Protein Science (1997), *6*:1467–1481. Cambridge University Press. Printed in the USA.

Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment

PETER J. MUNSON[1] AND RAJ K. SINGH[2]

[1]Analytical Biostatistics Section, LSB, DCRT, National Institutes of Health, Bldg. 12A, Room 2041, Bethesda, Maryland 20892-5626

[2]Department of Computer Science, 256 Sitterson Hall, University of North Carolina, Chapel Hill, North Carolina 27599-3175

(RECEIVED December 26, 1996; ACCEPTED March 27, 1997)

Abstract

Statistical potentials based on pairwise interactions between C^{α} atoms are commonly used in protein threading/fold-recognition attempts. Inclusion of higher order interaction is a possible means of improving the specificity of these potentials. Delaunay tessellation of the C^{α}-atom representation of protein structure has been suggested as a means of defining multi-body interactions.

A large number of parameters are required to define all four-body interactions of 20 amino acid types (20^4 = 160,000). Assuming that residue order within a four-body contact is irrelevant reduces this to a manageable 8,855 parameters, using a nonredundant dataset of 608 protein structures.

Three lines of evidence support the significance and utility of the four-body potential for sequence-structure matching. First, compared to the four-body model, all lower-order interaction models (three-body, two-body, one-body) are found statistically inadequate to explain the frequency distribution of residue contacts.

Second, coherent patterns of interaction are seen in a graphic presentation of the four-body potential. Many patterns have plausible biophysical explanations and are consistent across sets of residues sharing certain properties (e.g., size, hydrophobicity, or charge).

Third, the utility of the multi-body potential is tested on a test set of 12 same-length pairs of proteins of known structure for two protocols: Sequence-recognizes-structure, where a query sequence is threaded (without gap) through the native and a non-native structure; and structure-recognizes-sequence, where a query structure is threaded by its native and another non-native sequence. Using cross-validated training, protein sequences correctly recognized their native structure in all 24 cases. Conversely, structures recognized the native sequence in 23 of 24 cases. Further, the score differences between correct and decoy structures increased significantly using the three- or four-body potential compared to potentials of lower order.

Keywords: Delaunay tessellation; fold-recognition; high-order interactions; multi-body potential; protein folding; threading potential

Increasing attention has been paid to developing empirical potentials for use in de novo folding of a protein sequence and recognition of its correct fold in a library of folds. With a few exceptions, these efforts seek to measure the suitability of the three-dimensional (3D) environment of each residue or to evaluate contributions of pairwise interactions between residues based on their fixed backbone positions. Using such potentials, it is possible to scan a large database of folds with an amino-acid sequence and generate reasonable predictions of the 3D structure for that sequence. The ability of these threading methods to accurately distinguish the correct, folded structure from moderately distorted (misfolded) structures is limited (Sippl, 1995). A similar limitation affects the quality of alignments produced by threading methods that are perhaps no better in accuracy than about one turn of an alpha helix, or about three to four residues, on average (Bryant & Lawrence, 1993).

In an attempt to enhance these potential-based methods, we investigate the contributions not just from one-body (hydrophobic) and two-body (pairwise) terms, but from three- and four-body interactions, as well. One group (Godzik et al., 1992; Godzik & Skolnick, 1992) identified certain three-body interactions but did not find significant four-body interactions. However, this group used an alternative definition of multi-body contact and did not apply explicit statistical hypothesis tests.

Reprint requests to: Peter J. Munson, NIH, Bldg. 12A, Room 2041, Bethesda, Maryland 20892-5626; e-mail: munson@helix.nih.gov.

Using a geometric accounting based on the Delaunay tessellation for three- and four-body interactions, Singh et al. (1996) suggested that four-body interactions may indeed be important. Such multi-body contributions to protein potentials are reasonable from a geometric point of view, if we approximate each residue as a sphere centered on its C^{α} location. It is possible for three or even four closely packed spheres to make mutual contact, thus giving rise to three- or four-way interactions. Three-way interactions are clearly evident as a result of the formation of disulfide bridges joining a pair of cysteines. A third cysteine is not allowed covalent attachment to the other two, and hence there appears to be repulsive interaction among such triples. Close packing of hydrophobic side chains, thought to differentiate the molten globule state from the native folded protein (Ptitsyn, 1995), are largely volumetric in character. Replacing a valine by an isoleucine (addition of a single methyl group) in the protein core would not alter statistical pairwise potentials greatly (most hydrophobic pairs display similarly favorable energies), yet this would have a measurable effect on the total volume and stability of the protein (Lim & Sauer, 1991; Lee, 1993). Volumetric constraints implied by side-chain packing would thus be expected to produce multi-body interactions in the organization of residue types within the protein interior.

Here, we define a contact potential embodying higher-order terms (those involving sets of three or four residues), and test the significance of those terms. We decompose the potential into one-body, two-body, three-body, and four-body components and show that the multi-body components are indeed statistically significant. We provide a means to visualize the many terms of the potential, and show that the sign and magnitude of the estimated multi-body terms are organized in a rational way. Finally, we demonstrate the improved ability of multi-body potentials to discriminate the correct fold by using these higher order terms.

A major difficulty in defining multi-body contact potentials is finding a good definition of the multi-body contacts themselves. In the case of pairwise contacts, most investigators have simply used a distance cutoff (such as less than 10 Å) to define a putative contact between two residues. While apparently adequate for such purposes, such a simple definition is sometimes confounded by the effect of "shielding," where a third residue lies more or less between two residues declared to be in contact. This problem is magnified if we wish to deal with sets of three or four residues. Using a simple definition of contact will produce far too many potential multi-body contacts. Consider five mutually close residues. A combinatorial approach potentially would give five different four-way interactions, but, geometrically, at most two four-body contacts would actually occur.

Recently, a well-known geometric solution to this problem was proposed for use in studying protein packing: Delaunay tessellation (Singh et al., 1996). To explain this idea, we first describe a related geometric construct, the Voronoi diagram. In such a diagram, an entire 3D volume is divided into non-overlapping regions, each region defined as a set of points closest to one particular particle (the C^{α} for example) of the protein representation. The boundary points of the regions are thus equally distant from two or more particles. Particles whose regions share a boundary are said to interact. If we connect each pair of interacting particles with a line segment, we will have the Delaunay tessellation. Just as no more than four same-sized spheres can be in mutual contact in 3D space, no more than four Voronoi regions would generally be expected to meet at a point. (There is a possibility that five or more regions could intersect, but the probability is vanishingly small for a randomly distributed set of particles, and we neglect it here.) The point of intersection of four such regions is known as the Voronoi point, the unique point that is equidistant from the four defining particles. Each Voronoi point then, corresponds to a particular set of four interacting particles, which can be alternatively represented by a tetrahedron with the particles at its vertices. The complete set of tetrahedra divides up the interior space of the protein into non-overlapping volumes and is known as Delaunay tessellation. This tessellation uniquely defines all the internal multi-body contacts in the protein.

Computational geometry researchers have made available efficient computer code for calculating the Delaunay tessellation (Barber et al., 1995; Liang et al., 1996; QHULL v. 2.3, Geometry Center, U. Minn, 1996). However, the raw tessellation is not a satisfactory geometric representation of the protein. For one thing, the tessellation produces a geometrically convex object, while proteins have various surface irregularities, binding pockets, or other concave features. Also, particles connected in the tessellation may be too distant to have a realistic chance of side chain–to–side chain interaction. For these reasons, we have found it essential to filter the tessellation to produce a more realistic representation of the internal interactions within the protein. We have filtered on two geometric criteria. First, we require that the tetrahedral edges not be too long. Second, we eliminate many of the "surface" tetrahedra that are extremely distorted in shape (very flat, extremely long, etc.). After trying several approaches, the most satisfactory criterion was to require that the circumsphere of the tetrahedron have a radius of limited size. The qualitative effects of applying this criterion are displayed for a single protein in Figure 1. The unfiltered tessellation produces the convex hull of the C^{α} particles, and disguises the real surface of the protein. Overly stringent filtering produces gaps and holes within the protein, or in the extreme allows only neighboring particles in the main chain to be connected. We anticipate that the effect of underfiltering is to mask the real multi-body interactions by spurious interactions between distant residues on the protein surface. Ultimately, the goal of this work is to demonstrate the validity of the high-order interaction model and to improve the performance of sequence-structure recognition methods. We claim that most of the important known effects (such as hydrophobic burial, salt-bridges, etc.) are included in our current formulation. Further refinements of the geometric and graphical representation are possible and should be investigated. However, the limitation of the database size does not permit an exhaustive empirical investigation of all conceivable factors. For example, we do not include a buried/exposed indicator at each particle, as this effect is largely implicit in the coordination number (more tetrahedra include a buried residue than an exposed one). Although theoretical arguments regarding the necessity of multi-body interactions in the formation of protein structure are cogent, the statistical evidence for this approach and the ability of the potential to recognize correct native protein folds will be the ultimate test of the utility of our method.

