^0\ (r,s|d,p)$, in the form of relative contact counts N_{rsdp}/N^0_{rsdp}. Here N_{rsdp} are observed contact counts from the contact contingency table described above, and N^0_{rsdp} are contact counts expected for randomly permuted sequences R^p, assigned to motif X^p_{ijd}. They are defined by mass-action formulae:

$$N^o_{rsdp} \equiv \begin{cases} m^p_r m^p_s \times \sum_i \sum_j X^p_{ijd} & i, r \neq peptide, j, s \neq peptide \\ m^p_s \times \sum_i \sum_j X^p_{ijd} & i, r \neq peptide, j, s = peptide \\ m^p_r \times \sum_i \sum_j X^p_{ijd} & i, r \neq peptide, j, s = peptide \\ \sum_i \sum_j X^p_{ijd} & i, r = peptide, j, s = peptide \end{cases} \qquad (4)$$

where m^p_r and m^p_s are the mole fractions of residues of types r and s in protein p, and sums taken over X^p_{ijd} given the total numbers of contacts involving side chain and/or peptide sites. We assume in the analysis that each protein should contribute equally to estimation of best-fit values μ_{rsd}, and we thus weight individual contact observations by $1/N^0_{rsdp}$. This weighting prevents proteins which are large or rich in particular residue types from unduly affecting the estimates, and leads to improved specificity in alignment of sequence and motif for several test proteins (not shown).

Equation (3) corresponds to a well studied statistical model, the log-linear model for categorical data, and numerical methods for finding maximum likelihood parameter estimates are well known.[27-29] With weighting of observed counts by $1/N^0_{rsdp}$ we obtain maximum likelihood estimates by the method of iterative proportional fits,[27] applied to a table of relative counts N_{rsdp}/N^0_{rsdp}. We use a double-precision FORTRAN subroutine from the IMSL subroutine library, and extract μ_{rsd} *from the fitted table as* suggested by Feinberg,[27] using S[30] and FORTRAN subroutines. When any m^p_r or m^p_s is zero the corresponding cells in the table of relative counts are defined as a structural zero.[27] A protein with no residues of type r or s thus provides no information on contact preferences involving these residues, and μ_{rsd} is derived solely from data for the remaining proteins.

The log-linear model allows us to decompose contact potentials into a series of terms giving increments in log probability associated with individual categories:

$$\mu_{rsd} = \mu'_r + \mu'_s + \mu'_{rd} + \mu'_{sd} + \mu'_{rs} + \mu'_{rsd}. \qquad (5)$$

Here μ'_r and μ'_s give increments in log probability associated with the first and second residues of a pair being of types r and s, and μ'_{rd} and μ'_{sd} their variation with distance interval. These are interpreted as distance-dependent hydrophobicities:

$$\mu^H_{rd} = \mu'_r + \mu'_{rd}$$
$$\mu^H_{sd} = \mu'_s + \mu'_{sd}. \qquad (6)$$

We note that μ^H_{rd} and μ^H_{sd} are equal for a given residue type, since the data from which they are derived are symmetrical with respect to pairwise contacts. Parameters μ'_{rs} and μ'_{rsd} give further increments in log probability associated with residue pairs of a given type and their variation with distance interval. These define the pairwise component of the contact potential as departures from additivity of distance and residue-specific hydrophobicities:

$$\mu^P_{rsd} = \mu'_{rs} + \mu'_{rsd}. \qquad (7)$$

The derived contact potential is thus interpreted as a sum of two hydrophobicity terms and a pair-specific interaction term:

$$\mu_{rsd} = \mu^H_{rd} + \mu^H_{sd} + \mu^P_{rsd}. \qquad (8)$$

We note that the hierarchical model (5) does not imply orthogonality of parameters μ^H_{rd} and μ^P_{rsd}, but this interpretation is justified since μ^H_{rd} change very little as higher-order terms within μ^P_{rsd} are added in stepwise fashion.

Statistical significance of individual terms in (5) is assessed by means of likelihood ratio tests. The change in the statistic g^2 as new terms are added stepwise is asymptotically chi-squared, with degrees of freedom equal to the number of new independent parameters.[27,28] The agreement with the data achieved by the model may also be partitioned into contributions from individual terms by expressing the associated change in g^2 as a percentage of the total.[27] In conducting chi-squared tests, we allow for symmetry in the data and parameters with respect to contacts involving residues of types r, s and s, r.

Complementarity of Sequence and Folding Motif

Folding motifs are represented by residue contact matrices as described above. A sequence is threaded through a motif by considering alternative alignments of residues from the sequence with side chain sites from the motif. Each alignment represents a distinct model structure, and conformational energy may be calculated by summing contact potentials:

$$G_{R|M} = \sum_{\text{contacts } i,j\ d \in X^M_{ijd}} \mu_{r(i)s(j)d}. \qquad (9)$$

Here i and j indicate residue sites of folding motif M, and d the distance interval characteristic of that contact pair. $r(i)$ and $s(j)$ represent the residue types assigned to sites i and j, given the assigned sequence R. Contact potentials properly define the difference in $G_{R|M}$ and its expected value for random sequences with the same composition:

$$\Delta G_{R|M} = G_{R|M} - G^{0}_{R|M}. \tag{10}$$

The reference state $G^{0}_{R|M}$ corresponds to that used in parameterization of μ_{rsd}. It is defined as a sum over random contact frequencies using mass-action formulae (4), where residue mole fractions are now derived from the assigned sequence, and contact matrices from the folding motif. The expected value of the conformational energy $\Delta G_{R|M}$ is zero for a random sequence, regardless of the motif considered or the amino acid composition of that sequence.

For "core" folding motifs, we assign sequences by "gapped" alignment. Subsequences are aligned with specific side chain sites for segments of core secondary structure, but subsequences assigned to loops may vary in length, and are not aligned with specific sites. To compute conformational energies, we therefore omit contacts involving loop residues from contact matrices X^{M}_{ijd}, and carry out the summations (10) for core-core contacts only. The subsequences assigned to loops affect $\Delta G_{R|M}$, but only via the reference state $G^{0}_{R|M}$, inasmuch as they may concentrate hydrophobic residues in core segments, and lead to more favorable core-core interactions than one would expect for random assignment of core subsequences. In an alternative calculation of $G_{R|M}$ we further assume that loops shield the protein core from solvent to the same extent as do the loops in the known, complete backbone structure of core motif M:

$$G^{L}_{R|M} = \sum_{\text{core-core contacts } i,j,d} \mu_{r(i)s(j)d} + \sum_{\text{loop-core contacts } i,j,d} \mu^{H}_{r(i)d} + \bar{\mu}^{H}_{l(j)d}$$

$$\bar{\mu}^{H}_{l(j)d} \equiv \begin{cases} \displaystyle\sum_{\kappa \in \text{loop subsequence } l} \mu^{H}_{s(k)d}/L, & j, s \neq \text{peptide} \\ \mu^{H}_{s(j)}, & j, s = \text{peptide}. \end{cases} \tag{11}$$

Here L is the length of a given loop subsequence l, and $\bar{\mu}^{H}_{ld}$ is the mean hydrophobicity for side chains assigned to this loop. Contact potentials for loop-core contacts are defined from the hydrophobicity component of the contact potential for the core residue of each contact, $\mu^{H}_{r(i)d}$, and by an average hydrophobicity for the loop residue. Conformational energies $\Delta G^{L}_{R|M}$ are calculated by (10), where the reference state energy now allows for fractional occupancy of loop-residue sites as implied by (11). We note that neither of these calculations employs any gap penalty. The lengths of subsequences assigned to loops may be constrained to fall within chosen limits, but $\Delta G_{R|M}$ and $\Delta G^{L}_{R|M}$ vary with the composition of loop subsequences, not with their length.

To compare conformational energies of different model structures, we allow for the effects of sequence composition and length on the statistical distribution expected by chance, for random alignments of sequence and folding motif. The distribution of $\Delta G_{R|M}$ (or $\Delta G^{L}_{R|M}$) is normal, as expected from the central limit theorem and confirmed by simulation (not shown). Its mean is zero by definition, but its variance is a function of amino acid composition, as residues with greater or lesser absolute values μ_{rsd} may be present in a particular sequence assignment. To compare conformational energies for different assignments, we therefore compute Z-scores $Z_{R|M}$, which give in standard deviation units the departures of $\Delta G_{R|M}$ from the means of their random sequence distributions. Variances are calculated from conformational energies for 1,000 random permutations of R, a number sufficient for accurate determination (not shown). Sequence length affects Z-scores only inasmuch as more alternative assignments are considered when we thread a longer sequence through a motif, and the maximum Z-score expected by chance will be greater, simply due to the larger number of trials. Probabilities of observing a maximum Z-score by change within a certain number of trials are given by order statistics, which are well characterized for the standard normal distribution, and we may thus compute chance occurrence probabilities $E_{R|M}$ analytically,[31] using an S function.[30] We note that this use of Z-scores and order statistics is directly analogous to well known procedures in sequence alignment, where they similarly correct for effects of composition and length on the homology scores expected by chance in alignment of random sequences.[32]

In the threading trials below, we exhaustively enumerate all possible alignments of each sequence and motif. By enumeration, we exclude any possibility of missing favorable alignments, as may occur when more rapid, heuristic algorithms for optimal threading are employed. Enumeration of possible sequence alignments and evaluations of $\Delta G_{R|M}$ and related quantities are carried out by S and FORTRAN subroutines, which are available upon request from the authors. For motifs the size of 1HMQ or 1MCP, approximately 500,000 evaluations of $\Delta G_{R|M}$ may be performed per hour on a Silicon Graphics 4D-35 workstation.

RESULTS

Residue Contact Potential

The contact potential we derive may be viewed as the logs of pseudo-equilibrium constants relating pairwise contact frequencies in correctly versus incorrectly folded proteins. Its derivation differs from previous analyses[6,13,19,22–24,33,34] in the definition of summary coordinates for side chain and peptide groups, and in treatment of distance dependence. A major difference, however, is in the definition of the reference state, the contact frequencies assumed for incorrectly folded proteins. We treat each protein as

TABLE I. Likelihood Ratio Statistics for Stepwise Log-Linear Modeling of Relative Contact Frequencies

Model*	Components	g^2	df	Δg^2	Δdf	$p^{\dagger}$	% $\Delta g^{2\ddagger}$
	None	424163.9	210845				
1	μ_r', μ_s'	399975.4	210820	24188.5	25	<.001	62.8
2	$+\ \mu_{rd}', \mu_{sd}'$	397865.2	210720	2110.2	100	<.001	5.5
3	$+\ \mu_{rs}'$	392518.7	210511	5347.5	209	<.001	13.9
4	$+\ \mu_{rsd}'$	385641.6	209467	6877.1	1044	<.001	17.9
Total				38522.3	1378		100.0

*Model 1 assumes residue-specific hydrophobicity only, and model 2 adds distance dependence. Model 3 adds pairwise contact preferences, constant over the interval 0–10 Å, and model 4 assumes distance-dependent pairwise contact preferences.

†Column p indicates the probability of obtaining the reduction Δg^2 by chance, given the number of additional parameters Δdf that have been introduced.

‡%Δg^2 indicates the proportion of the overall goodness-of-fit that may be attributed to the components added in each step, as derived from the change in statistics g^2 between successive models.

a separate "evolutionary experiment," and compare observed contacts to those expected for other random sequences if they adopted the conformations of proteins in the data base. This parameterization is motivated by the hypothesis that natural selection of protein sequences results in a Boltzmann-like distribution of nonbonded interactions.[25,26]

Use of a log-linear model for categorical data also allows us to partition the derived potential into components whose interpretation and contribution to overall goodness-of-fit may be assessed separately. Likelihood ratio statistics summarizing this decomposition are listed in Table I. They indicate that the most important components in the potential are residue-specific "hydrophobicities," μ_{rd}^H. Their distance-independent components account for or 62.8% of the overall reduction in the goodness-of-fit measure g^2, with addition of distance dependence adding 5.5%. These improvements are highly significant, as indicated by the associated p values. Residue-specific hydrophobicities thus account for approximately two-thirds of the apparent specificity in pairwise residue interaction. Pairwise potential components μ_{rsd}^P contribute the remaining one-third of the overall goodness-of-fit achieved by the model. Their distance-independent components account for 13.9% of the overall reduction in g^2, and distance dependence adds 17.9%. These improvements are again highly significant. This result indicates that hydrophobicities alone are insufficient to account for the apparent specificity in pairwise interaction, even at the low resolution implied by a representation of two summary coordinates per residue and discrete, 1-Å distance intervals.

Hydrophobicity components μ_{rd}^H literally give the log odds of a residue type contacting other residues of any type, in correct versus random sequence assignments. They are equivalent to the log odds of each residue type not contacting surrounding solvent, if one assumes a fixed coordination geometry, common to all residue types. Values are listed in Table IIa. Hydrophobicities for aliphatic and aromatic side chains are the most negative, indicating preference for contact with other residues, while values for oxygen and nitrogen-containing side chains are most positive. Correlation of the distance-independent component μ_r' with hydrophobicities derived from solvent accessibility in known structures[35] is 0.84, indicating the general similarity of measures derived from contact numbers and solvent accessibilities.[20,36] Distance dependence appears to reflect differences in size of aliphatic and aromatic residues; values for alanine and valine are minimum within the interval 0–5 Å, for example, but within >7–8 Å for phenylalanine.