Results

Statistical comparison of potentials

The adequacy of these multi-body potentials and their associated log-linear models may be assessed by how well they explain the distribution of amino-acid four-tuples. We have fit the complete dataset with a hierarchy of five models of increasing complexity

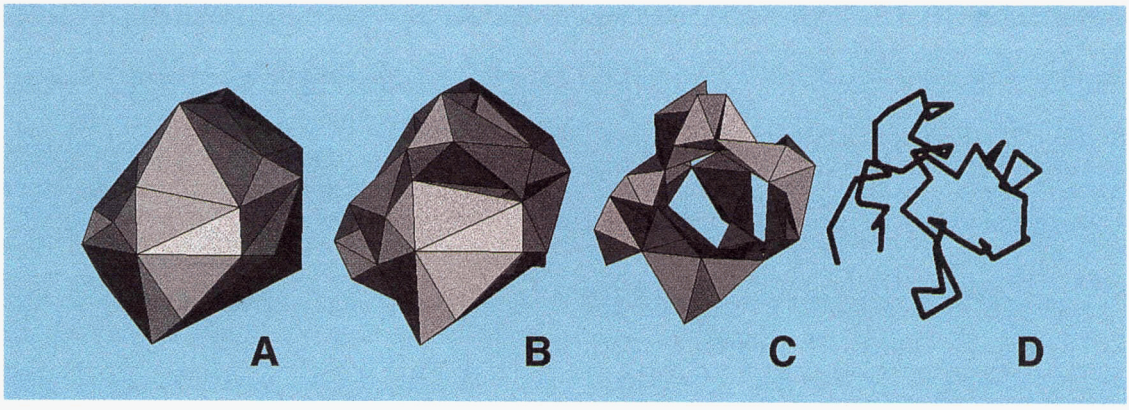

Fig. 1. A: Convex hull of ferredoxin (1fdx), using C^{α} representation. Note presence of large surface triangles with long edges representing infeasible residue interactions. **B:** Filtered Delaunay tessellation, using cutoffs described in Materials and methods, has more uniform tetrahedral face sizes. **C:** Over-filtered tessellation showing internal voids in molecule, caused by overly stringent edge length (7.5 Å) and circumsphere radius (8.0 Å) cutoffs. **D:** Main-chain backbone only.

(Table 1). Successively higher order models produce better fits to the data, as indicated by the decreasing NegLogLikelihood scores. These scores are exactly analogous to the residual sum-of-squares in ordinary linear models; lower values indicate improved fit. We also apportioned the improvement of fit to each successive order of the model. These improvements can be interpreted as "information" or "variation" explained by that level of the hierarchy.

From Table 1, we see that 49% (18,401/37,388) of the improvement is explained by the one-body terms. Such terms include the propensity of particular residue types to appear in four-body contacts, and are essentially a measure of the propensity of residues to be buried in the protein compared to the overall abundance of that residue. Not surprisingly, these terms are highly significant, achieving approximate Z-scores of 5,967 ($p < 0.001$). Adding the two-body terms explains an additional 35% (12,978/37,388) of the variation, which is also highly statistically significant ($p < 0.001$). This result coincides with well-known observations that amino acid residues interact in a pairwise specific manner, notably showing a hydrophobic-hydrophobic pair preference. This test clearly shows that this pairwise-specific interaction is not simply a consequence of hydrophobic burial, as the burial propensity is accounted for fully in the one-body model. Rather, pairwise interactions exist independently of burial status.

Next, we consider the addition of the remaining multi-body terms (three- and four-body interactions). Together, these account for 16% (6,009/37,388) of the available information and are highly significant ($p < 0.001$). As a proportion of the multi-body information (two-, three-, and four-body), the three- and four-body terms comprise 32%. The importance of the multi-body terms is thus clearly established.

Table 1. *Comparisons of hierarchical loglinear models of tetrahedra frequencies*

Model	DF[a]	NegLog likelihood	ΔLog likelihood	Multi-body information[b] (%)	ΔG^2	ΔDF	Z[c]	P-value
Baseline	1	37,388	—	—	—	—	—	—
One vs. baseline			18,401	—	36,803	19	5,967	<0.001
One-body	20	18,987						
Two vs. one			12,978	68	25,956	190	1,322	<0.001
Two-body	210	6,009						
Four vs. two			6,009	32	12,017	8,645	26	<0.001
Four-body	8,855	0						
				100				
Two-body	210	6,009						
Three vs. two			2,062	11	4,123	1,330	54	<0.001
Three-body	1,540	3,947						
Four vs. three			3,947	21	7,894	7,315	4.8	<0.001
Four-body	8,855	0						
				32				

[a]Degrees of freedom or number of parameters in model.
[b]Information, expressed as percentage of One-body NegLogLikelihood.
[c]Standard normal deviate, $Z = (\Delta G^2 - \Delta DF)/sqrt(2 * \Delta DF)$.

Finally, we attempt to dissect out the three- and four-body terms separately (Table 1, last rows). While the three-body terms are clearly significant ($Z = 54$, $p < 0.001$), the four-body terms alone produce a Z-score much closer to the nominal significance threshold (3.08 for $p = 0.001$, one-tailed test). Nevertheless, the four-body terms explain more (21%) of the multi-body information than do the three-body terms. One the other hand, four-body terms incur a larger number of parameters (7,315), compared to the three-body terms (1,330). Given the large number of parameters and the approximate nature of the statistical test used here, one might wish to attach only provisional significance to the four-body terms at this point. However, many of the three- and even four-body terms have plausible molecular explanations, adding weight to the overall argument for significance, as we now demonstrate.

Visualization of the potentials

To look in detail into a complex potential function is problematic when there are so many (8,855) distinct energy terms. Extending the work of Miyazawa and Jernigan (1983), we represent the energy terms as graphical array, extended to a four-way array to represent the four-body potential.

Four-body potential

Figure 2 plots of the full four-body potential as a "conditional plot" or *coplot*. The major configuration of the potential is clear: The upper leftmost window-panes of the plot indicate the favorable values associated with the hydrophobic residues. Although each hydrophobic residue pair shows a generally favorable potential value overall, there is an even more favorable potential value attained if three or all four residues in the tetrahedron are hydrophobic (upper left portion of individual panes in upper left of plot). In the lower right, there is a single, intensely colored pane, reflecting the strongly favorable energy associated with C-C pairs, mostly of disulfide bridged cysteins. Like-charged pairs (K-K, R-R) show an unfavorable potential value, with more favorable values shown for oppositely charged pairs, especially if the third and fourth residues are hydrophobic.

The pattern within each pane is highly variable as we scan across the full plot. These variations portray the higher-order in-

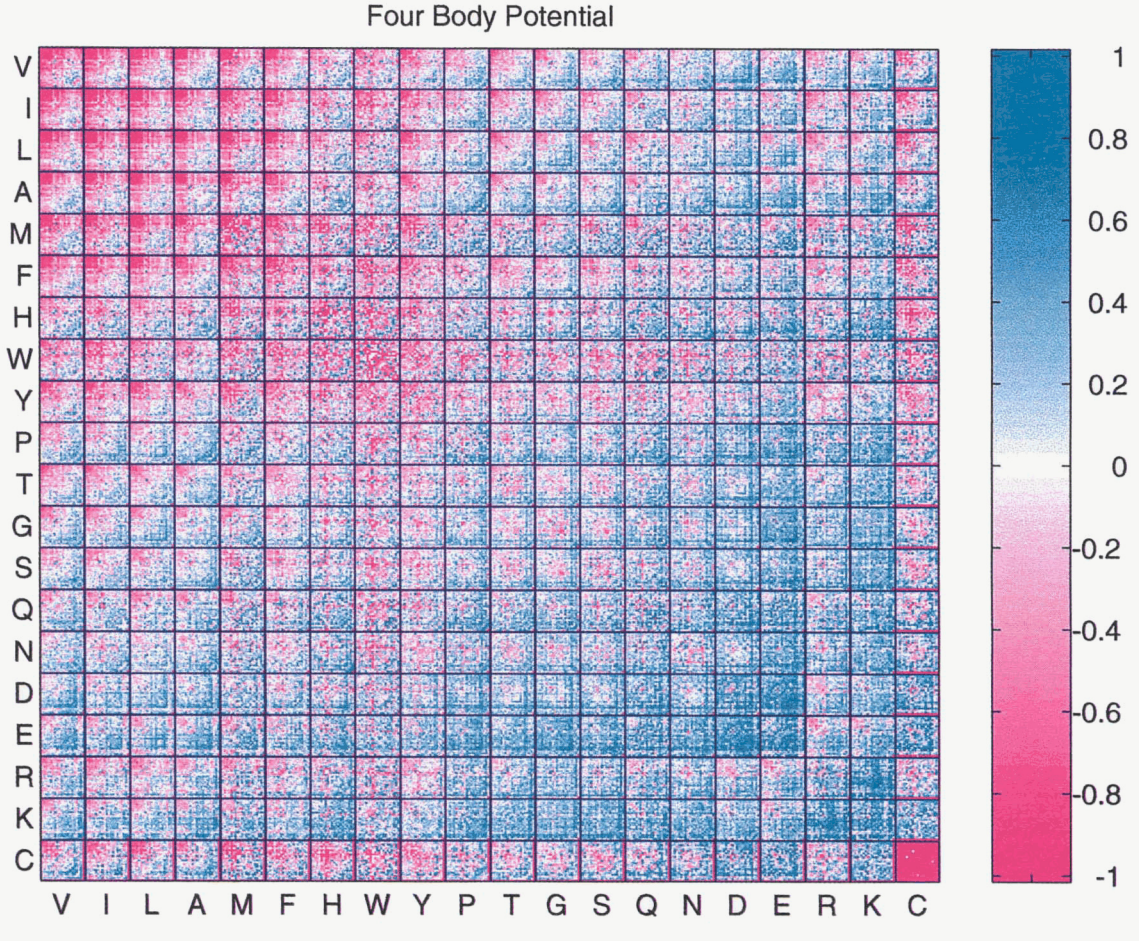

Fig. 2. Four-body potential color conditional plot. The co-plot is organized as a 20 by 20 array of "window panes" each containing a 20 by 20 pixel image. The panes are indexed by the first two amino acid residue in the tetrahedron, and the pixels within each pane are indexed by the third and fourth residue, in the same order. Potential values are colored: magenta represents favorable interactions, cyan represents unfavorable interactions. The C-C pane is solid magenta indicating a strong propensity for C-C pairs to form. Hydrophobic interactions are favorable, in upper left panes.

teractions and can be dissected from the full potential by subtraction of lower-order potentials.

Two-body potential minus one-body potential

We may isolate just the pairwise components of this potential by constraining the high order interaction terms to be zero. We remove the one-body hydrophobic burial effect by subtracting the one-body potential. The remainder represents the contribution of purely pairwise interactions of amino-acid residues, not just the general tendency for hydrophobic burial. This portion represents 36% of the total information in the full potential and is shown in Figure 3. The dominant effect is the C-C interaction term (lower right pane), resulting largely from disulfide bonded cystine pairs. The purely hydrophobic residues (V, I, L, A) show a definite preference to form contacts, even after removing the burial propensity of the individual residues. Oppositely-charged residues now show a strong tendency to pair up, while pairs bearing the same charge have an unfavorable potential value. There is also a tendency for the polar residues to show an attraction, as reflected in the brightly colored central panes of the plot. From the last row of panes in Figure 3, we note that, when not in a C-C pair, C shows a preference for other polar residues, and an antipathy for the hydrophobic residues.

Four-body potential minus two-body potential

We now turn to the higher-order terms in our potential. By subtracting all terms up to pairwise, we may study any new features arising from just these higher order interactions. Together, the three- and four-body interactions account for 17% of the available information (32% of the multi-body information, Table 1) and are clearly statistically significant overall.

The potential difference (four-body minus two-body) does not show any obvious patterns when visualized, owing to statistical variability. Some four-body combinations, especially those involving the rarer residues, show greater fluctuation in potential value than others. To remove this statistical artifact, we calculate the Freeman-Tukey standardized residual described in Materials and methods. This calculation produces a number which has nearly uniform variance, and facilitates finding patterns in the multi-body component of the potential. Selected panes of the full co-plot are shown in Figures 4 and 5. Figure 4 demonstrates the existence of multi-body interactions involving hydrophobic (VIL), small (AG) and cystein (C) residues. It is clear that a strong signal appears in the C-C pane, where we see an overall propensity for C-C-polar-polar tetrahedra (lower right quarter of C-C pane is magenta). Conversely, there is a tendency for tetrahedra involving a single C to include three other hydrophobics (e.g., C-L pane, upper right is magenta). There is also a notable under-representation (cyan color)

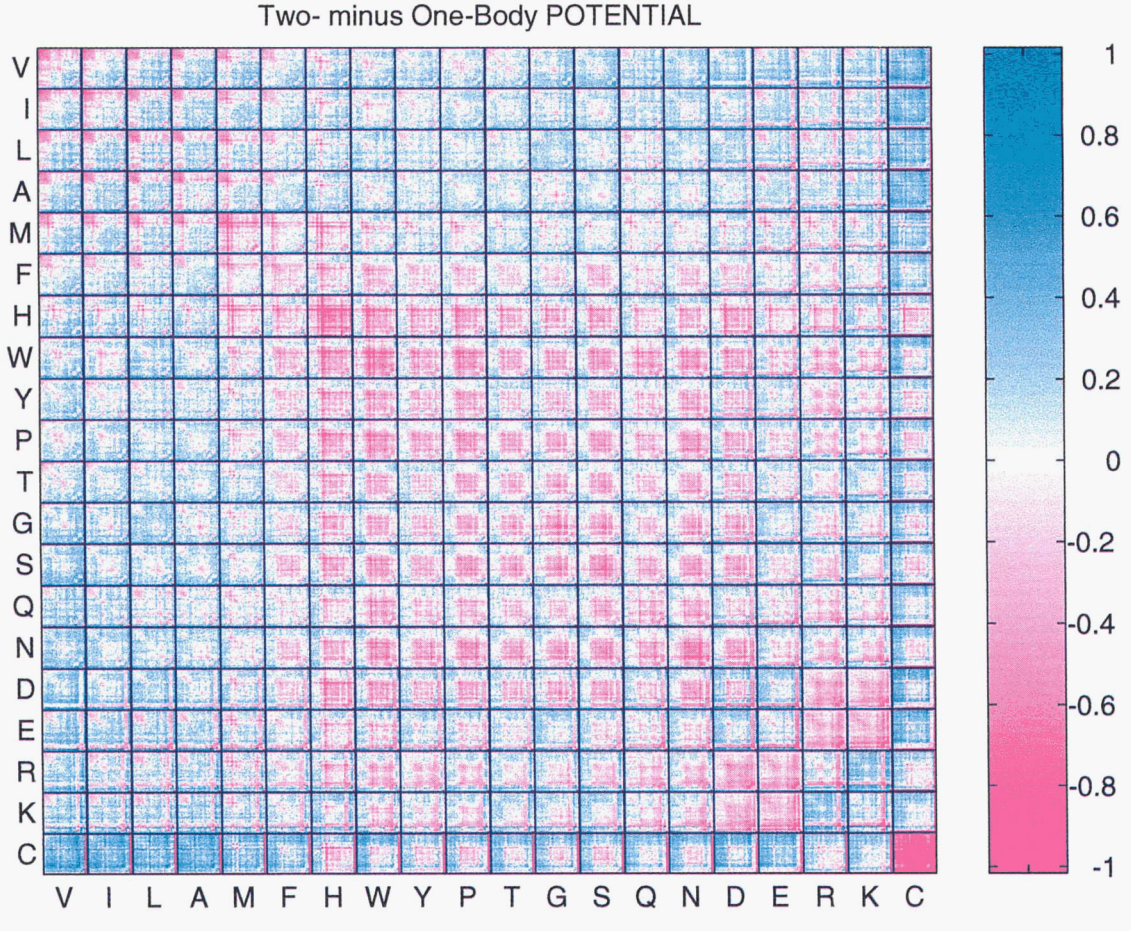

Fig. 3. Two-body minus one-body potential co-plot, arranged as in Fig. 2. The pure pairwise component of the potential is represented showing favorable pairwise interactions among hydrophobics, among polar residues and a very strong pairwise interaction in C-C pairs.

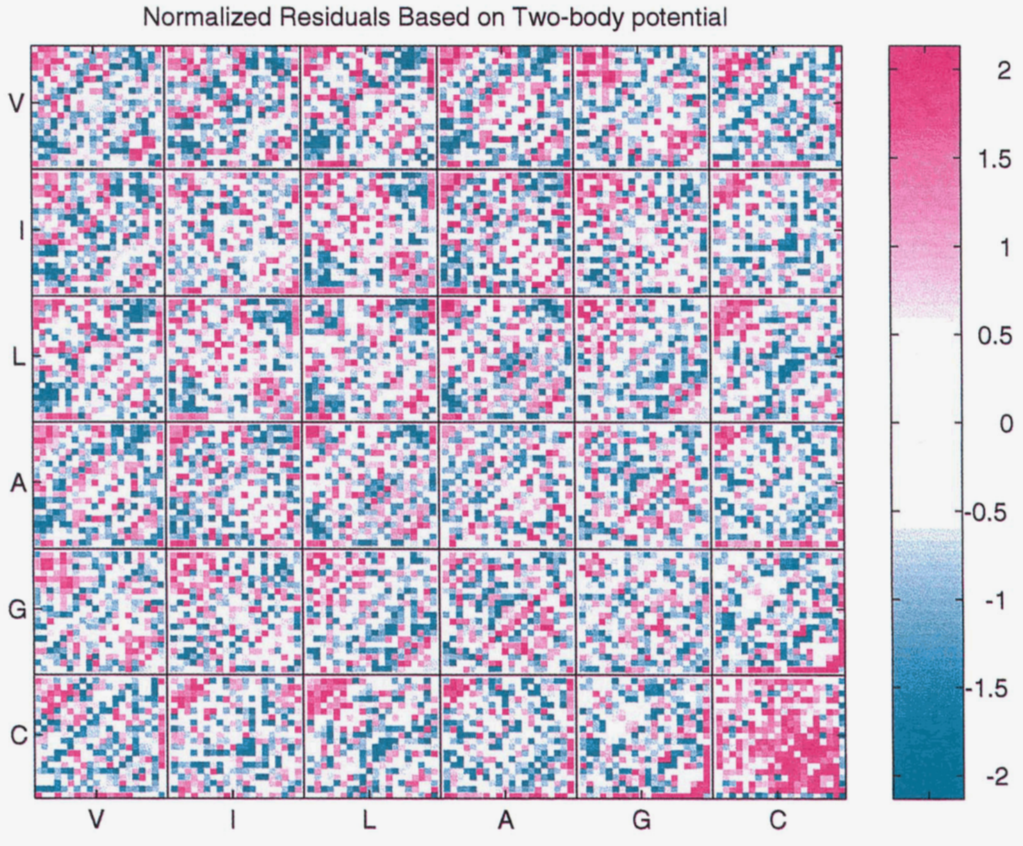

Fig. 4. Normalized residuals based on two-body potential. Positive values (magenta) represent tetrahedra which are significantly more common than would be expected from two-body interactions. Negative values (cyan) indicate under-represented tetrahedra. Arrangement of pixels within each pane is as in Fig. 2. Only 36 of the 400 panes are shown, emphasizing the interactions among hydrophobic (VIL), small (AG), and cystein (C) residues.

of the C-C-C-x patterns, compared to predictions based only on the two-body potential.

Several very interesting patterns can also be found in the hydrophobic panes of Figure 4. First, we can see a general tendency to form hydrophobic quartets (e.g., upper left portion of the V-V pane is magenta). This is clearly an indication of the formation of hydrophobic clusters beyond the prediction of the pairwise potential. Second, there are some remarkable and interesting patterns even within the associations of four hydrophobic residues. We see a stronger propensity to form hydrophobic clusters when one of the four residues is the small amino acid A or G (alanine or glycine). This is seen as the bright magenta upper-left portion of Figure 4, panes A-V, A-I, A-L G-V, G-I and G-L, but not in the A-A or G-G panes.

Figure 5 demonstrates the presence of multi-body interactions involving the charged residues DERK and the residues QN. Paired, oppositely charged residues tend to avoid other polar or charged residues, and prefer patterns with hydrophobic residues (e.g., in the R-E pane, the upper left is magenta, the lower right is predominantly cyan).