Pairwise potential components μ_{rsd}^P give the log odds of different contact types in correct versus random sequence assignments as departures from the log odds predicted by residue-specific hydrophobicities alone. Values are listed in Table IIb. As a potential of mean force, these reflect a variety of factors distinguishing native and nonnative protein structures, and they do not have simple interpretations in terms of elementary forces. Differences in pairwise potentials often agree with what one would expect from the chemistry of nonbonded interaction, however. Two examples are shown in Figure 1. Differences associated with Asn-to-Asp substitution in the vicinity of charged neighbor groups suggest electrostatic interaction: this difference is negative when the neighbor group is positively charged, and vice versa, and approaches zero with increasing distance. This potential difference agrees with an earlier, detailed analysis of ion-pair interaction.[26] Potential differences associated with Val-to-Thr substitution suggest the energetic costs of the methyl-to-hydroxyl exchange in different environments: this difference is positive in the vicinity of an apolar neighbor Val, an unfavorable environment for the hydroxyl group, and negative in the vicinity of the polar neighbor Asp, where favorable hydrogen bonding is possible, and both differences approach zero as distance increases to 10 Å.

To compare model structures, we consider differ-

TABLE IIa. Hydrophobic Potential*

	0–5	5–6	6–7	7–8	8–9	9–10
A	−0.241	0.079	−0.017	−0.067	−0.074	−0.072
R	0.370	0.201	0.137	0.111	0.062	0.095
N	0.325	0.222	0.260	0.182	0.238	0.172
D	0.378	0.444	0.275	0.223	0.193	0.215
C	−0.327	−0.380	−0.271	−0.264	−0.427	−0.279
Q	0.291	0.211	0.240	0.178	0.192	0.123
E	0.465	0.393	0.348	0.290	0.238	0.219
G	−0.008	0.184	0.044	0.075	0.082	0.108
H	−0.045	0.031	−0.044	−0.084	−0.049	−0.067
I	−0.290	−0.444	−0.252	−0.257	−0.233	−0.266
L	−0.133	−0.352	−0.277	−0.143	−0.196	−0.187
K	0.498	0.480	0.439	0.321	0.283	0.253
M	−0.197	−0.136	−0.335	−0.243	−0.223	−0.168
F	−0.257	−0.230	−0.315	−0.382	−0.213	−0.212
P	−0.017	0.218	0.062	0.122	0.143	0.119
S	0.101	0.147	0.183	0.101	0.117	0.078
T	0.060	0.084	0.114	0.073	0.034	0.031
W	−0.180	−0.159	−0.247	−0.216	−0.239	−0.153
Y	−0.224	−0.228	−0.236	−0.248	−0.178	−0.180
V	−0.286	−0.275	−0.175	−0.129	−0.217	−0.207
p	0.024	−0.052	−0.008	−0.006	0.000	−0.015

*Pairwise potential components μ^{P}_{rsd}. Residue types follow standard 1-letter codes, with the peptide group indicated by lower-case "p." Distance intervals are 0–5; >5–6; <6–7; <7–8; <8–9; and <9–10 Å.

ences in the sum of contact potentials. Since differences in contact potentials have plausible physical interpretations, it seems reasonable to interpret these sums as a conformational energy associated with nonlocal, nonbonded interaction. This approximates a free energy, inasmuch as we derive contact potentials as a potential of mean force, and the structural data base represents an "evolutionary ensemble" which may be similar to a true conformational ensemble.[25,26] We specifically compare quantities $\Delta G_{R|M}$ (or $\Delta G^{L}_{R|M}$), the expected work for substitution of a specific sequence R for a random sequence with the same composition, in the context of folding motif M.

Threading Sequence Through Complete Domain Folding Motif

To test whether contact potentials can identify the sequence most compatible with a folding motif, we generate many incorrect sequence assignments for each complete domain folding motif in the known structure data set. Sequences are assigned by threading the sequences of other proteins through each motif. For the 1HMQ motif, for example, we draw all subsequences of length 113 from the sequences of other proteins. While ungapped alignment cannot detect homologous sequences with insertions or deletions, it allows us to test the specificity of "self" recognition for many folding motifs in reasonable computer time.

Results of threading various sequences through the folding motifs of 1HMQ and 1MCP are shown in Figure 2. The models generated in this way are misfolded proteins as originally suggested by Novotny and colleagues,[10,11] and the set includes assignments of the 1HMQ sequence to the 1MCP motif and vice versa. It may be seen from the distibutions that conformational energies for incorrect models span a range of values, indicating varying degrees of complementarity of sequence and motif. Conformational energies for the correct models are removed from the distributions, however, and are lower than the energies of any incorrect model. Probabilities of chance occurrence $E_{R|M}$ further distinguish the correct models. The energies observed for the correct models are roughly 100,000 times less likely to have arisen by chance than those for the best models derived from an incorrect sequence. Differences in the pairwise interactions of correct structures and the misfolded models considered by Novotny and colleagues[10,11] are illustrated in Figure 3. It may be seen that misfolded models are most easily distinguished by their lack of favorable hydrophobic interactions within the protein interior.

Results of threading various sequences through each complete-domain folding motif in the data set are summarized in Figure 4. In the plot folding motifs are indexed in order of decreasing conformational energy for the correct model, values which show a correlation coefficient of 0.93 with the total number of contacts in the motif. This correlation indicates that energies per contact are roughly constant for known structures, and that $\Delta G_{R=native|M}$ increases regularly with molecular weight. It may be seen that conformational energies for correct models are generally lower than those for the best incorrect models, and that this difference is greatest for the folding motifs of large proteins. This pattern indicates that contact potentials generally identify the correct sequence among thousands of alternatives, and that one may expect greatest specificity for recognition of large folding motifs, where the number of pairwise contacts is greatest.

Seven folding motifs in Figure 4a have lowest conformational potential when assigned an incorrect sequence. They fall in the lower-left region of the plot, indicating that they are among the smallest motifs, with the fewest non-bonded contacts, and they are also distinguished by their occurrence as subunits within macromolecular complexes. It may be seen in Figure 4b that only one of these incorrect-sequence models is a false positive, with chance occurrence probability $E_{R|M}$ lower than that of the corresponding correct-sequence model. This is an alternative sequence for motif 2PLV-4, subunit VP4 from polio virus. This subunit is generated by proteolytic cleavage of the capsid polyprotein,[37] and it has an extended, aglobular structure when viewed in isolation. This is an extreme example, but it is reasonable to assume that the sequences for each of these 7 motifs have features which contribute to specific docking interactions in their native oligomers,

TABLE IIb. Pairwise Potential*

	A-A	R-A	R-R	N-A	N-R	N-N	D-A	D-R	D-N	D-D	C-A	C-R	C-N
0-5	0.230	0.237	-0.145	0.159	-0.303	-0.206	0.140	-0.527	-0.534	0.539	0.159	0.174	0.105
5-6	0.350	0.338	0.024	-0.170	-0.258	-0.372	0.068	-0.618	-0.481	0.128	0.006	0.183	0.616
6-7	0.223	0.101	0.215	0.042	0.219	0.221	0.178	-0.609	-0.403	-0.006	-0.037	0.242	0.510
7-8	0.284	0.266	0.249	0.010	-0.331	-0.360	-0.026	-0.243	-0.144	0.036	-0.110	0.001	-0.101
8-9	0.020	0.068	0.195	-0.071	0.115	0.015	0.058	-0.266	-0.321	0.021	0.094	-0.172	0.033
9-10	0.017	0.050	-0.035	-0.150	-0.086	-0.194	-0.121	0.130	-0.075	0.033	0.089	-0.081	-0.044

	C-D	C-C	Q-A	Q-R	Q-N	Q-D	Q-C	Q-Q	E-A	E-R	E-N	E-D	E-C
0-5	-0.137	-0.524	-0.008	0.222	-0.187	-0.128	0.487	-0.430	0.182	-0.717	-0.262	-0.178	0.121
5-6	0.788	0.841	-0.029	-0.297	-0.224	-0.194	-0.158	-0.001	0.194	-0.594	-0.669	0.301	-0.007
6-7	0.509	1.422	0.059	-0.084	-0.296	-0.241	-0.491	0.316	0.155	-0.602	-0.293	-0.123	0.106
7-8	0.456	0.459	-0.030	-0.022	-0.201	-0.143	0.058	0.039	-0.031	-0.112	-0.138	-0.173	-0.117
8-9	0.130	0.507	0.190	-0.273	0.080	-0.043	0.017	-0.200	0.092	-0.083	-0.235	0.248	0.422
9-10	-0.024	1.329	-0.066	-0.222	-0.125	0.001	0.327	0.184	0.117	-0.412	-0.123	-0.157	0.052

	E-Q	E-E	G-A	G-R	G-N	G-D	G-C	G-Q	G-E	G-G	H-A	H-R	H-N
0-5	-0.728	1.060	0.282	-0.037	-0.097	-0.336	-0.187	-0.072	-0.041	-0.068	-0.079	0.111	-0.141
5-6	0.081	-0.216	-0.165	-0.236	0.271	-0.157	0.094	0.086	0.020	-0.188	-0.052	-0.013	-0.219
6-7	0.087	-0.429	-0.099	-0.089	-0.178	-0.166	0.028	-0.177	0.164	-0.210	0.223	0.255	-0.325
7-8	-0.200	0.501	0.059	-0.102	-0.078	-0.167	-0.082	-0.125	0.141	-0.298	-0.046	-0.257	-0.196
8-9	-0.116	-0.086	-0.061	-0.087	-0.170	-0.106	0.105	-0.062	-0.052	-0.122	0.132	-0.238	-0.039
9-10	-0.052	0.062	-0.017	0.017	0.074	-0.067	-0.124	-0.065	-0.089	-0.113	0.183	0.149	0.084

	H-D	H-C	H-Q	H-E	H-G	H-H	I-A	I-R	I-N	I-D	I-C	I-Q	I-E
0-5	-0.270	-0.187	0.247	0.072	0.127	0.392	-0.535	-0.014	0.456	0.229	-0.415	0.020	0.385
5-6	-1.299	-0.421	0.374	-0.166	0.077	0.740	-0.386	-0.111	0.881	0.099	-0.161	0.375	0.537
6-7	-0.189	-0.085	-0.311	-0.616	0.026	0.818	0.108	0.037	0.570	0.275	-0.262	0.059	0.464
7-8	-0.266	0.322	0.056	-0.429	0.247	0.528	-0.041	-0.040	0.199	0.447	-0.298	0.239	0.158
8-9	-0.123	-0.400	-0.034	-0.180	-0.144	0.104	-0.094	0.181	0.081	0.293	-0.091	0.027	-0.019
9-10	-0.067	-0.378	0.067	-0.003	0.072	-0.067	0.032	0.116	0.079	0.146	-0.283	0.078	0.122

	I-G	I-H	I-I	L-A	L-R	L-N	L-D	L-C	L-Q	L-E	L-G	L-H	L-I
0-5	0.127	-0.062	-0.051	-0.245	0.374	0.430	0.316	-0.061	0.391	0.480	-0.004	0.029	-0.223
5-6	0.585	0.060	-0.531	-0.232	0.042	0.325	0.141	-0.072	-0.022	0.286	0.154	0.196	-0.518
6-7	0.268	0.437	-0.698	-0.163	0.113	0.490	0.501	-0.259	0.061	0.308	0.311	0.018	-0.702
7-8	0.114	0.250	-0.238	0.047	0.288	0.360	0.322	-0.317	0.266	0.179	0.115	0.207	-0.347
8-9	0.081	0.233	0.225	0.047	0.050	0.114	0.237	0.031	-0.015	0.084	0.174	0.136	-0.071
9-10	0.181	-0.276	0.092	-0.034	0.052	0.077	0.124	-0.093	-0.054	0.073	0.066	0.252	-0.030

	L-L	K-A	K-R	K-N	K-D	K-C	K-Q	K-E	K-G	K-H	K-I	K-L	K-K
0-5	-0.070	0.063	0.734	0.222	-0.759	1.115	-0.181	-0.782	0.065	-0.235	-0.058	-0.015	0.567
5-6	-0.265	0.429	0.643	-0.124	-0.674	-0.279	-0.173	-0.897	0.233	0.620	-0.075	0.142	0.277
6-7	-0.594	-0.044	0.592	-0.052	-0.615	0.089	0.222	-0.736	-0.325	0.369	0.086	0.278	0.340
7-8	-0.542	-0.016	0.020	-0.044	-0.369	0.021	-0.225	-0.382	-0.199	0.320	0.321	0.342	-0.138
8-9	-0.124	-0.018	0.118	-0.002	-0.263	0.001	0.051	-0.493	0.077	0.188	-0.001	0.072	-0.072
9-10	0.003	0.029	0.211	-0.115	0.044	-0.014	-0.059	-0.118	0.017	-0.039	0.086	-0.017	-0.262

	M-A	M-R	M-N	M-D	M-C	M-Q	M-E	M-G	M-H	M-I	M-L	M-K	M-M
0-5	-0.364	0.509	0.405	0.097	-0.032	0.013	0.188	0.084	0.293	-0.103	-0.253	0.347	-0.006
5-6	0.129	0.324	-0.120	0.044	-0.307	-0.014	0.163	0.126	0.918	-0.209	-0.171	0.188	-0.280
6-7	0.033	-0.528	-0.144	0.525	-0.177	0.543	0.187	0.276	0.452	-0.221	-0.231	-0.089	-0.454
7-8	-0.262	0.262	0.152	0.121	0.163	0.005	0.611	0.062	-0.137	-0.314	-0.178	0.138	0.473
8-9	-0.172	0.058	0.006	0.495	-0.224	0.311	0.155	0.026	-0.016	-0.317	-0.059	0.123	0.332
9-10	-0.087	0.085	0.139	0.035	0.095	0.284	0.230	-0.034	-0.330	-0.247	-0.118	0.105	0.283