Conversely, clusters of three hydrophobic residues tend to avoid charged residues (e.g., Fig. 4, pane V-L, upper right corner is cyan). Thus, as previously noted by Godzik (Godzik et al., 1992), while charged residues are generally exposed to solvent and should thereby avoid hydrophobic residues, paired charges of opposite sign show an increment in favorable energy toward other hydrophobics.

Pure four-body components

The pure four-body interactions were inspected in co-plots analogous to Figures 4 and 5. However, few patterns emerged. One exception is the pattern in the C-C pane, where a stabilizing term for the C-C-C-C tetrahedron and destabilizing terms for C-C-C-*polar* tetrahedra could be found. The CCCC term is based on 93 tetrahedra occurring in 38 distinct proteins, which suggests that such four-body interactions involving four cysteins are broadly based, and not an artifact present in just a few proteins.

Although no other simple patterns were evident in this co-plot, we cannot easily discount the remaining large deviations. For example, a suggestive pattern of both strong positive and negative components appeared in tetrahedra with the catalytically active residues S-H (data not shown), which might be a reflection of the tendency to organize these residues in specific patterns around active sites of the protein.

In Table 2, we have ranked all pure four-body interactions according to statistical significance. One four-body component (CCCC) stands out with a positive residual value of 5. Inspection of the top (positive residuals, favorable component) and bottom (negative residuals, unfavorable component) of this list reveals other suggestive patterns. Several entries are examples of the patterns seen in Figures 4 and 5. For example, ADFK, ERVV, and EIKV all show the favorable component arising from packing an oppositely charged pair next to two hydrophobic residues. For

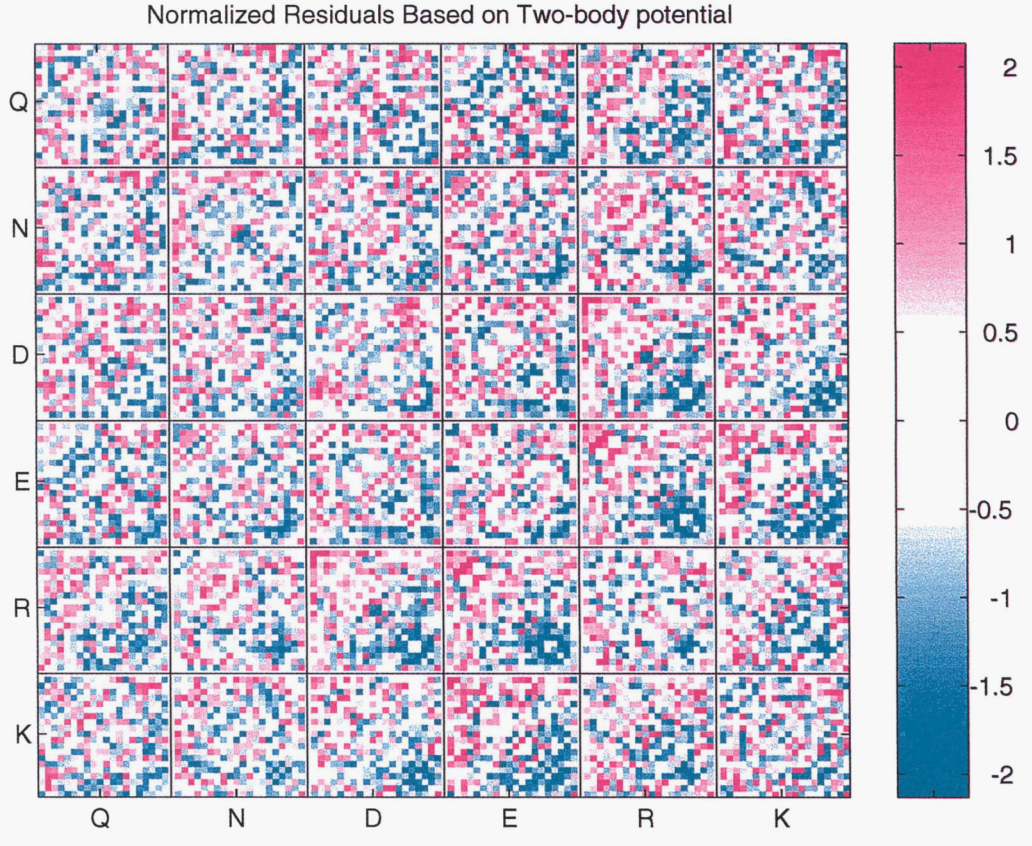

Fig. 5. Normalized residuals based on two-body potential. Arranged as in Fig. 4, but for the charged and polar residues (QNDERK).

these three tetrahedra, there are 457 occurrences where only 358 are expected. The quartet EIMV is an example of the propensity of an unpaired charge (E) avoiding a hydrophobic cluster (IMV). Only 34 such tetrahedra are observed where 56 would be expected based on pairwise interactions. Although the statistical significance for many of the terms in this table is borderline, these residue combinations may still reflect important packing tendencies whose importance may emerge with detailed analyses of their molecular interactions.

Sequence-structure recognition

To assess the true utility of the four-body potential as a discriminator of correctly folded structures, we performed two structure/sequence discrimination tests. These tests were performed in a fully cross-validated manner, i.e., sequences with significant homology to the test set proteins were removed from the training set. The test uses the potential function to first recognize the correct structure (sequence-recognizes-structure) and then to recognize the correct sequence for a given structure (structure-recognizes-sequence). The first problem is more relevant for choosing the correct threading target from a library of protein folds. The second is useful for designing sequences which have unusual stability in a particular structure. Twelve pairs of same-length proteins were used, with no sequence or structural homology evident within pairs.

The four-body potential was successful at recognizing the correct structure of the two presented to it in all 24 cases (Table 3). In recognizing the correct sequence, the potential failed only once. Do the high-order terms contribute significantly to the success of the multi-body potential? To answer this, we performed the structure recognition test using the two-body, three-body, and four-body potentials and compared the score difference of the native and decoy structures. We observe that the four-body scores generally decrease with sequence length (Fig. 6), meaning that the correct fold for larger structures should be more easily distinguished than for smaller structures. In fact, all the potentials tested (one-, two-, three-, and four-body) could correctly distinguish all pairs of structures. However, the higher-order potentials yielded more negative scores, indicating a greater discriminatory power. By subtracting the score differences (four-body minus three-body, three-body minus two-body) we measure the degree of improvement. There is clearly an improvement due to the inclusion of three-body terms (Fig. 7) of about -3 units averaged over proteins ($t = 4.2$, 23 *df*, $p < 0.001$). A smaller improvement of an additional -1.2 to -1.9 units is obtained using the four-body potential, making the total improvement for the multi-body potential about -4 to -6 units. While modest, this improvement still represents an additional 28% of the improvement obtained using the two-body potential in the same circumstances.

Discussion

In summary, we have presented three lines of evidence for the existence of three- and four-body interactions governing the packing of residues into three-dimensional protein structures. The statistical evidence follows from the formal comparison of the adequacy of higher-order log-linear models to explain the observed frequen-

Table 2. *Most important pure four-body components*

	Standardized residual	Observed frequency	Expected frequency	Obs./Exp.	Identity
		Favorable interactions			
8855	5.04[a]	93	50.88	1.828	CCCC
8854	3.62[b]	33	15.58	2.118	MQRS
8853	3.28	48	28.09	1.709	GPWY
8852	3.17	145	109.5	1.324	ADFK
8851	3.01	150	115.6	1.297	ADEV
8850	2.91	30	16.3	1.841	AHSW
8849	2.83	66	45.19	1.46	AKKY
8848	2.80	59	39.61	1.489	DHLN
8846	2.79	101	75.08	1.345	DIKN
8847	2.79	111	83.73	1.326	ERVV
8845	2.72	36	21.65	1.662	ACDF
8844	2.72	44	27.96	1.574	EQSS
8843	2.70	25	13.42	1.862	DEMQ
8842	2.69	201	164.9	1.219	EIKV
8841	2.69	73	52	1.404	ADKM
8840	2.65	163	131.1	1.244	GLNS
8839	2.63	51	34.11	1.495	KLMN
		Unfavorable interactions			
15	−2.97	178	220.1	0.8088	ADGV
14	−2.97	5	14.41	0.347	ACEH
13	−2.99	8	19.19	0.4169	EEGG
12	−3.00	22	38.75	0.5677	GMRV
11	−3.03	1	7.148	0.1399	CPPR
10	−3.08	44	67.15	0.6553	LLPR
9	−3.14	2	9.617	0.208	FPQQ
8	−3.24	34	55.89	0.6083	EIMV
7	−3.28	29	49.77	0.5827	EKQS
6	−3.30	5	15.67	0.319	HIQY
5	−3.30	8	20.58	0.3886	CGNQ
4	−3.43[b]	10	24.28	0.4118	MSTY
3	−3.53[b]	18	36.57	0.4922	ANNP
2	−3.58[b]	18	36.82	0.4888	GLNW
1	−3.67[b]	6	18.97	0.3163	ETYY

[a]Indicates highly significant standardized residual.
[b]Indicates a residual with absolute value greater than expected in sample of 8855 normally distributed values.

cies of four-body contacts in known protein structures. Visualization of these complex interactions leads to identification of specific interactions and groups of interactions for which plausible biophysical explanations can be given. Finally, testing the potentials in a cross-validated setting shows that the high-order terms augment the discriminatory power to a small but significant degree. This finding should be useful in protein-fold recognition.

Among the significant multi-body interactions, many involve cystein pairs, triples, and quadruples. Almost certainly these statistical interactions have a molecular basis in the formation of disulfide bridges. Since such bridges do not form between triples, there is an apparent "repulsive" three-way interaction and a compensating "attractive" four-way interaction. The placement of cystein pairs within the protein is not uniform but appears to prefer polar rather than hydrophobic neighbors. We have found interactions involving charged and hydrophobic residues. That such residues interact strongly is easy to understand. That they interact statistically in a three-body and four-body manner is quite intriguing. The difference of the multi-body interaction between same-charged pair plus hydrophobic versus oppositely charged pair plus hydrophobic ought to be rationalizable in terms of electrostatics, but the origin of the attraction of charged pairs for hydrophobics is still obscure. We have detected a clear hydrophobic clustering signal, beyond that accounted for by pairwise interactions. This feature is compatible with the observed hydrophobic character of the protein interior, but is clearly omitted from any potential which includes only pairwise terms. We also see evidence for a preferred alternation of hydrophobic side-chain size within the protein core, namely a multi-body potential preference for residue quadruples obeying the big-big-small rule, such as V-I-G-x or L-I-A-x. Although the evidence for this rule (Fig. 4) is preliminary, it could become important for engineering proteins which pack tightly into unique structures or for recognizing the correct packing arrangements in protein folding experiments. Finally, we have produced a list (Table 2) of over- and under-represented quartets of residue types. Molecular-based explanations for these propensities should be sought, as they may reveal some undisclosed aspects of the protein-folding code.

Investigation of multi-body interactions requires careful attention to the methodology. In our work, we have utilized a clear consistent geometric representation of the three-dimensional structure of the protein. Our definition of contacts is largely insensitive to the choice of length cutoff, as the Delaunay tessellation is essentially a nearest-neighbor approach. Use of the filtering has made the tessellation significantly more representative of important residue interactions, as sterically infeasible interactions are removed. Our definition of the four-body potential incorporates all lower-order interactions, and implicitly includes the very important hydrophobic burial tendencies of certain residues. We have carefully partitioned the information attributable to each level of model complexity through use of a hierarchical statistical model. This partitioning facilitated the calculation of statistical significance and the visualization of the potential components. Lastly, we used a carefully cross-validated training/testing procedure to assure that the improvements is structure recognition for the multi-body potential were not artifactual.

There are more steps to be taken to refine and enhance this empirical potential, and further development of the underlying methodology is needed. Many of these topics are under active investigation in our group. Our current protein representation is crude; one particle per residue. Other atoms beyond the C^{α} could be included. Comparison of our filtered tessellation with the geometrically more elegant alpha complex approach (Liang et al., 1996) is needed to confirm the adequacy of our method. Other factors about the 3D environment of each residue could also be incorporated such as the local main-chain configuration, local hydrogen-bonding pattern, the secondary structure status, the side-chain orientation, and most importantly, the solvent-exposure status of each residue. Long-range (second-order contacts, third-order contacts, etc.) might also be investigated in the multi-body potential. Judicious choice among these factors is required due to the limited data set size but some combinations may permit still better distinction between correctly and incorrectly folded structures.

The statistical methodology also needs to be advanced in part to account for the low frequencies in many of the categories. Suitable Bayesian priors may play an important role here. The underlying theory, which forms the basis for estimating the potential, needs to be addressed. We are considering alternate estimation strategies more suited to combining the various factors described above.

Table 3. *Structure and sequence recognition test for 12 pairs of same-length proteins*

PDB id.	Protein name (class)	Length	Four-body potential scores			Two-body scores	
			A[a]	B[b]	C[c]	D[d]	E[e]
1cbh	Cellulase tail domain	36	−61	2	−63	−77	−51
1ppt	Avian pancreatic polypeptide	36	0	16	−17	−2	−16
1fdx	Ferredoxin ($\alpha + \beta$ class)	54	−69	−5	−64	−99	−64
5rxn	Rubredoxin	54	−31	30	−61	−26	−56
2ci2	Chymotrypsin inhibitor	65	−19	12	−31	−32	−32
2cro	Cro repressor (α class)	65	−9	13	−22	−21	−23
1hip	High-potential iron protein (β)	85	−6	39	−45	−48	−37
2b5c	Cytochrome b5	85	−4	42	−46	−43	−42
2cdv	Cytochrome c3(α class)	107	12	55	−43	−4	−41
2ssi	Subtilisin inhibitor($\alpha + \beta$ class)	107	−34	16	−51	−89	−47
1bp2	Phospholipase A2($\alpha + \beta$ class)	123	−68	11	−80	−116	−70
2paz	Pseudoazaurin (b class)	123	−63	48	−111	−74	−104
1p2p	Phospholipase A2($\alpha + \beta$ class)	124	−78	−21	−57	−118	−42
1rn3	Ribonuclease	124	−54	40	−95	−33	−87
2i1b	Leghaemoglobin (α class)	153	11	67	−56	7	−53
1lh1	Interleukin 1b(α class)	153	−46	4	−49	−113	−47
1rei	Bence-Jones protein (β class)	214[f]	−21	50	−72	−65	−83
5pad	Papain (multi-domain)	214	−90	44	−134	−140	−130
1rhd	Rhodanese (α/β class)	293	−46	85	−132	−185	−133
2cyp	Cytochrome peroxidase (multi)	293	0	139	−139	−85	−136
1abe	Arabinose-binding protein(α/β)	306	−86	55	−141	−240	−128
1pmb	Myoglobin dimer(α class)	306	−10	154	−164	−65	−168
2tmn	Thermolysin (multi-domain)	317	−31	59	−90	−117	−75
2tsl	Tyrosine-tRNA synthetase (α/β)	317	−28	86	−114	−87	−105

[a]Potential energy for protein sequence threaded through native structure.
[b]Potential energy for sequence threaded through decoy structure.
[c]Difference: D−E is negative if sequence recognizes structure.
[d]Difference: D−decoy E, native compared to decoy sequence in same structure, is negative if structure recognizes sequence.
[e]Two-body potential energy difference for sequence recognizes structure.
[f]Protein actually has two identical chains of length 107.

A fuller demonstration of the power of multi-body potentials also requires more computationally intense optimal alignment algorithms such as simulated annealing, branch-and-bound, or Gibbs sampler. It would also require threading a larger test database of structures. Substantial headway has been made in these areas for pairwise potentials, and many of those solutions should be directly applicable to the four-body case, as well.

More immediately, we intend to explore the details of the three-body interactions, in light of the classification of tetrahedra by edge-bonding pattern (Singh et al., 1996). We anticipate that much stronger multi-body interactions will be found as the filtering criteria are further optimized, and as the interplay of interaction distance and main-chain configuration are explored. Likewise, the reduction of the 20-letter amino acid alphabet to fewer letters will permit exploration of the other factors described above. A recent paper (Zheng et al., 1997) and our own preliminary work suggests this to be a promising idea.

We conclude that the influence of the three- and four-body interactions on fold recognition is clearly significant. Although this influence may currently seem small, we anticipate that taken in combination with other structural features (secondary structural state, surface exposure, local vs. distant packing), multi-body rules will be found to be increasingly significant. The packing preferences of side chains clearly involve more than residue pairwise terms. In particular, volumetric constraints are essentially multi-body in character. Our approach provides a means to incorporate such higher-order interactions consistently into threading approaches.

Multi-body potentials are also important for de novo protein folding (Godzik et al., 1993). Identification of the truly significant multi-body contacts might have a big influence upon attempts to fold proteins using hierarchical condensation (e.g., Srinivasan & Rose, 1995) as the course of the dynamic simulation seems to be heavily determined by the early formation of clusters of mutually contacting residues. Understanding the interplay of multi-body, pairwise, and one-body potential energy terms may prove to be crucial to this important problem.

Materials and methods

There are many facets to this method for developing a useful and satisfactory potential function with multi-body terms. First, we need a particle representation of the protein structure, i.e., a finite

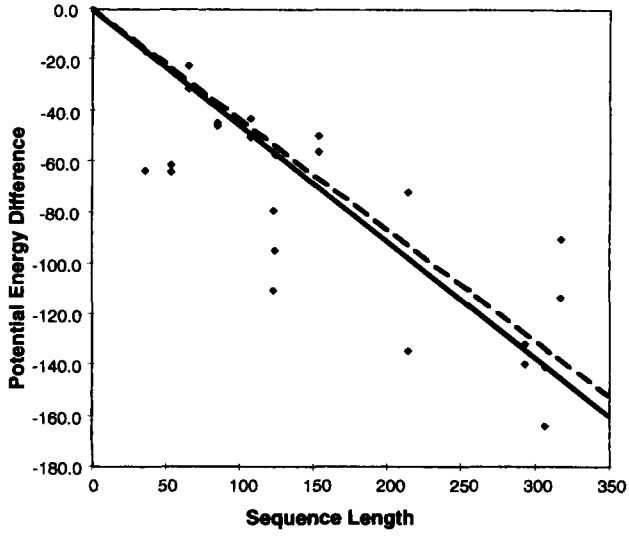

Fig. 6. Score difference (sequence threaded through native structure minus that through decoy structure) vs. sequence length for 24 pairs of protein structures using four-body potential (diamonds). Solid line is regression, showing increase of absolute difference with length of sequence. Dashed line is regression for two-body potential applied to same 24 pairs. Difference between lines represents increased discriminatory power of four-body potential.