	F-A	F-R	F-N	F-D	F-C	F-Q	F-E	F-G	F-H	F-I	F-L	F-K	F-M
0-5	0.132	-0.018	0.336	0.118	-0.095	0.283	0.270	0.457	-0.141	-0.348	-0.460	0.259	-0.501
5-6	-0.378	0.036	0.453	0.703	-0.407	0.252	0.048	0.071	-0.191	-0.313	-0.184	-0.279	-0.087
6-7	-0.244	-0.157	0.362	0.237	-0.439	0.273	0.402	0.357	0.081	-0.326	-0.333	0.020	-0.438
7-8	-0.002	0.014	0.328	0.270	-0.031	0.099	0.069	0.378	-0.334	-0.304	-0.233	0.157	-0.288
8-9	-0.005	0.194	0.130	0.239	-0.055	0.188	0.205	-0.065	-0.198	-0.269	-0.198	0.255	-0.225
9-10	0.007	-0.043	0.031	-0.022	-0.175	0.011	0.186	0.036	0.237	-0.061	-0.201	0.119	-0.153

	F-F	P-A	P-R	P-N	P-D	P-C	P-Q	P-E	P-G	P-H	P-I	P-L	P-K
0-5	-0.163	0.022	-0.176	0.109	0.175	-0.199	-0.545	-0.064	0.127	0.037	0.382	-0.170	0.068
5-6	-0.162	0.128	0.243	-0.050	0.264	0.777	-0.496	0.130	-0.133	-0.242	-0.035	-0.047	0.049
6-7	-0.198	0.080	-0.200	-0.105	0.065	-0.100	0.044	-0.346	-0.293	-0.057	-0.088	0.396	-0.279
7-8	-0.224	-0.090	0.093	-0.062	-0.073	0.222	0.228	-0.316	0.059	-0.265	0.133	-0.054	-0.055
8-9	-0.040	-0.065	-0.589	-0.047	-0.337	0.180	-0.076	-0.225	-0.053	0.466	0.145	-0.058	0.044
9-10	-0.127	0.000	0.052	-0.211	-0.027	-0.043	-0.118	0.008	0.033	-0.167	0.196	0.034	-0.050

(continued)

TABLE IIb. Pairwise Potential (continued)

	P-M	P-F	P-P	S-A	S-R	S-N	S-D	S-C	S-Q	S-E	S-G	S-H	S-I
0-5	-0.104	-0.083	0.033	0.115	-0.263	-0.334	-0.565	-0.164	0.298	-0.207	-0.100	0.080	0.254
5-6	-0.100	-0.018	0.132	0.187	0.121	-0.358	-0.339	-0.188	-0.175	-0.249	-0.076	-0.365	0.411
6-7	0.096	0.458	0.134	-0.116	0.126	-0.585	-0.199	-0.199	-0.126	0.077	-0.255	-0.336	0.476
7-8	-0.252	0.143	0.088	0.056	-0.115	0.065	-0.036	-0.220	0.117	-0.044	-0.166	-0.215	0.133
8-9	-0.167	-0.093	0.281	-0.069	0.011	-0.133	-0.043	-0.015	-0.104	0.003	0.006	-0.071	0.011
9-10	0.023	0.203	-0.124	-0.051	0.105	0.021	-0.030	-0.353	0.083	-0.066	0.012	-0.006	-0.012

	S-L	S-K	S-M	S-F	S-P	S-S	T-A	T-R	T-N	T-D	T-C	T-Q	T-E
0-5	0.390	-0.246	0.259	0.357	0.057	-0.150	0.002	0.250	-0.018	-0.433	0.021	0.060	-0.006
5-6	0.308	-0.099	0.089	0.162	-0.087	0.021	-0.197	-0.187	-0.121	-0.124	0.206	-0.093	0.094
6-7	0.451	0.107	0.062	0.030	-0.022	-0.255	-0.003	-0.040	-0.182	-0.034	-0.010	-0.041	0.061
7-8	0.151	-0.292	-0.011	0.029	-0.117	-0.050	0.005	0.137	-0.003	-0.179	-0.074	0.008	0.089
8-9	0.003	0.049	-0.056	0.341	0.132	-0.027	-0.023	0.085	-0.068	-0.061	-0.244	-0.012	0.060
9-10	0.030	0.057	0.081	0.118	0.010	-0.002	0.084	-0.003	0.126	0.036	0.000	-0.054	-0.056

	T-G	T-H	T-I	T-L	T-K	T-M	T-F	T-P	T-S	T-T	W-A	W-R	W-N
0-5	-0.106	0.061	0.167	0.075	-0.170	0.087	0.058	0.108	-0.430	-0.315	0.117	-0.244	0.538
5-6	-0.306	0.011	0.017	0.205	-0.190	0.296	0.411	-0.151	0.033	-0.359	-0.048	-0.050	0.424
6-7	-0.044	-0.264	-0.073	0.020	0.159	0.483	0.250	-0.038	-0.169	-0.289	-0.213	0.242	-0.120
7-8	-0.252	-0.053	0.097	0.003	-0.059	0.089	-0.061	0.018	-0.056	0.125	-0.181	-0.288	0.062
8-9	-0.117	-0.074	-0.053	0.045	-0.024	0.172	-0.051	0.079	-0.020	0.209	-0.083	0.051	0.141
9-10	-0.165	0.094	0.026	0.007	0.086	-0.209	0.038	0.069	-0.050	-0.085	0.056	-0.221	0.265

	W-D	W-C	W-Q	W-E	W-G	W-H	W-I	W-L	W-K	W-M	W-F	W-P	W-S
0-5	1.287	0.080	0.459	-0.089	0.036	-0.098	-0.505	-0.635	-0.822	-0.783	-0.185	-0.312	0.298
5-6	0.386	-0.761	0.712	-0.020	-0.138	-0.397	-0.210	-0.008	0.461	-0.331	0.103	-0.181	-0.067
6-7	-0.126	-0.650	0.204	0.153	0.147	-0.700	0.233	-0.283	-0.036	-0.036	-0.240	-0.017	0.396
7-8	0.263	-0.442	-0.244	-0.222	0.277	0.244	-0.105	-0.116	0.218	-0.293	-0.243	0.186	0.231
8-9	-0.148	-0.203	0.112	-0.014	0.144	0.016	-0.037	-0.128	0.009	-0.338	-0.053	0.194	0.025
9-10	0.108	0.113	-0.297	0.170	0.046	0.061	-0.237	-0.009	-0.179	0.093	-0.041	0.020	-0.108

	W-T	W-W	Y-A	Y-R	Y-N	Y-D	Y-C	Y-Q	Y-E	Y-G	Y-H	Y-I	Y-L
0-5	0.040	0.707	-0.030	0.044	-0.491	0.694	0.069	-0.036	-0.004	0.121	-0.305	-0.100	-0.289
5-6	0.383	-0.206	-0.035	0.030	-0.023	0.380	-0.240	-0.092	0.693	-0.482	-0.107	0.034	-0.226
6-7	0.516	0.743	-0.102	-0.186	-0.181	-0.089	0.302	-0.211	0.187	0.047	0.113	0.001	0.085
7-8	0.188	0.715	0.075	0.091	0.126	-0.257	0.253	-0.059	0.100	0.018	0.024	-0.063	-0.087
8-9	-0.032	0.302	-0.030	0.158	0.085	-0.188	-0.215	0.042	0.146	0.184	0.037	-0.057	-0.134
9-10	-0.107	0.473	-0.082	0.072	0.106	-0.197	-0.135	0.132	-0.072	-0.015	0.093	0.039	-0.003

	Y-K	Y-M	Y-F	Y-P	Y-S	Y-T	Y-W	Y-Y	V-A	V-R	V-N	V-D	V-C
0-5	-0.294	-0.218	-0.335	0.071	0.442	0.492	-0.105	0.211	-0.359	-0.011	0.158	0.442	-0.026
5-6	-0.536	-0.296	-0.040	0.041	0.381	0.187	0.036	0.121	0.069	0.458	0.311	0.378	-0.325
6-7	0.011	-0.119	-0.088	0.119	0.135	-0.068	-0.152	0.269	-0.067	0.383	0.385	0.356	-0.163
7-8	0.197	-0.226	0.185	0.032	0.111	-0.128	-0.069	-0.251	-0.037	-0.006	0.179	0.122	0.051
8-9	0.039	-0.069	-0.068	0.002	-0.042	0.144	0.081	0.032	-0.028	0.301	0.234	0.105	0.019
9-10	0.056	-0.172	-0.010	-0.076	-0.013	0.102	0.028	0.229	-0.061	0.047	0.100	0.112	-0.114

	V-Q	V-E	V-G	V-H	V-I	V-L	V-K	V-M	V-F	V-P	V-S	V-T	V-W
0-5	-0.191	0.255	0.141	0.006	-0.253	-0.454	0.253	-0.265	-0.288	0.358	0.316	0.251	-0.034
5-6	0.148	0.159	0.122	0.499	-0.633	-0.592	0.427	-0.358	-0.264	-0.111	0.186	0.039	-0.224
6-7	0.130	0.597	0.098	0.107	-0.453	-0.589	0.035	-0.364	-0.215	0.136	0.403	-0.024	-0.148
7-8	0.026	0.105	0.030	0.188	-0.234	-0.227	0.027	-0.068	-0.116	0.164	0.280	0.056	-0.081
8-9	-0.026	0.036	0.165	0.185	-0.103	-0.043	-0.142	-0.009	-0.168	0.224	-0.077	-0.122	0.012
9-10	-0.005	0.121	0.025	-0.054	-0.068	-0.026	0.015	0.036	-0.078	0.082	0.005	-0.015	-0.093

	V-Y	V-V	p-A	p-R	p-N	p-D	p-C	p-Q	p-E	p-G	p-H	p-I	p-L
0-5	0.075	-0.641	-0.231	-0.133	-0.256	-0.052	-0.365	0.150	0.077	-0.520	0.097	0.512	0.289
5-6	0.122	-0.405	-0.189	-0.014	0.012	-0.031	-0.194	0.005	0.128	0.052	0.065	0.119	0.213
6-7	-0.068	-0.357	-0.124	-0.001	0.089	0.238	-0.218	-0.030	0.189	0.160	0.067	-0.198	0.032
7-8	-0.059	-0.162	0.010	0.090	0.164	0.114	-0.154	0.094	0.201	0.009	-0.069	-0.176	-0.237
8-9	-0.137	-0.341	0.009	0.100	0.075	0.108	0.078	-0.026	0.109	0.105	-0.012	-0.197	-0.174
9-10	-0.074	0.051	0.028	0.005	0.062	0.024	0.021	0.005	0.045	0.102	0.012	-0.088	-0.142

	p-K	p-M	p-F	p-P	p-S	p-T	p-W	p-Y	p-V	p-p
0-5	-0.021	0.291	0.232	0.033	-0.356	-0.154	0.241	-0.045	0.230	0.000
5-6	-0.062	-0.054	0.004	-0.037	0.069	-0.050	-0.030	-0.004	-0.071	0.000
6-7	-0.083	0.098	0.063	0.001	0.006	-0.126	0.049	0.027	-0.252	0.000
7-8	0.090	-0.027	0.073	-0.070	0.126	0.067	-0.089	-0.011	-0.235	0.000
8-9	0.021	-0.072	-0.114	0.020	0.064	0.088	-0.073	-0.033	-0.070	0.000
9-10	0.018	-0.115	-0.105	0.071	0.121	0.053	-0.069	-0.025	-0.034	0.000

*Pairwise potential components μ^{H}_{rsd}. Residue types follow standard 1-letter codes, with the peptide group indicated by lower-case "p." Distance intervals are 0–5; >5–6; <6–7; <7–8; <8–9; and <9–10 Å.

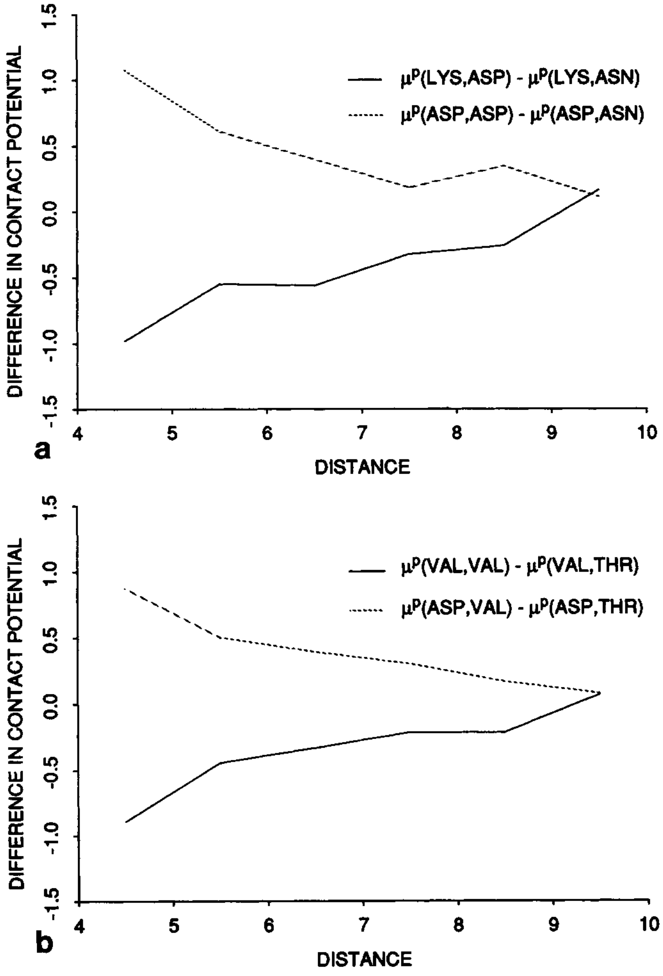


Fig. 1. Differences in pairwise contact potentials μ^P_{rsd} for (**a**) Asn-to-Asp substitution with neighboring Lys or Asp, and (**b**) Val-to-Thr substitution with neighboring Val or Asp. Potential differences are given in *kT* units (see text), and plotted at the midpoints of distance intervals >5–6; >6–7; >7–8; >8–9; >9–10 Å, and at 4.5 Å for the interval 0–5 Å.

and it is perhaps not surprising that complementarity with monomeric motifs is difficult to recognize. Allowing these exceptions, it is fair to say that contact potentials show perfect specificity in matching the correct sequence with each globular folding motif in the known-structure data base.