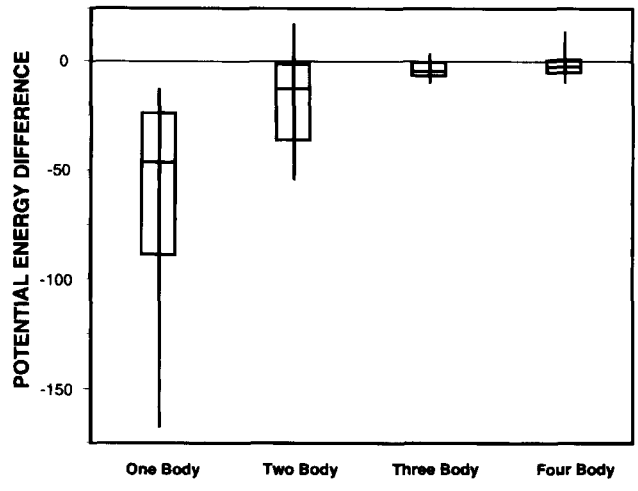

Fig. 7. Contribution to score difference between sequences threaded onto native and decoy structures, from each step of model complexity. Box plots show median (central line), middle 50% (box), and range (extent of vertical line) of values for 24 proteins given in Table 2. The discrimination is improved at each step, as the median values are all less than zero. Each contribution is statistically significant (t-test results: one-body, $p < 0.001$; two-body, $p < 0.001$; three-body, $p < 0.001$; four-body, $p < 0.03$ using the signed rank test). Mean $\pm$ standard errors are: one-body, -59 ± 9; two-body, -15 ± 4, three-body, -3 ± 0.7, four-body, -1.2 ± 0.99. One value in the four-body contribution is clearly an outlier (arising from threading 2 half-length chains, 1REI, onto full-length decoy structure, 5PAD). Ignoring this value, the mean becomes -1.9 ± 0.75.

list of labeled points in space which define the protein. Second, we deal only with a contact potential, so that per se pairwise distances will not appear in the potential terms other than as a cutoff value. There are several ways in which distance is incorporated into high-order terms (namely as surrogates for area or volume), and we investigated some of them. Third, we need a consistent means of deciding which four-tuples of particles in fact most significantly interact within the structure. This is supplied by Delaunay tessellation and its filtered version. Significant computational geometry difficulties must be surmounted here. Fourth, we need a means of estimating the potential function for each possible order of interaction, from a database of proteins. Our approach here is to build a log-linear model (Bishop et al., 1975) of the observed frequencies of each type of four-tuple (tetrahedron) of associated residues. The maximum likelihood estimates of the parameters of this model are related to the energy terms in the potential function. The log-linear model permits a hierarchical decomposition of the frequency variation into a constant term and first-order through fourth-order interaction terms. The first order terms of the log-linear model measure the propensity of particular amino acid residues to appear in four-tuples. Comparing these propensities to the prevalence of residues in the overall database measures the tendency for certain residue types to be buried.

After developing the new potential, we also justify its complexity relative to the more commonly used pairwise potential. We do this formally in two ways. First, we compare the contribution of each level of the hierarchy in the log-linear model, either in terms of the information it provides, or equivalently, in terms of the likelihood of observing the data, given that level of description (singles, doubles, triples, etc.). For this comparison, an approximate statistical test is applied. Second, we evaluate sensitivity of the potential for differentiating the fit of a native sequence to its native structure from its fit to an alternate structure (or conversely, to distinguish the fit of native-sequence to native structure from that of another sequence to the same structure). Again, we test whether the high-order terms are warranted. In both tests, we are acutely aware of the problem of memorization of the database faced by machine-learning tools. In the statistical literature, this is known as the problem of over-parameterization. We deal with this problem by accounting for the number of parameters estimated (in the case of the log-linear model hierarchy) or by using a training-set/test-set approach (for the sequence-structure recognition) where the training set has been purged of proteins bearing significant homology to test-set proteins.

Structure representation

Following a now standard practice, we use a single particle to represent each amino acid residue of the polypeptide chain, with the pseudo-atom placed at the C^{α} position. Alternatively, we could have placed the particles at the C^{β} positions, or used a combination of such particles. These possibilities will be investigated in the future. Each particle is labeled by the residue type it represents, using the standard 20 letter code. No distinction is made between cross-linked cystines and the free cysteine residue. The influence of metal ions, heme groups, etc. is also omitted from this analysis.

Contact potential

A contact potential measures the overall energy of a system, where the system is described by the presence or absence of contacts of specific types. Because we are considering four-way contacts, our potential energy for a specific protein of size M can be calculated as

$$E = \sum_{i=1}^{M-3} \sum_{j>i}^{M-2} \sum_{k>j}^{M-1} \sum_{l>k}^{M} \Gamma_{ijkl} E(s_i, s_j, s_k, s_l)$$

where Γ_{ijkl} is an indicator of a four-way contact between residues i, j, k, and l, s_i is the amino acid type of the i^{th} residue, and $E(s,t,u,v)$ is the energy associated with the tetrahedron having amino acid types s, t, u, and v at its vertices. The sums may be reordered and terms collected to give

$$E = \sum_{i=1}^{20}\sum_{j=1}^{20}\sum_{k=1}^{20}\sum_{l=1}^{20} n^{o}_{ijkl} E_{ijkl}$$

where n^{o}_{ijkl} is the number of four-body contacts observed among four residues ordered along the backbone of the protein, of types i, j, k, and l, respectively, where i, j, k, and l are now one of the twenty naturally occurring amino acids. Assuming symmetry in the energy over all permutations of the four subscripts ($E_{ijkl} = E_{ijlk} = \ldots = E_{lkji}$), we may write the full potential as a sum of not 160,000, but only 8,855 terms:

$$E = \sum_{i=1}^{20}\sum_{j=i}^{20}\sum_{k=j}^{20}\sum_{l=k}^{20} n_{ijkl} E_{ijkl}$$

where $n_{ijkl} = \sum n^{o}_{\sigma(ijkl)}$, the sum over all distinguishable permutations σ of the four indices.

Of course, high order contacts imply contacts of all lower orders, so by definition, a four-body contact potential implies at least the complexity of pairwise and three-body contact potentials. The division of variation at different levels is accomplished by the use of the hierarchical log-linear model described below.

The contact potential is a relative energy, where the baseline is chosen to be a random protein with average composition corresponding to a database. However, the statistical expectation of the potential is not necessarily zero, although the expectation can be adjusted by simply adding a constant.

Which four-tuples?

Filtering the Delaunay tessellation

Our formulation for a four-body contact potential requires a definition of Γ_{ijkl} for each protein. As discussed above, simply requiring short pairwise distances among a set of four particles is not sufficient to avoid the possibility of overlapping sets of four residues, i.e., the tetrahedron connecting one set of four particles might overlap the tetrahedron connecting another set of four. To avoid this geometric difficulty, illustrated in two-dimensions in Figure 8, we resort to the Delaunay tessellation to define a complete set of such contacts Γ_{ijkl} for each protein. This tessellation includes only the closest sets of four particles in the sense that if a set of particles *ijkl* is included, then the circumsphere defined by those four is guaranteed to be devoid of any other particles (i.e., no other particle is mutually closer to the other three than the fourth particle in the set). This is tantamount to assuming that only the nearest four-tuples actually contribute to the contact interaction, which is quite reasonable.

Formally, the Delaunay tessellation is the mathematical dual of the Voronoi diagram. The Voronoi diagram is a partitioning of three-dimensional space into regions or neighborhoods of the original particles (e.g., the C^{α} carbons). Each region represents that portion of space that is closest to one particular particle. The Voronoi regions form polyhedral cells analogous to those in a beehive, but of irregular size and shape. Two Voronoi regions that share a boundary are connected in the Delaunay tessellation. Sets of four particles that are mutually connected (all six pairwise connections) represent four-body contacts, and correspond to $\Gamma_{ijkl} = 1$. Thinking of the pairwise connections as straight line-segments between the C^{α} carbon locations, the four-body contacts are the tetrahedra with these line segments as edges. Interestingly, the probability of having any contacts of order higher than four is zero. To picture why, consider close-packed same-sized spheres. It is easy to picture two, three, or four spheres in mutual contact. But it is impossible to place a fifth sphere in mutual contact with a set of four mutually contacting spheres. Thus, four-body contacts exhaust the possible complexity of this representation.

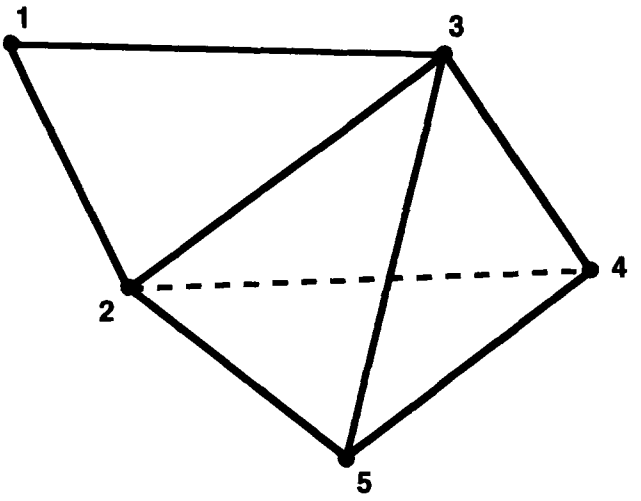

Fig. 8. Delaunay tessellation of five points in 2D space divides space into non-overlapping triangles. The triangles 2-3-5 and 3-4-5 are chosen rather than 2-3-4 and 2-4-5 since the line segment 3-5 is shorter than 2-4. Accordingly, the interaction between particles 3 and 5 is said to "screen out" the potential interaction between 2 and 4. An analogous situation pertains to 3D space, which is divided into non-overlapping tetrahedra by the Delaunay tessellation.

It is immediately apparent from Figure 1 that not all high-order contacts implied by the Delaunay tessellation are suitable for inclusion in our study. Many connections in the tessellation are far too long for residue–residue interaction to be plausible. Thus, we filter the raw Delaunay tessellation. The first criterion is to remove any edges of length greater than 9.5 Å. The value was chosen as a compromise, after observing the effect of the cutoff on the number of available tetrahedra and the appearance of typical proteins under different cutoffs. Table 4 shows this effect for cutoffs from 4–12 Å. Excessively short cutoffs create too many apparent voids inside protein cores, while long cutoffs allow too many tetrahedra of extensive shape on the surface of the protein (Fig. 1C vs. Fig. 1A).