To test whether contact potentials can also identify the folding motif most compatible with a specified sequence we generate many incorrect backbone conformations for each complete-domain sequence in the known-structure data set. Incorrect folding motifs are drawn from continuous-chain backbone segments in the structures of other proteins. All backbone segments of length 113 are assigned the sequence of 1HMQ, for example. While many alternative motifs chosen in this way are aglobular, and will necessarily have values $\Delta G_{\mathrm{R|M}}$ near zero, others correspond to compact domains with non-native structure, and this trial provides an indication of whether contact potentials can discriminate among alternative globular conformations.

It may be seen in Figure 5 that the sequences of 1HMQ and 1MCP have lower conformational energy in their correct folding motifs than in any of roughly 20,000 others. The complete-domain sequences of most other proteins also have lowest conformational energy when assigned their native conformation, and differences increase with molecular weight as was seen above. Only 10 sequences have lowest conformational potential in a non-native motif. These include sequences of the 7 motifs identified above as having lowest complementarity with their native sequence, and 3 others also drawn from small structures whose native motifs are aglobular as monomers, or in one case when considered in the absence of large prosthetic groups.[47] Only 1 of 10 is a false positive by chance occurrence probability $E_{R|M}$, an alternative conformation for the polio virus subunit

VP4 sequence, whose folding motif similarly preferred an alternative sequence. Aglobular subunits aside, it is fair to say that contact potentials also show perfect specificity in matching the correct folding motif with each sequence in the known-structure data base.

The observation that contact potentials can discriminate among alternative folding motifs suggests that $\Delta G_{R|M}$ *reflects the relative stability of different* conformers, even though it formally gives the expected work for substitution of specific for random sequences. In this context it is interesting to note that the random-sequence reference state we employ is similar to the "random-collapsed" state defined in a thermodynamic scheme by Dill.[48] It follows that we may also interpret $\Delta G_{R|M}$ as a rearrangement energy, the work for transition from a random-collapsed to a specifically-structured state, and probabilities $E_{R|M}$ as the odds that this energy difference would occur by chance in random rearrangements. From the results it would appear that rearrangement energies defined in this way are correlated with conformer stabilities, and are perhaps a major component.

Threading Sequence Through Core Folding Motif

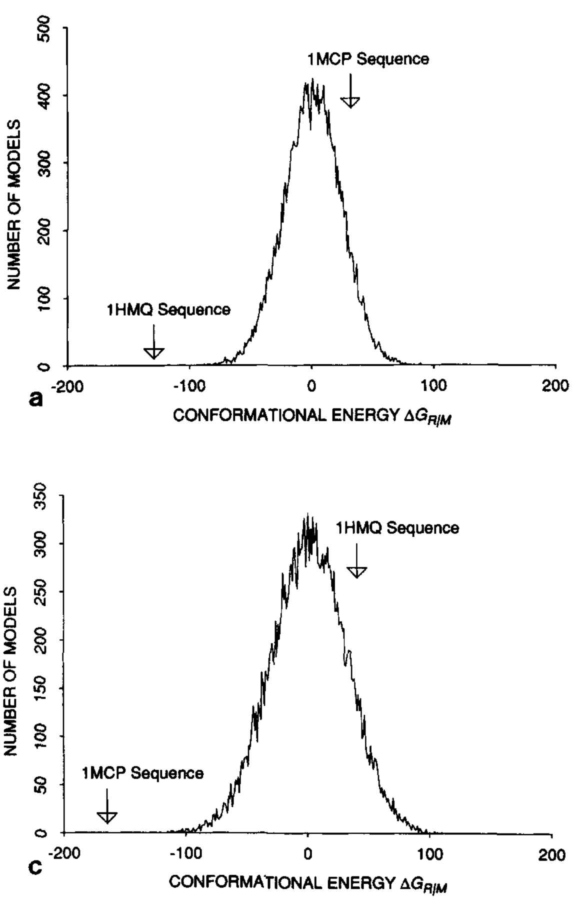
Practical applications of motif recognition must consider "generic" folding motifs that represent the core structure one expects to be conserved in diverse members of a protein family.[3,4] Core motifs will contain "gaps", segments of the polypeptide backbone which form loops in known structures, and which one expects to vary in length and conformation among members of the protein family. Due to the introduction of gaps there will be many ways in which a given sequence can be aligned with a core folding motif, and motif recognition will involve choosing among alternative alignments, as well as choosing among alternative sequences and/or folding motifs. To test whether contact potentials can detect complementarity of sequence and core folding motif we examine models generated by threading the sequences of other proteins through core motifs of 1HMQ and 1MCP.

The core folding motifs we consider are described in Figure 6. The structures of 1HMQ and 1MCP are divided into core and loop segments, and limits on the lengths of subsequences which may be assigned to loop segments are specified. The core segments correspond to elements of secondary structure which are conserved in hemerythrin and myohemerithrin,[51] and among members of the immunoglobulin V-domain family.[52] Loop-length limits are based on variations observed in the respective protein families. They encompass length variations in known sequence homologs of 1HMQ, and in most immunoglobulin V_h or V_l domains. Results below are similar for core definitions which alter the precise boundaries of core segments, or the specific limits placed on loop lengths (not shown).

Sequences are threaded through core motifs by considering all possible placements of core segments along the chain, as constrained by sequence length, core segment length, and limits on allowed loop lengths. The first side chain of 1HMQ core segment 1 might be assigned residue 10 of a sequence, for example. This specifies the assignment of all side chains in segment 1, residues 10–23. The first side chain of core segment 2 might then be assigned residue 28, since this is downstream and within the allowed loop-length limits. This assignment specifies all side chains of core segment 2, and as well the length and composition of the intervening loop, residues 24–27. Once all side chains of the motif are assigned, conformational energy $\Delta G^L_{R|M}$ may be calculated as described above. This threading algorithm effectively considers all possible ways of folding a query sequence into a conformation with a given core motif.

Results of threading different sequences through core motifs of 1HMQ and 1MCP are shown in Figure 7. The number of alternative models is far greater than for complete-domain motifs, but the distribution of conformational energies is otherwise similar, and energies for the correct models still fall outside the distribution defined by incorrect sequence assignments. The energy for the best model derived from the 1MCP sequence in the 1HMQ motif is higher than that of the correct model, and vice versa, indicating that contact potentials can distinguish the misfolded models originally proposed as test cases[10,11] even when only a core substructure is considered, and when an "optimal" misfolded model is constructed by searching all possible alignments of incorrect sequence and core motif. Chance occurrence probabilities $E^L_{R|M}$ may be used to quantitate these preferences. By these criteria the correct sequence for the 1MCP core motif is 111,200,000 more likely than the best derived from 1HMQ, for example, and the correct model with the 1MCP core motif is 34,040 times more likely than the best derived by threading the 1MCP sequence through the 1HMQ core motif. Pairwise interactions in "optimal" misfolded models are illustrated in Figure 8. These models show a number of favorable interactions, since gapped alignment of sequence and folding motif allows positioning of many residues in complementary positions. A number of unfavorable interactions remain, however, and even "optimal" misfolded models lack the continuous network of favorable interactions that characterizes native structures.

The 10 best models derived by threading the sequences of other proteins though the 1HMQ core motif are listed in Figure 7b. We have included in the trial the sequence of myohemerithrin (2MHR), a homolog of 1HMQ for which a high-resolution crystal



b

Chain	CS1	$\Delta G_{R\|M}$	$Z_{R\|M}$	$E_{R\|M}$	Log Odds
1HMQ A	1	-129.36	5.497	<.0001	0.000
1RNH 1	32	-85.80	3.984	0.0015	-4.877
3CLA 1	4	-92.07	3.788	0.0076	-5.597
3GAP A	82	-84.85	3.645	0.0129	-5.825
4FXN 1	23	-87.45	3.259	0.0145	-5.874
8DFR 1	7	-90.09	3.554	0.0145	-5.875
3ADK 1	50	-84.74	3.530	0.0171	-5.947
2TIM A	93	-82.82	3.618	0.0203	-6.021
2BP2 1	5	-69.36	2.887	0.0344	-6.251
3RN3 1	12	-63.68	2.746	0.0356	-6.266

d

Chain	CS1	$\Delta G_{R\|M}$	$Z_{R\|M}$	$E_{R\|M}$	Log Odds
1MCP L	1	-164.60	5.936	<.0001	0.000
4RHV 1	148	-125.01	4.071	0.0041	-6.450
4TNC 1	9	-124.97	3.682	0.0058	-6.596
5TNC 1	9	-120.40	3.552	0.0095	-6.813
3CNA 1	104	-102.67	3.638	0.0170	-7.065
2HMG A	126	-105.77	3.724	0.0209	-7.155
2LBP 1	224	-110.31	3.694	0.0255	-7.242
2LTN A	47	-98.64	3.321	0.0305	-7.319
1LZ1 1	5	-82.79	2.881	0.0351	-7.379
4SBV A	134	-105.18	3.490	0.0354	-7.383

Fig. 2. Conformational energies for models with the complete-domain folding motifs of hemerythrin (**a,b**), and immunoglobulin McPC603 V_l domain (**c,d**). (a) and (c) are histograms giving the number of model structures falling within unit intervals of conformational energy. The 1HMQ motif is defined by backbone coordinates of the A chain, all 113 residues, and the 1MCP motif is defined by backbone coordinates of the L chain, residues 1–113. Sequences are drawn from all unique subsequences of length 113 that are found in the 161-structure data set listed under Methods, to generate a total of 23,824 alternative models. (b) and (d) describe the 10 best models found by threading the sequences of other proteins through either motif, listing conformational energies, Z-scores, and chance occurrence probabilities (see text). Columns labeled CS1 indicate the starting point of the subsequence assigned to each motif relative to the first residue of that chain in the corresponding data bank entry. Log odds log $(E_{R=\text{native}|M}/E_{R|M})$ give chance occurrence probabilities for incorrect sequence models as multiples of those for the correct sequence model.

structure is available.[53] It may be seen that a few incorrect models have conformational energies lower than that of the correct model, but both 1HMQ and 2MHR are distinguished from all incorrect models by chance occurrence probabilities $E^L_{R|M}$. By this criterion there are no false positives, and we may conclude that contact potentials identify the correct models for 1HMQ and 2MHR in a field containing over 20,000,000 alternatives. The 30 best models derived from threading different sequences through the 1MCP core motif are listed in 7d. We have included the sequences of 21 immunoglobulin V-domains whose known structures are compatible with our definition of the 1MCP core motif. It may be seen that the correct model for 1MCP is ranked above the first non-immunoglobulin by criterion $E^L_{R|M}$, indicating that contact potentials can identify the "self" model in a field containing over 100,000,000 incorrect alternatives. Twenty of 21 immunoglobulin V-domain sequences are also identified among the top 30 models, indicating a sensitivity of 20/21 = 0.95 at a threshold sufficient to give a specificity of 20/30 = 0.67. This is a stringent test, however, because we search for subsequences compatible with the V-domain core motif anywhere within the lengths of the non-immunoglobulin sequences. All false positives are eliminated if we restrict the search to sequences which have lengths near those of immunoglobulin light or heavy chains, and ask that the V-domain fall near their N-terminus. By this measure threading is both perfectly sensitive and specific for identification of sequences which fold with the 1MCP core motif.

The true structures of myohemerythrin and the immunoglobulin V-domains included in this test differ from 1HMQ and 1MCP by up to 1.5 Å root mean square super position residual for core backbone atoms. There is no apparent correlation between the values in Figure 6 and superposition residual with 1HMQ or 1MCP (not shown), and we may conclude that sums of contact potentials are relatively insensitive to structural variations of this magnitude. The sequences of these proteins display between 25% and 68% identity with 1HMQ and 1MCP, for core residues, and there is again no significant correlation of conformational energies and percent residue identity. This observation indicates that our representation folding motifs contains no apparent "memory" of sequence, and that conformational energies provide a measure of similarity that is apparently independent of sequence homology.

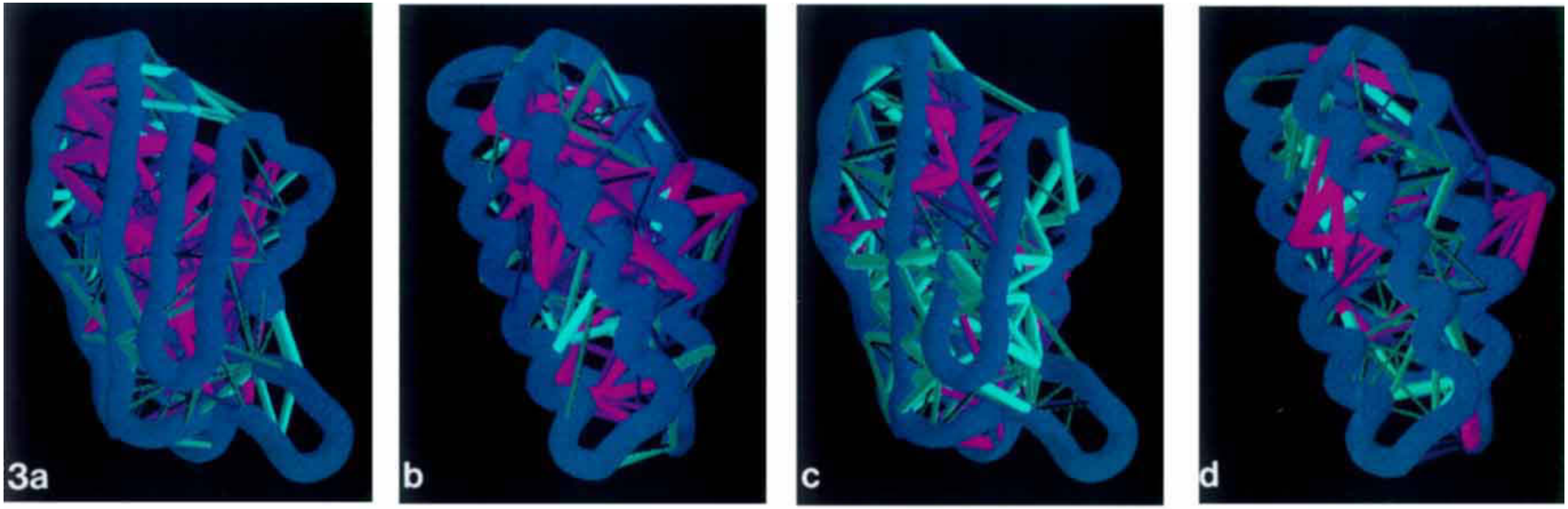
It is interesting to note that the sequence of myohemerythrin (2MHR) has a lower conformational energy in the 1HMQ motif than does the cognate, hemerythrin sequence. 2MHR is monomeric, while 1HMQ forms an octamer,[51,53] and this energy difference probably reflects solvent exposure of oligomer-interface residues from 1HMQ when its sequence is assigned to the monomeric core motif we consider. Similarly, we find that a dimeric 1MCP motif which includes fixed V_h-domain contacts leads to improved sensitivity in identification of other immunoglobulin sequences (not shown). These observations suggest that inclusion of docking contacts in definition of a core motif will be generally useful. It is also interesting that the subsequences forming the best incorrect models are often derived from larger proteins. The best incorrect model with the 1HMQ core motif, for example, derives from 1R1A, Rhinovirus Coat Protein,[54] a fragment which occurs on the exterior of two β-sheet domains and which is itself aglobular. The best incorrect model with the 1MCP motif derives from 3BLM, β lactamase,[55] an extended fragment falling largely at the interface of two domains. It would appear that the sequence features required to "dock" these fragments with the remainder of their structures partially mimic those required to fold an isolated chain into the 1HMQ or 1MCP motifs. This observation suggests that it will be important to determine possible domain structures of long sequences, and to consider core motifs whose sizes are compatible with this domain structure.

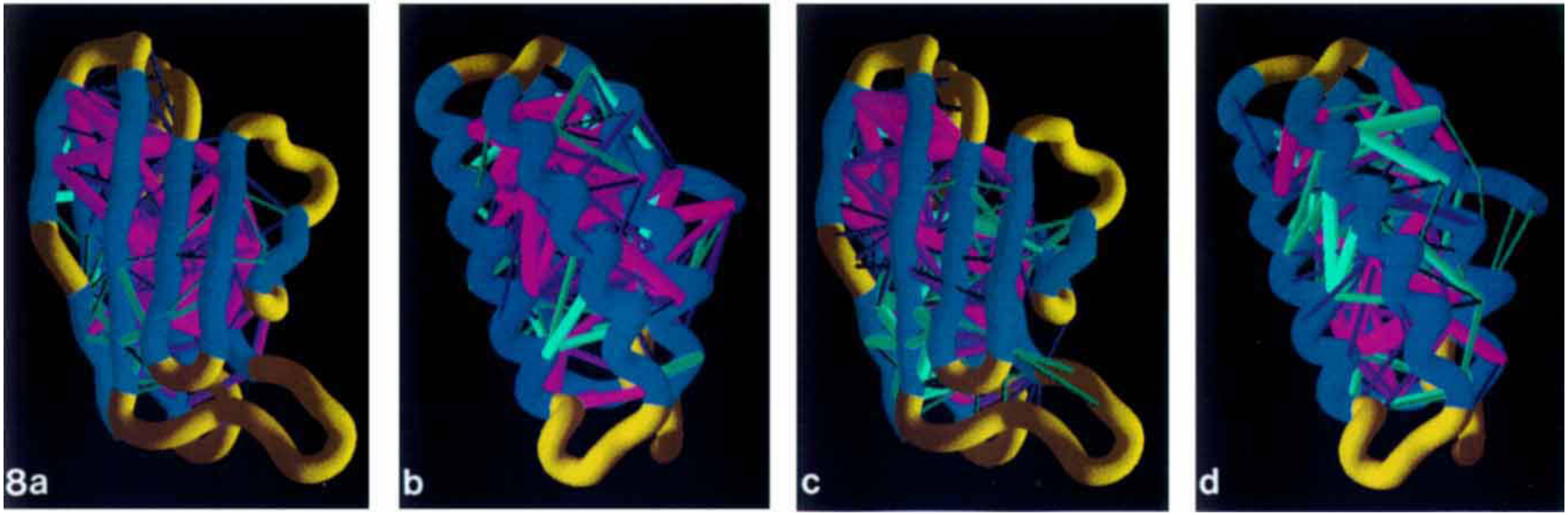
The specificity with which contact potentials identify correct alignments of sequence and core folding motif is illustrated in Figure 9. Threading of the 1HMQ sequence through its core motif generates 1,477 alternative models, but only a few have low conformational energy and significant probability in an ensemble defined by the alternatives. The correct model begins core segments 1 through 5 with residues 2, 18, 41, 69 and 90, respectively, and it may be seen in Figure 9b that this "thread" is identified as the best. The next-best alignment places residue 94 at the start of core segment 5. This corresponds to a model in which helix 4 is moved axially by one helical repeat, preserving hydrophobic contacts with the remainder of the 4-helix bundle. The remaining low-energy threads similarly displace one or two core segments by amounts consistent with the periodicity of their secondary structure. Threading of the 1MCP sequence through its core motif generates 3,697 alternative models, and it may be seen in Figure 9d that the correct alignment is identified as the best. The next-best alignment displaces only core segment 10, and corresponds to a model in which the carboxy-terminal β-strand undergoes a "register shift" of 2 residues. The other low-energy alignments similarly "overlap" the correct model, displacing a few of the shorter β-strands, but preserving most residues in their correct positions it the three-dimensional structure.

Accurate alignments are also obtained when the sequences of structurally similar proteins are threaded through the core motifs of 1HMQ and 1MCP. Threading of the myohemerythrin sequence through the 1HMQ core motif identifies a small group of low-energy alignments, and the best, listed in Figure 7b, corresponds exactly with the optimal structural superposition. Correct structural alignments are also identified as the best for 13 of 20 other immunoglobulin V-domain sequences, when threaded through the 1MCP core motif. In the remaining 7 cases the correct alignment falls within the low-energy group, and differs from the best by displacement of shorter β-strands, as seen for other

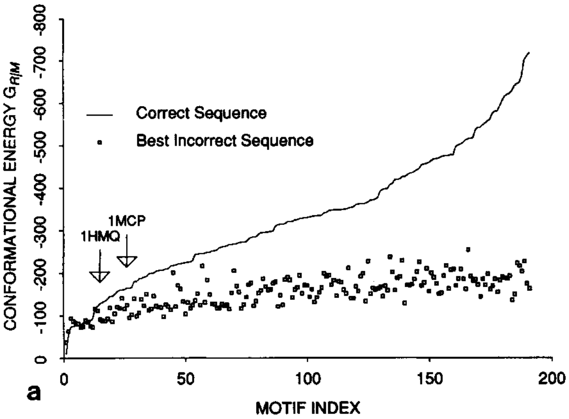
Fig. 3. Pairwise interactions in the correct structures of (**a**) 1MCP and (**b**) 1HMQ, and in misfolded models generated by assigning (**c**) the sequence of 1HMQ to the folding motif of 1MCP, and (**d**) the sequence of 1MCP to the folding motif of 1HMQ. The most favorable interactions are indicated by magenta-colored cylinders linking C_α coordinates within the backbone "worm". Unfavorable interactions are indicated by cyan-colored cylinders. The diameter and brightness of the cylinders indicates the strength of interactions, with the largest, brightest cylinders indicating interactions with absolute values greater than 1 energy unit. Side-chain to side-chain and side-chain to peptide-group interactions are summed, if they occur between the same residue pair. The figures are produced by the graphics program GRASP, by Anthony Nicholls.[72] Most of the unfavorable interactions visible in the correct structures correspond to ligand binding and/or macromolecular docking sites, where pairwise interactions are favorable only when the ligand and/or neighboring protein domain is included in the conformational energy calculation.

Fig. 8. Pairwise interactions for correct models with the core folding motifs of (**a**) 1MCP and (**b**) 1HMQ, and for "optimal" misfolded models generated by threading (**c**) the sequence of 1HMQ through the 1MCP core motif, and (**d**) the sequence of 1MCP through the 1HMQ core motif. Pairwise interactions are color-coded as in Figure 3. Loop regions of the core folding motifs are indicated by yellow coloring of the backbone "worm." Loop regions are ignored in calculation of conformational energy, except for evaluation of core-side-chain to loop-peptide-group interactions as described under Methods.





Figures 3 and 8. Legends appear on page 104.



b

Motif	$\Delta G_{R\|M}$	$Z_{R\|M}$	$E_{R\|M}$	$\Delta G_{n\|M}$	$Z_{n\|M}$	$E_{n\|M}$	Log Odds
2PLV 4	-37.64	3.960	0.0037	-10.44	1.339	0.0903	1.391
1CSE I	-87.34	4.667	0.0002	-75.73	3.786	0.0001	-0.466
2GN5 1	-94.59	3.655	0.0095	-74.67	3.086	0.0010	-0.970
1UTG 1	-63.96	3.994	0.0137	-57.75	3.462	0.0003	-1.708
2SNI I	-81.38	3.703	0.0259	-78.14	3.339	0.0004	-1.789
2CI2 I	-89.97	3.693	0.0252	-83.85	3.592	0.0002	-2.185
1TEC I	-84.33	3.870	0.0122	-77.52	3.777	0.0001	-2.187

Fig. 4. Conformational energies for models with the folding motifs of proteins in the known-structure data set. Plot (**a**) shows energies for correct-sequence models as a solid line, and the lowest energy found for incorrect-sequence models as an open square. The figure includes 191 motifs drawn from unique monomer or subunit structures in the data set listed under Methods. Incorrect models are generated by assigning to each folding motif all unique subsequences of the same length, drawn from the sequences of other, longer proteins in the data set. Homologous sequences sharing >20% residue identity with the correct sequence of each motif are omitted. The number of alternative models varies with motif length, ranging from 31,753 for 1CSE-1, to 258 for 7CAT-A. (**b**) describes the 7 motifs for which an incorrect model has lower conformational energy than the correct model, listing conformational energies, Z-scores, and chance occurrence probabilities for the best incorrect models ($G_{R/M}$, *etc.*) and for the correct models ($G_{n/M}$, *etc.*). Also listed are log odds of chance occurrence for the correct vs. best incorrect model, log ($E_{R=\text{native}/M}/E_{R/M}$). 1CSE-I[38] and 1TEC-I[39] are derived from eglin C, a protease inhibitor from leeches, observed in complex with two different proteases. 2CI2-I[40] and 2SNI-I[41] are derived from barley CI-2, a protease inhibitor observed free and as a protease complex. Motifs 2GN5-1[42] and 1UTG-1[43] are monomers from the dimeric structures of bacteriophage fd gene-5 protein and rabbit uteroglobin, respectively. Motif 1PLV-4[44] is subunit VP4 of the polio virus capsid protein.

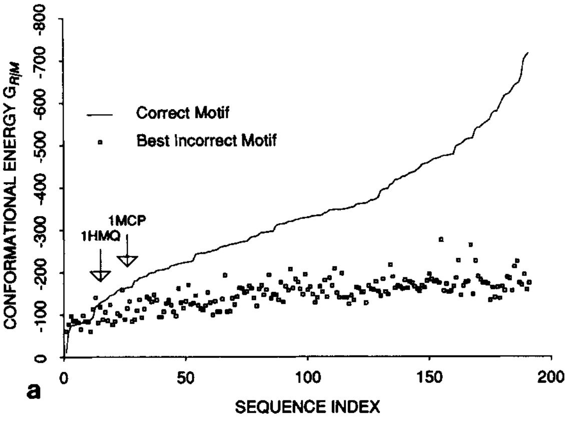
low-energy "self" threads of 1MCP. These observations suggest that contact potentials generally detect complementarity of sequence and core folding motif by identifying a small group of related, low-energy alignments. This group includes the correct structural alignment in all cases considered, suggesting that accurate alignments as required for homology modeling are likely to be identified by "threading".