The second filtering criterion arises naturally from the properties of the Delaunay tessellation applied to proteins. One characteristic of the tetrahedra in the tessellation is the circumsphere radius. This is the distance from the Voronoi point to each of the four vertices. A small radius implies that the four residues are intimately associated while larger radii suggest that the association between the four is tenuous. Figure 9 shows a distribution of radii for our dataset, and indicates that the majority (77%) of circumspheres have radii less than 9.0 Å. Inspection of tetrahedra with excessively large radii shows them to be primarily on the surface of the protein. Thus, to avoid these spurious interactions, tetrahedra with circumsphere radii larger than 9.0 Å were filtered out.

Table 4. *Proportion of 689,549 tetrahedra excluded by the maximum pairwise distance cutoff or by the circumsphere radius cutoff*

Maximum pairwise distance (Å)	Proportion excluded (%)	Circumsphere radius (Å)	Proportion excluded (%)
4	100	4	72
5	100	5	49
6	91	6	34
7	76	7	28
8	64	8	25
9	50	9	23
9.5	44	9.5	22
10	38	10	21
11	29	11	19
12	23	12	18
Inf.	0	Inf.	0

An alternative to the filtered Delaunay tessellation is provided by the alpha complex of Edelsbrunner (Liang et al., 1996). This construction begins by placing spheres of radius alpha centered at each particle. The spheres are inflated (starting at alpha = 0) until the first contact between spheres is obtained. The contacting sphere centers are joined by a line segment into a *1-simplex* and added to the growing list or *simplicial complex*. The process is continued and eventually three particles are joined into a triangle. If the circumcircle of this triangle is less than alpha, this *2-simplex* is added to the alpha complex. Tetrahedra with circumsphere radius less than alpha are added as *3-simplices*. Thus, at any stage, we have a list of points, 1-simplices, 2-simplices, and 3-simplices satisfying a uniform criterion for closeness, alpha. The list of 1-simplices with alpha equal to infinity is equivalent to the Delaunay tessellation. In the current work, we are using the equivalent of the 3-simplices of the alpha-complex for a specified alpha.

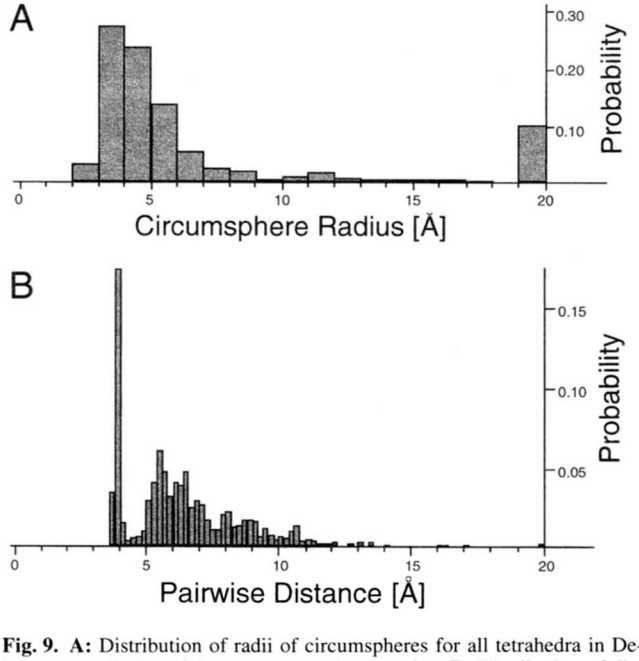

Fig. 9. A: Distribution of radii of circumspheres for all tetrahedra in Delaunay tessellation of three representative proteins. **B:** Distribution of distances between pairs of vertices in tetrahedra from A. Peak at 3.8 Å reflects vertices linked by the protein backbone.

Reduction of filtered tessellation to tetrahedra

As described above, we filter the tessellation in two stages; first on the basis of pairwise distances. Then, potential order-four contacts (tetrahedra) are checked for small circumsphere radius. The first cutoff is 9.5 Å for pairwise distance, and the second is 9.0 Å for circumsphere radius. Thus, the entire complexity of the 3D organization of all proteins in the dataset is reduced to a list of tetrahedra and the identities of each vertex.

The use of just the tetrahedral interactions is supported by a number of factors: First, virtually all relevant pairwise and three-way interactions appear as components of the filtered tetrahedra. Thus, our representation is a true generalization of commonly used pairwise potentials. Second, with reference to the theory of Markov Random Fields, which describe the process of randomly placing amino acid residues of various sizes at fixed backbone positions under potential energy constraints, it has been noted (Kindermann & Snell, 1980; Chellappa & Jain, 1993) that the limiting distribution of such a process may be adequately described by the distribution of residues at all *cliques* in the connection graph. In our case the *cliques*, or maximal connected subgraphs, are the selected tetrahedra. Thus, knowing the statistical distribution for the tetrahedra, we know the distribution for the entire system. Third, the extensive data reduction allowed by this tessellation argues for the need to investigate its utility as a means of recognizing correct versus incorrect sequence-structure matches. Finally, an additional reduction is achieved by assuming complete symmetry under permutation of the vertices of each tetrahedron. At this stage, it is not possible to test the adequacy of this assumption owing to the limited data set size. However, we argue that the symmetry is at least reasonable, a priori. At worst, we might be ignoring some chiral properties of 4 interacting residues. The beneficial effect of assuming symmetry is to reduce the $20^4 = 160{,}000$ categories of tetrahedra to a manageable 8,855, an approximate 24-fold reduction. With the current dataset, this implies an average category size of 44.

Coordination number

Coordination number (number of contacts or neighbors in a graph for a particular vertex) is known to be an important factor in determining the environment of residue side chains. High coordination number implies that the residue position is buried within the protein core, while low coordination number implies that a surface position is likely. In our scheme, coordination number is generalized to mean the number of tetrahedra in which a particular particle is included (rather than the number of simple connections to that particle). However, the effect of coordination number is largely unchanged: It remains a measure of the surface exposure of each residue environment.

Neither coordination number nor burial/exposed character is explicitly included in our tessellation description of protein organization. However, the effect of this factor is nonetheless felt by the resulting statistical descriptions of four-body residue interactions. Buried residues (high coordination number) contribute to the data set more often than do surface residues; thus, residues that prefer burial will have a higher propensity to appear in tetrahedra than surface-preferring, given their overall prevalence in the database.

Potential estimation and log-linear hierarchical model

Assuming that the presence of a particular combination of four residues occurs in the database in relationship to its potential energy, we can compute the appropriate energy terms from the frequencies.

The contribution to the potential energy of a single tetrahedron with vertex residue types $i \le j \le k \le l$ is estimated to be

$$E_{ijkl} = -\ln \frac{n_{ijkl}/K}{Np_ip_jp_kp_l}$$

where n_{ijkl} is the observed frequency of such tetrahedra summed over all K distinguishable permutations of the subscripts, p_i is the proportion of residue type i in the database, and N is the total number of tetrahedra in the database. The number of distinguishable permutations depends on the particular residue types, in particular on the equivalence relations which exist in a particular case, such as $i = j = k$. In general there are $K = k!/(q_1!q_2!\ldots q_d!)$ distinguishable permutations of k objects, where we have d equivalence classes of sizes $q_1,\ldots,q_d$ and $q_1 + q_2 + \ldots + q_d = k$. In our situation, $k = 4$, so $K(i,i,i,i) = 1$, $K(i,i,i,j) = 4$, $K(i,i,j,j) = 6$, $K(i,i,j,k) = 12$, $K(i,j,k,l) = 24$, for distinct vertices $i < j < k < l$. The numerator effectively averages the frequencies over permutations, ensuring satisfaction of the assumption that the potential function be fully symmetric. The denominator removes the average residue composition effect from the potential.

The four-body potential energy may also be estimated with the symmetrized observed frequencies n_{ijkl}/K replaced by the *predicted* frequency m_{ijkl}, where the prediction is based on a reduced model, one which postulates only low-order interactions. This technique allows us to keep the form of the potential (a quadruple sum over all tetrahedra in the tessellation) fixed, while we investigate the properties of the various interactions. Such predicted tetrahedral frequencies are elegantly provided by the log-linear model.

Log-linear model for four-way table

The potential energy function estimated above includes contributions from all interaction orders (pairwise to four-body); every tetrahedron includes six two-body and four three-body interactions, as well as four one-body terms. To separate these lower order interactions, we build a hierarchical log-linear model. Following Bishop et al. (1975), we construct a four-way table ($20 \times 20 \times 20 \times 20$) of the symmetrized, observed frequencies $n_{ijkl}/K(i,j,k,l)$. Using the iterative proportional fitting algorithm, we find maximum likelihood estimates for all the u terms in the model,

$$\begin{aligned}
\ln m^{(4)}_{ijkl} = u &+ u_i + u_j + u_k + u_l && \text{one-body effects}\\
&+ u_{ij} + u_{ik} + u_{il} + u_{jk} + u_{jl} + u_{kl} && \text{two-body interaction}\\
&+ u_{ijk} + u_{ijl} + u_{ikl} + u_{jkl} && \text{three-body interaction}\\
&+ u_{ijkl} && \text{four-body interaction}
\end{aligned}$$

for the predicted frequencies m_{ijkl}, subject to the symmetry constraint ($m_{ijkl} = m_{\sigma(ijkl)}$ for all permutations σ). Hierarchical submodels are also estimated:

$$\begin{aligned}
\ln m^{(3)}_{ijkl} &= u + u_i + u_j + u_k + u_l + u_{ij} + u_{ik} + u_{il} + u_{jk}\\
&\quad + u_{jl} + u_{kl} + u_{ijk} + u_{ijl} + u_{ikl} + u_{jkl},\\
\ln m^{(2)}_{ijkl} &= u + u_i + u_j + u_k + u_l + u_{ij} + u_{ik} + u_{il} + u_{jk}\\
&\quad + u_{jl} + u_{kl},\\
\ln m^{(1)}_{ijkl} &= u + u_i + u_j + u_k + u_l,\\
\ln m^{(0)}_{ijkl} &= u.
\end{aligned}$$