To examine their contribution to alignment specificity, we separate in Figures 9b and 9d energies derived from the hydrophobic and pairwise components of the contact potential. It may be seen that the hydrophobic component of conformational energy is generally larger in absolute value, consistent with observations above concerning the importance of residue-specific hydrophobicities. For the 1HMQ motif the correct thread is identified as the best by both hydrophobic and pairwise energies, but the pairwise component distinguishes it from the low-energy threads to a greater extent. The 1MCP motif contains more core segments, some quite short, and the correct thread is identified only as one of the best by both pairwise and hydrophobic energies. In combination, however, they identify the correct thread as that with minimum chance occurrence probability. These observations suggest that the pairwise component of contact potentials may be particularly important for identification of correct alignments. We also note that correct alignments are less often identified as the best within the low-energy group by conformational energies $\Delta G_{R|M}$, which do not take into account shielding of core residues by loops that is included in $\Delta G^{L}_{R|M}$ (not shown). This indicates that approximate treatment of loop-residue contacts is useful in choosing among alignments within this group, and that improved representation of loop-core interactions may improve specificity.

These tests make no use of sequence features one might expect in proteins that are members of the hemerythrin or immunoglobulin V-domain families. The best alignment for 1MCP reproduces a conserved disulfide bond, for example, but we do not constrain alternative models to place cysteine at the appropriate residue sites, nor include terms sensitive to disulfide bonding in the potential. The best alignment for 1HMQ similarly reproduces the correct placement of active-site histidines, but we do not impose this as a constraint. In practical applications one might easily make use of local sequence features expected in members of a protein family, or more generally combine sums of contact potentials with measures of sequence similarity.

DISCUSSION

These computer experiments demonstrate that it is possible to recognize complementarity of protein sequence and folding motif by means of a simple empirical energy function. They suggest that conformational energies can be successfully represented as a sum over residue contact potentials and that useful estimates of these energetic parameters may be derived by statistical analysis of the structural data base. The accuracy of the current energy function is



b

Chain	$\Delta G_{R\|M}$	$Z_{R\|M}$	$E_{R\|M}$	$\Delta G_{R\|n}$	$Z_{R\|n}$	$E_{R\|n}$	Log Odds
2PLV 4	-60.68	3.121	0.0595	-10.44	1.339	0.0903	0.181
2GN5 1	-97.66	3.747	0.0198	-74.67	3.086	0.0010	-1.289
1CSE I	-86.26	4.194	0.0037	-75.73	3.786	0.0001	-1.687
1TEC I	-86.65	4.182	0.0039	-77.52	3.777	0.0001	-1.695
1UTG 1	-77.65	3.492	0.0639	-57.75	3.462	0.0003	-2.377
2SNI I	-81.82	3.252	0.1053	-78.14	3.339	0.0004	-2.399
2HMG B	-140.41	3.566	0.0286	-119.49	3.874	0.0001	-2.728
2CI2 I	-84.77	3.293	0.0911	-83.85	3.592	0.0002	-2.744
2WRP R	-98.53	3.629	0.0395	-83.35	4.360	<.0001	-3.785
2CDV 1	-113.42	3.969	0.0080	-94.89	5.113	<.0001	-4.705

Fig. 5. Conformational energies for models with the sequences of proteins in the known structure data set. Plot (**a**) shows energies for correct-motif models as a solid line, as in Figure 3, and the lowest potential found for any incorrect-motif model as an open square. The figure includes 191 sequences drawn from unique monomer or subunit structures in the data set, and incorrect models are generated by assigning to each sequence the conformations of all unique backbone segments which have the same length, as drawn from the structures of other, longer proteins. Backbone segments whose native sequences share >20% residue identity are omitted. (**b**) describes the 10 sequences for which an incorrect model has lower conformational energy than the correct model, listing conformational energies, Z-scores, and chance occurrence probabilities for the best incorrect model ($G_{R|M}$, *etc.*) and for the correct model ($G_{R|n}$, *etc.*). Also listed are log odds of chance occurrence for the best incorrect model relative to the correct model, log ($E_{R|M=native}/E_{R|M}$). Seven of these sequences are from proteins described in the caption to Figure 3. Sequence 2HMG-B is from one chain of influenza virus hemagglutinin[45], sequence 2WRP-R from E. coli trp repressor,[46] and sequence 2CDV-1 from cytochrome c3 of Desulfovibrio vulgaris.[47]

such that it can distinguish correct structures not only from grossly misfolded proteins, but as well from models which are only partially incorrect, differing by displacement of a helix by one turn, for example, or by "register shifts" of a β-strand. It seems likely that contact potentials will generally identify correct alignments of sequence and folding motif as one of a few low-energy alternatives, and that "threading" may indeed prove useful for structure prediction by recognition of folding motif.

To evaluate alternative models, we find it useful to compare not only conformational energies, but also the probabilities that these values could occur by chance among random alignments of sequence and folding motif. This statistical criterion corrects for effects of sequence length and composition, and more generally reflects uncertainties inherent in use of an approximate energy calculation, and in representation of alternative conformations by incomplete, "core" substructures. It is interesting that this chance occurrence probability is as successful in distinguishing alternative motifs, given a sequence, as it is in distinguishing among sequences, given a motif. Our calculation formally refers to energy differences expected for random sequence rearrangement and/or random conformational rearrangement to an equally compact state, with equivalent backbone-backbone interactions, but it appears that we may nonetheless distinguish among folding motifs which differ in these respects. It has been suggested that backbone conformations from native proteins are generally representative of stable, compact conformers,[49,50] and this may account for our success in identifying correct motifs with a statistical criterion which refers only to sequence complementarity as determined by side chain interactions.

Other workers have recently addressed problems in motif recognition using empirical scoring functions, and it interesting to compare these methods. Bowie, Eisenberg and colleagues[56–58] and Overington et al.[59] represent folding motifs as a linear "profile" containing properties of the local environment at each residue site, and they derive residue preferences for each environment from substitutions observed in aligned sequence families where a structure is known. Solvent accessibility is important among these properties, and these workers' methods are loosely analogous to "threading" with the hydrophobic component of a contact potential. Local environment profiles are strongly dependent on the side chain types originally present in the folding motif, however, and this "memory" of sequence leads to Z-scores which are highly correlated with those from conventional alignment.[57,59] Bowie, Eisenberg and colleagues[56–58] and Overington et al.[59] detect sequences complementary with the profiles of several protein families, including cases with low homology. It seems likely that their environment preferences reflect hydrophobic complementarity of sequence and motif, but they are also sensitive to sequence similarities, and it is difficult to draw further conclusions concerning the accuracy of their empirical potentials.

Sippl and colleagues[60–62] and Jones et al.[63] derive potentials of mean force from pairwise contacts in known structures. These workers do not define a

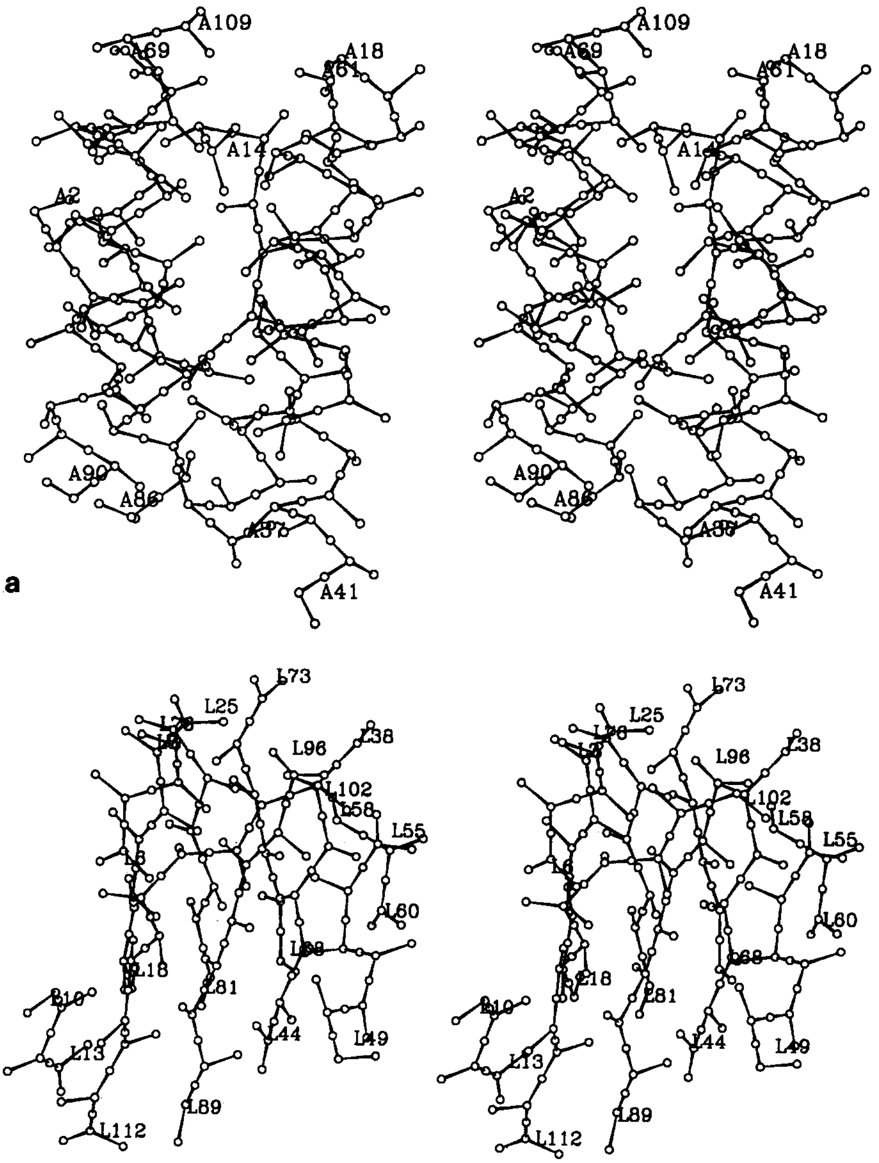
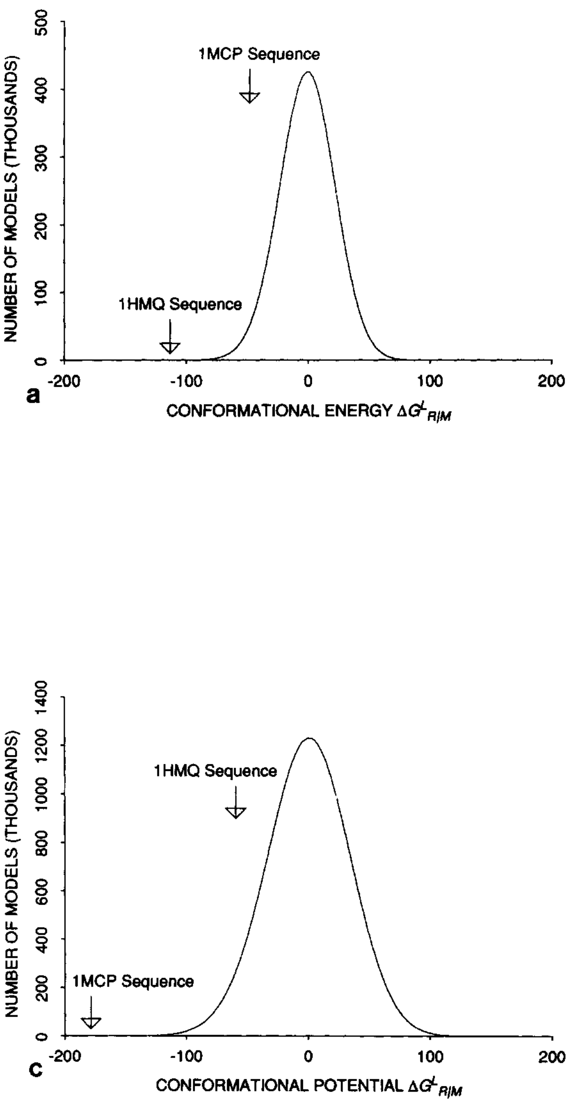


Fig. 6. Core folding motifs of (**a**) hemerythrin (1HMQ), and (**b**) immunoglobulin McPC603 V_l domain (1MCP). Stereo drawings show C_α and summary coordinates for each residue included in the core motif definition (see text). Five 1HMQ core segments (CS) are defined by residues 2–14, 18–37, 41–61, 69–86, and 90–109. CS2 through CS5 correspond to helices of the four-helix bundle, and CS1 to an N-terminal segment of irregular conformation which packs against the exterior of the bundle. Loop-length limits for threaded sequences are 3–5, 3–9, 3–9, and 3–9, for loops CS1–CS2 through CS4–CS5, respectively. Ten 1MCP core segments are defined by residues 3–16, 10–13, 18–25, 38–44, 49–55, 58–60, 68–73, 76–81, 89–96, and 102–112. These correspond to β-strands conserved among most immunoglobulin V_l domains. Loop-length limits are set to 2–3, 4, 5–12, 4, 2–9, 7, 2–4, 7, and 3–15, for CS1–CS2 through CS9–CS10, respectively. Alignments shown below indicate only the starting residues of CS1, CS2, CS4, CS6, CS8, and CS10, since the lengths of interventing loops are fixed. The 1MCP core definition is compatible with the structures of V-domains in the immunoglobulins listed under Methods, with the exceptions of 1CD4-1 and 3FAB-L, where β-strands conserved in other V-domains are deleted.

protein-specific reference state, as we do, but instead sum contact counts from different proteins, and compare totals by pair category to their mean value. With this procedure, they cannot define residue-specific hydrophobicities, and the derived potentials may be affected by between-protein differences in size and amino acid composition. Sippl and colleagues[60–62] and Jones et al.[63] also employ a large number of pair categories, based on separation in the sequence and/or backbone atom types, and their potentials contain roughly 80,000 parameters. Sippl and colleagues test their potential by searching for the folding motifs which match each sequence in a known-structure data set, using ungapped alignment,[61] and they report a specificity slightly lower than we show above for the same problem. It seems clear from this comparison that a large number of pair categories is unnecessary if the contact poten-



b

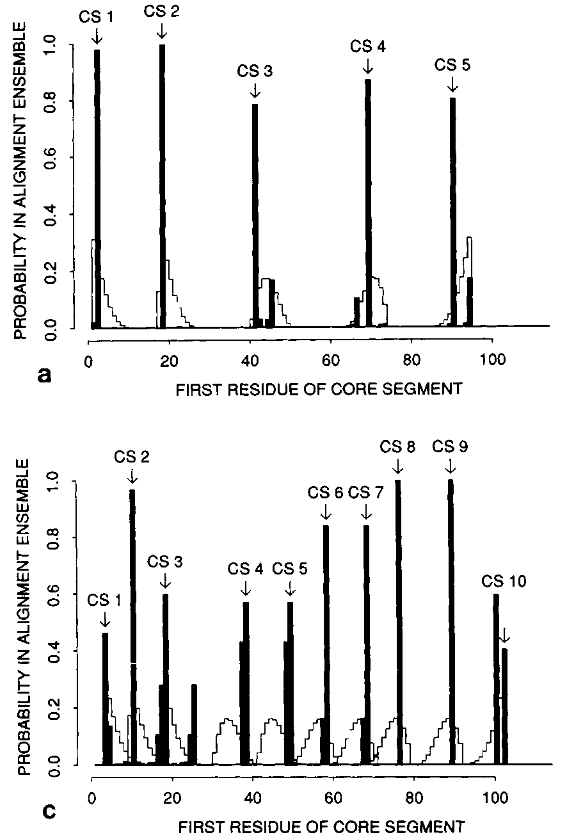
	Chain	CS1	CS2	CS3	CS4	CS5	$\Delta G^L_{R\|M}$	$Z^L_{R\|M}$	$E^L_{R\|M}$	Log Odds
+	2MHR 1	2	18	41	69	95	-143.95	5.716	<.0001	1.464
+	1HMQ A	2	18	41	69	90	-113.37	4.868	0.0008	0.000
	1R1A 1	163	179	207	231	252	-120.39	5.606	0.0019	-0.346
	5TNC 1	53	69	95	122	143	-129.34	5.144	0.0068	-0.909
	2HMG B	36	53	82	111	133	-120.07	5.180	0.0071	-0.928
	7CAT A	385	401	429	458	480	-111.58	5.417	0.0122	-1.164
	3GPD R	143	160	188	212	233	-118.98	5.230	0.0191	-1.361
	4TNC 1	53	69	95	122	143	-123.97	4.933	0.0202	-1.385
	3PFK 1	92	109	133	161	186	-115.68	5.202	0.0207	-1.395
	5RUB A	181	197	222	250	277	-118.60	5.312	0.0208	-1.396

d

	Chain	CS1	CS2	CS4	CS6	CS8	CS10	$\Delta G^L_{R\|M}$	$Z^L_{R\|M}$	$E^L_{R\|M}$	Log Odds
+	1F19 L	3	10	32	52	70	96	-169.66	5.744	<.0001	0.283
+	1MCP L	3	10	38	58	76	102	-174.57	6.138	<.0001	0.000
+	3HFM L	3	10	32	52	70	96	-166.37	5.426	<.0001	-0.514
+	1FDL L	3	10	32	52	70	96	-156.10	5.401	<.0001	-0.573
+	2HFL L	3	10	31	51	69	94	-162.83	4.980	<.0001	-0.744
	3BLM 1	144	151	172	194	213	244	-226.04	6.577	<.0001	-1.035
+	2RHE 1	3	9	33	53	71	95	-163.27	5.453	<.0001	-1.234
+	3HFM H	3	9	33	57	77	101	-168.47	5.491	0.0001	-1.519
+	3MCG 1	3	9	34	54	72	97	-150.82	5.210	0.0001	-1.815
+	1REI A	3	10	32	52	70	96	-139.43	4.506	0.0006	-2.572
+	4FAB L	3	10	37	57	75	101	-146.24	4.974	0.0008	-2.735
+	2FB4 H	3	9	33	58	78	102	-170.17	5.212	0.0011	-2.841
+	2FB4 L	3	9	33	53	71	95	-132.48	4.504	0.0012	-2.877
+	1MCW W	3	9	34	54	72	97	-136.82	4.625	0.0020	-3.112
+	2FBJ H	3	9	32	58	78	106	-160.81	5.020	0.0030	-3.284
+	2FBJ L	3	10	31	51	69	95	-127.67	3.927	0.0033	-3.324
+	1FDL H	3	9	35	57	77	104	-146.18	4.820	0.0049	-3.504
	2GBP 1	126	132	160	184	203	229	-174.67	5.710	0.0054	-3.544
+	3FAB H	3	9	32	59	77	105	-142.84	4.795	0.0073	-3.676
	1CD4 1	50	56	81	106	125	154	-162.71	5.462	0.0080	-3.714
	1BMV 1	54	61	89	111	131	156	-165.21	5.403	0.0132	-3.931
+	4FAB H	3	9	33	60	80	106	-150.91	4.693	0.0155	-3.999
	3GBP 1	126	132	160	184	203	229	-170.03	5.518	0.0161	-4.017
	1I1B 1	39	45	67	94	114	140	-179.67	5.189	0.0191	-4.090
	1LAP 1	125	131	156	182	200	227	-186.59	5.552	0.0258	-4.222
+	2HFL H	3	9	33	58	78	104	-141.15	4.462	0.0277	-4.252
	1ACX 1	3	9	30	51	69	97	-104.58	3.735	0.0325	-4.321
	2LTN A	49	56	84	104	123	159	-149.97	5.180	0.0352	-4.356
	5RUB A	15	21	48	75	93	123	-165.03	5.431	0.0508	-4.516
+	1MCP H	3	9	33	60	79	106	-141.19	4.563	0.0613	-4.597

Fig. 7. Conformational energies for models with the core folding motifs of hemerythrin (**a,b**) and immunoglobulin McPC603 V_l domain (**c,d**). Models are generated by assigning different sequences to each motif. Plots (a) and (c) are histograms giving the number of model structures falling within unit intervals of conformational energy. Conformational energies for the correct models are indicated, as are the lowest conformational potentials found for any alignment of the 1MCP sequence with the 1HMQ core motif, and vice versa. (b) and (d) further describe the best models found for either motif, listing conformational energies, Z-scores, chance occurrence probabilities, and log odds of chance occurrence relative to the native-sequence model, log $(E^L_{R=\text{native}|M}/E^L_{R|M})$. A " + " in the first column indicates a true positive, chains whose native structure is similar to 1HMQ or 1MCP, as determined by structural alignments (not shown). Columns labeled CS1 through CS10 indicate the residues assigned to the first position of each core segment. For the 1HMQ core motif sequences are derived from 175 unique monomers or subunits that have sufficient lengths, to generate total of 24,347,782 alternative model structures. The 1MCP core motif is similarly threaded by the sequences of 171 unique chains, to generate a total of 104,673,011 alternative model structures. By criterion log $(E^L_{R=\text{native}|M}/E^L_{R|M})$ the correct sequence for the 1HMQ core motif is 16,070 times less likely to have occurred by chance than the best assignment derived from the 1MCP sequence, and the correct sequence for the 1MCP core motif is 111,200,000 times more likely than the best derived from 1HMQ. By criterion log $(E_{R|M=\text{native}}/E_{R|M})$ the correct model with the 1HMQ core folding motif is 1199 times more likely than the best model derived by threading the 1HMQ sequence through the 1MCP core motif, and the correct model with the 1MCP core motif is 34,040 times more likely than the best model derived by threading the 1MCP sequence through the 1HMQ core motif. The differences in specificity reflect the β-sheet structure of 1MCP, which produces a greater number of non-local contacts per residue, and also allows a greater number of alternative alignments, due to the greater number of core segments.

tial takes account of hydrophobicity, and is estimated from the data base in a different manner. Jones et al.[63] add an empirically derived hydrophobicity term and gap penalties, and report accurate, gapped alignments of sequence and core folding motif. One may infer that the hydrophobicity term increases specificity, in agreement with our conclusions above concerning its importance. Gap penalties based on original loop lengths obviously carry some element of sequence "memory," however.



b

CS1	CS2	CS3	CS4	CS5	$\Delta G^h_{R\|M}$	$\Delta G^p_{R\|M}$	$\Delta G^L_{R\|M}$	$Z^L_{R\|M}$	Log Odds
2	18	41	69	90	-85.77	-27.60	-113.37	4.868	0.000
2	18	41	69	94	-83.87	-21.62	-105.49	4.604	-0.565
2	18	45	69	90	-85.08	-20.55	-105.63	4.408	-0.966
2	18	41	66	90	-82.54	-21.33	-103.87	4.296	-1.188
2	18	42	69	90	-76.47	-19.73	-96.20	4.119	-1.528
1	18	41	69	90	-69.43	-23.64	-93.07	4.027	-1.700
2	18	45	69	94	-83.06	-14.67	-97.73	4.017	-1.718
2	18	44	69	90	-77.02	-19.53	-96.55	3.979	-1.788
2	18	45	73	94	-78.39	-14.14	-92.54	3.918	-1.899
2	18	41	66	89	-76.36	-16.44	-92.80	3.794	-2.119

d

CS1	CS2	CS4	CS6	CS8	CS10	$\Delta G^h_{R\|M}$	$\Delta G^p_{R\|M}$	$\Delta G^L_{R\|M}$	$Z^L_{R\|M}$	Log Odds
3	10	38	58	76	102	-142.18	-32.38	-174.57	6.138	0.000
3	10	38	58	76	100	-141.08	-36.19	-177.28	6.050	-0.236
3	10	37	58	76	102	-143.59	-31.14	-174.73	6.024	-0.306
3	10	37	57	76	102	-139.76	-32.52	-172.29	6.013	-0.338
4	10	37	58	76	102	-135.84	-32.42	-168.26	5.939	-0.533
3	10	37	58	76	100	-142.48	-34.16	-176.64	5.938	-0.537
4	10	38	58	76	102	-134.44	-33.62	-168.05	5.932	-0.553
4	10	38	58	76	100	-133.31	-38.05	-171.36	5.931	-0.556
4	10	37	57	76	102	-132.01	-33.80	-165.82	5.890	-0.663
3	10	37	57	76	100	-138.65	-35.56	-174.21	5.889	-0.665

Fig. 9. "Self" alignments of the hemerythrin (**a**,**b**) and immunoglobulin McPC603 V_l domain (**c**,**d**) sequences with their core folding motifs. Solid bars in (a) and (c) show probabilities that the indicated residues will be assigned as the first of each core segment, and open bars show the values expected by chance. The correct assignments are indicated by arrows. (b) and (d) list the ten best "self" alignments and their log odds of chance occurrence relative to the correct assignment, log $(E^L_{R=native|M}/E^L_{R|M})$. They also list conformational energy, Z-score, and the hydrophobic $(\Delta G^h_{R|M})$ and pairwise components $(\Delta G^p_{R|M})$ of conformational energy. Probabilities that a core segment begins at a certain residue are calculated by summing probabilities for all sequence assignments that place this residue at that position. These probabilities are calculated from the Boltzmann law, $p(R) = (\Delta G^L_{R|M}/t)/(\sum_R \Delta G^L_{R|M}/t)$, where the partition function in the denominator defines the conformational ensemble of all alternative "self threads." The effective temperature t is here set to 5 kT units. Chance probabilities are defined by p(R) for infinite temperature t, and they show the effect of alignment constraints imposed by non-overlap of core segments and loop-length limits. A few residues in the 1MCP sequence have finite probabilities of beginning more than one core segment, and the solid bars correspond in this case to the sum of these probabilities.

Our results suggest that gap penalties are unnecessary for accurate alignment of sequence and core folding motif.

Godzik et al.[64,65] identify residue contacts on the basis of the closest approach of side chain atoms, rather than backbone or virtual C_β coordinates. These workers do not define a protein-specific reference state, but they derive residue hydrophobicities from the overall proportion of residues buried beyond a threshold, and pairwise potentials by comparing total counts across proteins to expected values defined by products of residue hydrophobicities. They similarly define potentials for three-way "triple" contacts, and add gap penalties, but the total number of parameters remains small since only a single distance interval is considered. Godzik et al.[64,65] show specificity comparable to ourselves for recognition of correct sequences, by ungapped alignment, and also for recognition of core motifs by gapped alignment. The Z-scores they compute appear to be correlated with sequence similarity, however. Myoglobin shows an energy (equivalent to a minus Z-score[64]) of −15 against its own core motif, for example, while hemoglobin, whose structure is similar to 1.5 Å but whose sequence is only 30% identical, shows an energy of only −5. These scores may be influenced by gap penalties, but it seems likely that representation of folding motifs on the basis of side chain contacts present in the original sequence entails some "memory" of that sequence. Sites with many contacts are likely to have been occupied by large residues, for example, and a contact potential defined on the basis of side chain atomic contacts will tend to put large residues back in these positions. Because of this question, it is difficult to evaluate the novel features of the empirical potential proposed by Godzick et al.,[64,65] in particular, their interesting use of residue "triples."

All of these "threading" methods may be viewed as limited searches of the configuration space available to polypeptide species, and in this respect they are similar to other current research in tertiary structure prediction. Search methods have been applied to the problem of side chain packing, given an alignment of sequence and core folding motif, and they require potentials sensitive to details of side chain size and shape,[66,67] and solvent exposure of nonpolar atoms.[68] Threading considers a "lower resolution" problem, that of choosing the core motif and alignment. An "ideal" potential should reflect only constraints on the chemical and steric properties of side chain groups that are common to all sequences compatible with a folding motif, as distinct from their similarity to a particular, known sequence. Lattice models address a still "lower resolution" problem, that of generating protein secondary structure and topology, and they must employ potentials which reflect the "homopolymer" components of conformational energy, backbone hydrogen bonding and torsional flexibility.[69–71] It remains to be seen how these approaches may best be combined, and lead to general methods for tertiary structure prediction. The unique advantage of threading, perhaps, is that one may substitute a data base of known motifs for an encoding of rules governing protein backbone topologies, and may attempt to recognize known folding motifs in new sequences.

ACKNOWLEDGMENTS

We thank Geoff Barton, David States, and the referees for critical readings of this manuscript, and Jeff Skolnick and Janet Thornton for sending preprints of their work on protein threading.

REFERENCES

1. Chothia, C. One thousand families for the molecular biologist. Nature 357:543–544, 1992.
2. Richardson, J.S. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34:167–339, 1981.
3. Blundell, T.L., Sibanda, B.L., Sternberg, M.J., Thornton, J.M. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326:347–352, 1987.
4. Greer, J. Comparative modeling of homologous proteins. Methods Enzymol. 202:239–252, 1991.
5. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
6. Gregoret, L.M., Cohen, F.E. Novel method for the rapid evaluation of packing in protein structures. J. Mol. Biol. 211:959–974, 1990.
7. Clark, D.A., Shirazi, J., Rawlings, C.J. Protein topology prediction through constraint-based search and the evaluation of topological folding rules. Protein Engineering 4:751–760, 1991.
8. Finkelstein, A.V., Reva, B.A. A search for the most stable folds of protein chains. Nature 351:497–499, 1991.
9. Taylor, W.R. Towards protein tertiary fold prediction using distance and motif constraints. Protein Engineering 4:853–870, 1991.
10. Novotny, J., Bruccoleri, R.E., Karplus, M. An analysis of incorrectly folded protein models. Implications for structure predictions. J. Mol. Biol. 177:787–818, 1984.
11. Novotny, J., Rashin, A.A., Bruccoleri, R.E. Criteria that discriminate between native proteins and incorrectly folded models. Proteins 4:19–30, 1988.
12. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. Nature 319:199–203, 1986.
13. Bryant, S.H., Amzel, L.M. Correctly folded proteins make twice as many hydrophobic contacts. Int. J. Peptide Protein Res. 29:46–52, 1987.
14. Baumann, G., Frömmel, C., Sander, C. Polarity as a criterion in protein design. Protein Engineering 2:329–334, 1989.
15. Chiche, L., Gregoret, L.M., Cohen, F.E., Kollman, P.A. Protein model structure evaluation using the solvation free energy of folding. Proc. Natl. Acad. Sci. USA 87:3240–3243, 1990.
16. Holmes, M.A., Stenkamp, R.E. Structures of met and azidomet hemerythrin at 1.66 Å Resolution. J. Mol. Biol. 220: 723–737, 1991.
17. Satow, W., Cohen, G.H., Padlan, E.A., Davies, D.R. Phosphocholine binding immunoglobulin FAB McPC603. An X-ray diffraction study at 2.7 Å. J. Mol. Biol. 190:593–604, 1987.
18. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J.C. Protein data bank. In: "Crystallographic Databases: Information Content, Software Systems, Scientific Applications." Allen, F.H., Bergerhoff, G., Sievers, R., eds. Bonn, Chester, Cambridge: Int. Union of Crystallography, 1987:107–132.
19. Miyazawa, S., Jernigan, R.L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18:534–552, 1985.
20. Colonna-Cesari, F., Sander, C. Excluded volume approximation to protein-solvent interaction. The solvent contact model. Biophysical J. 57:1103–1107, 1990.
21. Bryant, S.H. PKB: A program system and data base for analysis of protein structure. Proteins 5:233–247, 1989.
22. Tanaka, S., Scheraga, H.A. Medium- and long-range interaction parameters between amino acids for prediction three-dimensional structures of proteins. Macromolecules 9:945–950, 1976.
23. Warme, P.K., Morgan, R.S. A survey of amino acid sidechain interactions in 21 proteins. J. Mol. Biol. 118:289–304, 1978.
24. Narayana, S.V., Argos, P. Residue Contacts in protein structures and implications for protein folding. Int. J. Pept. Protein Res. 24:25–39, 1984.
25. Berg, O.G., von Hipple, P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J. Mol. Biol. 193:723–750, 1987.
26. Bryant, S.H., Lawrence, C.E. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. Proteins 9:108–119, 1991.
27. Fienberg, S.E. "The Analysis of Cross-Classified Categorical Data." Cambridge, Massachusetts: MIT Press, 1980.
28. Bishop, Y.M.M., Fienberg S.E., Holland, P.W. "Discrete Multivariate Analysis: Theory and Practice." Cambridge, Massachusetts: MIT Press, 1975.
29. Kendall, M., Stuart, A. "The Advanced Theory of Statistics. Vol. 2. Inference and Relationship." New York: Macmillan, 1979.
30. Becker, R.A., Chambers, J.M., Wilks, A.R. "The New S Language. A Programming Environment for Data Analysis and Graphics." Pacific Grove, California: Wadsworth and Brooks/Cole, 1988.
31. Gupta, S.S. Percentage points and modes of order statistics from the normal distribution. Annals of Mathematical Statistics 32:888–893, 1961.
32. Karlin, S., Bucher, P., Brendel, V., Altschul, S.F. Statistical methods and insights for protein and DNA sequences. Annu. Rev. Biophys. Biophys. Chem. 20:175–203, 1991.
33. Wilson, C., Doniach, S. A computer model to dynamically simulate protein folding: Studies with crambin. Proteins 6:193–209, 1989.
34. Crippen, G.M. Prediction of protein folding from amino acid sequence over discrete conformational spaces. Biochemistry 30:4232–4237, 1991.

35. Janin, J. Surface and inside volumes in globular proteins. Nature 277:491–492, 1979.
36. Viswanadhan, V.N. Hydrophobicity and residue-residue contacts in globular proteins. Int. J. Biol. Macromol. 9:39–48, 1987.
37. Ypma-Wong, M.F., Filman, D.J., Hogle, J.M., Semler, B.L. Structural domains of the polioviris polyprotein are major determinants for proteolytic cleavage at Gln-Gly pairs. J. Biol. Chem. 263:17846–17856, 1988.
38. Bode, W., Papamokos, E., Musil, D. The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin C, an elastase inhibitor from the leech Hirudo medicinalis. Structural analysis, subtilisin structure, and interface geometry. Eur. J. Biochem. 166:673–692, 1987.
39. Gros, P., Fujinaga, M., Dijkstra, B.W., Kalk, K.H., Hol, W.G. Crystallographic refinement by incorporation of molecular dynamics: Thermostable serine protease thermitase complexed with eglin C. Acta. Crystallogr. B45:488–499, 1989.
40. McPhalen, C.A., James, M.N.G. Crystal and molecular structure of the serine proteinase inhibitor Cl-2 from barley seeds. Biochemistry 26:261–269, 1987.
41. McPhalen, C.A., James, M.N.G. Structural comparison of two serine proteinase-protein inhibitor complexes. Eglin-C-subtilisin Carlsberg and Cl-2 subtilisin novo. Biochemistry 27:6582–6598, 1988.
42. Brayer, G.D., McPherson, A. A model of intracellular complexation between gene-5 protein and bacteriophage fd DNA. Eur. J. Biochem. 150:287–296, 1985.
43. Morize, I., Surcouf, E., Vaney, M.C., Epelboin, Y., Buehner, M., Frindlansky, F., Milgrom, E., Mornon, J.P. Refinement of the C222$_1$ crystal form of oxidized uteroglobin at 1.34 Å Resolution. J. Mol. Biol. 194:725–739, 1987.
44. Hogle, J.M., Chow, M., Filman, D.J. Three-dimensional structure of poliovirus at 2.9 Å resolution. Science 229: 1358–1365, 1985.
45. Weis, W.I., Brunger, A.T., Skehel, J.J., Wiley, D.C. Refinement of the influenza virus hemagglutinin by simulated annealing. J. Mol. Biol. 212:737–761, 1990.
46. Lawson, C.L., Zhang, R.G., Schevitz, R.W., Otwinowski, Z., Joachimiak, A., Sigler, P.B. Flexibility of the DNA-binding domains of trp repressor. Proteins 3:18–31, 1988.
47. Higuchi, Y., Kusunoki, M., Matsuura, Y., Yasuoka, N., Kakudo, M. Refined structure of cytochrome c3 at 1.8 Å resolution. J. Mol. Biol. 172:109–139, 1984.
48. Dill, K.A. Theory of folding and stability of globular proteins. Biochemistry 24:1501–1509, 1985.
49. Ptitsyn, O.B. Protein as an "edited" statistical copolymer? In: "Conformation in Biology." Srinivasan, R., Sarma, R.H., eds. Guilderland, New York: Adenine Press, 1983: 49–58.
50. Lau, K.F., Dill, K.A. Theory for protein mutability and biogenesis. Proc. Natl. Acad. Sci. 87:638–642, 1990.
51. Sheriff, S., Hendrickson, W.A., Stenkamp, R.E., Sieker, L.C., Jensen, L.H. Influence of solvent accessibility and intermolecular contacts on atomic mobilities in hemerythrins. Proc. Natl. Acad. Sci. USA 82:1104–1107, 1985.
52. Lesk, A.M., Chothia, C. Evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domains. J. Mol. Biol. 160:325–342, 1982.
53. Sheriff, S., Hendrickson, W.A., Smith, J.L. Structure of myohemerythrin in the azidomet state at 1.7/1.3 Å resolution. J. Mol. Biol. 197:273–296, 1987.
54. Kim, S.S., Smith, T.J., Chapman, M.S., Rossmann, M.C., Pevear, D.C., Dutko, F.J., Felock, P.J., Diana, G.D., McKinlay, M.A. Crystal structure of human rhinovirus serotype 1A (HRV1A). J. Mol. Biol. 210:91–111, 1989.
55. Herzberg, O. Refined crystal structure of beta-lactamase from Staphylococcus aureus PC1 at 2.0 Å resolution. J. Mol. Biol. 217:701–719, 1991.
56. Bowie, J.U., Clarke, N.D., Pabo, C.O., Sauer, R.T. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. Proteins 7:257–264, 1990.
57. Bowie, J.U., Lüthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
58. Lüthy, R., McLachlan, A.D., Eisenberg, D. Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins 10:229–239, 1991.
59. Overington, J., Donnelly, D., Johnson, M.S., Sali, A., Blundell, T. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. Protein Sci. 1:216–226, 1992.
60. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. 213:859–883, 1990.
61. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, M.J. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J. Mol. Biol. 216: 167–180, 1990.
62. Sippl, M.J., Weitckus, S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 13:258–271, 1992.
63. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature 358:86–89, 1992.
64. Godzik, A., Kolinski, A., Skolnick, J. Topology fingerprint approach to the inverse folding problem. J. Mol. Biol. 227: 227–238, 1992.
65. Godzik, A., Skolnick, J. Sequence-structure matching in globular proteins: Applications to supersecondary and tertiary structure determination. Proc. Natl. Acad. Sci. USA, 89:98–102, 1992.
66. Ponder, J.W., Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:773–791, 1987.
67. Lee, C., Levitt, M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. Nature 352:448–451, 1991.
68. Sander, C., Holm, L. Evaluation of protein models by atomic solvation preference. J. Mol. Biol. 225:93–105, 1992.
69. Covell, J.G., Jernigan, R.L. Conformations of folded proteins in restricted spaces. Biochemistry 29:3287–3294, 1990.
70. Taketomi, H., Kano, F., Go, N. The effect of amino acid substitution on protein-folding and -unfolding transition studied by computer simulation. Biopolymers 27:527–559, 1988.
71. Skolnick, J., Kolinski, A., Yaris, R. Monte carlo simulations of the folding of β-barrel globular proteins. Proc. Natl. Acad. Sci. 85:5057–5061, 1988.
72. Nicholls, A., Sharp, K.A., Honig, B. Protein folding and association: Insights from the thermodynamic properties of hydrocarbons. Proteins 11:281–296, 1991.