Comparison of observed frequencies to those based on hierarchical submodels allows for testing the significance of the omitted high-order terms. Terms of the log-linear model appear simply as components of the potential function:

$$\begin{aligned}
E_{ijkl} = \ln(Np_ip_jp_kp_l) - (u &+ u_i + u_j + u_k + u_l + u_{ij} + u_{ik}\\
&+ u_{il} + u_{jk} + u_{jl} + u_{kl} + u_{ijk} + u_{ijl}\\
&+ u_{ikl} + u_{jkl} + u_{ijkl}).
\end{aligned}$$

So, for a particular protein, the relevant potential value or score is a weighted sum of the relevant components:

$$\begin{aligned}
E = n\ln(N) + 4n\sum_{i=1}^{20}\ln(p_i) - nu - 4\sum_{i=1}^{20} n_iu_i - 6\sum_{i=1}^{20}\sum_{j=i}^{20} n_{ij}u_{ij}\\
- 4\sum_{i=1}^{20}\sum_{j=i}^{20}\sum_{k=j}^{20} n_{ijk}u_{ijk} - \sum_{i=1}^{20}\sum_{j=i}^{20}\sum_{k=j}^{20}\sum_{l=k}^{20} n_{ijkl}u_{ijkl}.
\end{aligned}$$

The u terms within the log-linear model represent components of the energy due to particular interactions of sets of particles; we should be able to interpret them in light of physical interactions. It is, however, more convenient to isolate terms of particular orders (pairwise, three-way, four-way) by subtracting the lower-order from the higher order potential values. We shall visualize these potential differences as a means of interpreting the underlying physical significance of the components.

Our potential function values and differences are presented in a four-way color-coded conditional plot or *co-plot* (Tukey & Tukey, 1983; Cleveland et al., 1992; Cleveland, 1993). Similar plots have also been termed *trellis* plots. Appropriate ordering of the 20 amino acid types enhances the appearance of patterns in the data. After investigating numerous orderings, we have chosen the following: VILAMFHWYPTGSQNDERKC. Here, the pure hydrophobic residues (VILAMF) are clustered, as are the positively (RK) and negatively (DE) charged ones. Cysteine (C) is located last so that its special properties can be gleaned easily. In the co-plot, a 20 by 20 array of window panes is presented, each pane labeled by the letters of the first two amino acid residues in the tetrahedron. Within each pane is a 20-pixel by 20-pixel image; each pixel representing the potential value for a particular combination of four residues. In this way, five dimensions of data can be represented in a two-dimensional plot, and all 160,000 terms of the potential may be represented. The imposed symmetry of the potential can easily be detected.

Owing to the small number of counts in some cells in the four-way table, we expect to see considerable statistical noise in the corresponding high-order terms of the potential. This is especially noticeable for the rarer residues (M, F, H, W). To aid in discerning the truly significant high order terms, we present the *normalized residuals*, or difference between observed and expected frequencies, adjusted for the expected variance of this difference. We use the Freeman-Tukey residual (Bishop et al., 1975, p. 136). $z = \sqrt{x} + \sqrt{x+1} - \sqrt{4m+1}$, where x is the observed frequency and m is the log-linear model predicted frequency. These residuals theoretically should have approximately a normal distribution with zero mean and variance equal to the number of residual degrees of freedom. They have slightly better properties for the low frequency case than do the more familiar components of chi-square residuals ($z = (x - m)/m^{1/2}$). When $x = 0$, we convert the Poisson proba-

bility for zero counts, given mean m, into the equivalent standard normal deviate. This method effectively highlights the most significant terms in the estimated potential differences.

Comparison of hierarchy

Four relevant log-linear models (first-, second-, third-, and fourth-order) may be compared by inspecting the degree to which the predicted frequencies (m_{ijkl}) agree with the observed frequencies. We apply the G^2 statistic for this purpose (Bishop et al., 1975), a form of the more familiar chi-square goodness-of-fit statistic, applicable to log-linear models. The G^2 statistic compares the difference between log-likelihood scores for model pairs, accounting for the complexity of each model (in terms of the number of parameters u required). Normal approximations to the chi-square statistic are satisfactory in this case since the chi-square degrees-of-freedom is high. Some caution must be used in interpreting the results of these tests as the occurrences of tetrahedra in the database are not statistically independent events as required by the G^2 test. Rather, there is a complex correlation within the observed frequencies since each residue contributes to several tetrahedra.

An intuitive appreciation of the relative merit of these models may be obtained by noting the equivalence of the log-likelihood values and theoretical information. The proportion of the total log-likelihood explained by each successive stage in the hierarchy is a good measure of the importance of that new step of complexity.

Sequence-structure recognition tests

A more incisive test of the performance of these potentials requires them to distinguish pairs of protein sequences threaded onto the same structure. The native sequence should produce a significantly lower energy value (score) than a sequence from another protein of equal length if the potential is performing well. This is the *structure-recognizes-sequence* test. Twelve pairs of same-length proteins from Holm and Sander (1992) were evaluated. All sequences in the calibration data set showing detectable homology (BLAST p-value < 0.1) to any of these 24 proteins were removed. Conversely, a useful potential will also correctly choose the correct structure for new sequences. Thus, a *sequence-recognizes-structure* test was also made, wherein the same sequence was threaded through two structures having the same length. The native structure for the sequence should again have significantly lower score. Adequacy of the potentials may be judged by how many pairs are correctly distinguished in either test, and the degree to which the *native* energy differs from the *incorrect* energy.

Dataset

The calibration dataset consisted of 608 proteins of known structure, having less than 35% pairwise sequence identity (Hobohm et al., 1992; Hobohm & Sander, 1994). These structures all had a resolution less than 3.0 Å, and an R-factor of less than 28%, and are listed in Table 5. An updated list can be obtained via anonymous ftp to site ftp.embl-heidelberg.de/pub/databases/protein_extras/pdb_select. The test data set consisting of 12 pairs is listed in Table 3. The following proteins were found to be significant homologues to one of these 24 and were removed from the 608 proteins to provide a new calibration dataset of 581 proteins to be used with the test set: 1cis, 1cse, 1ezm, 1fca, 1gdm, 1hip, 1huc, 1ilr, 1plc, 1pmy, 1poa, 1ppa, 1ppn, 1ppt, 1rhd, 2cbh, 2cdv, 2cro, 2cy3, 2tld, 2ts1, 3b5c, 6fab, 7ccp, 8i1b, 8rxn, 8tln.

Table 5. *List of 608 PDB identifiers*

125d, 1aaf, 1aaj, 1aak, 1ab2, 1abk, 1abrb, 1add, 1ads, 1aep, 1afp, 1agx, 1alka, 1amg, 1amp, 1aora, 1aoza, 1aplc, 1apme, 1arb, 1ars, 1ash, 1asza, 1atna, 1aty, 1baa, 1babb, 1bam, 1bara, 1bbl, 1bbpa, 1bbt1, 1bbt2, 1bcfa, 1bet, 1bgeb, 1bgu, 1bip, 1bmda, 1bpb, 1bsaa, 1bvp1, 1bw4, 1c5a, 1caua, 1caub, 1cbn, 1cbp, 1ccf, 1ccr, 1cdta, 1cela, 1cewi, 1cfb, 1cgt, 1chl, 1chma, 1chra, 1cid, 1cksa, 1clc, 1cmca, 1cola, 1cpca, 1cpcb, 1cpt, 1crl, 1csei, 1cska, 1csp, 1ctaa, 1ctl, 1ctm, 1ctn, 1ctt, 1cus, 1ddt, 1dec, 1dfna, 1dhr, 1dlc, 1dmc, 1dsba, 1dts, 1dyna, 1eca, 1ede, 1eft, 1eng, 1enh, 1epab, 1epta, 1erd, 1erh, 1eria, 1erl, 1erp, 1etb1, 1fbaa, 1fca, 1fcda, 1fcdc, 1fct, 1fha, 1fiab, 1fkf, 1fnc, 1fod4, 1frpa, 1frub, 1fxra, 1gal, 1gara, 1gata, 1gbs, 1gca, 1gdha, 1gdm, 1ghsa, 1gky, 1glcg, 1glt, 1gmfa, 1gof, 1gox, 1gp1a, 1gpb, 1gph1, 1gpr, 1gps, 1grj, 1gsq, 1gtra, 1har, 1hbq, 1hce, 1hcgb, 1hcna, 1hcnb, 1hcra, 1hdca, 1hdgo, 1hex, 1hfc, 1hfi, 1hft, 1hgja, 1hjra, 1hks, 1hlb, 1hlea, 1hleb, 1hmcb, 1hmpa, 1hmy, 1hnf, 1hns, 1hoe, 1hph, 1hrf, 1hrnb, 1hrti, 1hsla, 1htmd, 1htp, 1htrp, 1hucb, 1huma, 1huw, 1hvd, 1iae, 1iag, 1ica, 1ifa, 1ifc, 1igp, 1ilk, 1ilr1, 1irk, 1isca, 1isua, 1ivd, 1kab, 1kana, 1knb, 1knt, 1ktx, 1l92, 1lba, 1lena, 1lgaa, 1lis, 1lki, 1llda, 1lobd, 1lpbb, 1lpe, 1ltsa, 1ltsc, 1ltsd, 1lyp, 1mat, 1mdc, 1mina, 1minb, 1mmob, 1mmog, 1mrb, 1mrj, 1msc, 1msec, 1mup, 1mylb, 1mypc, 1nar, 1nbaa, 1nfp, 1nhkl, 1nipa, 1nnt, 1nrd, 1olba, 1oma, 1omp, 1oxy, 1oyb, 1paa, 1pbe, 1pbp, 1pcl, 1pcrh, 1pcrm, 1pdc, 1pdga, 1peta, 1pfia, 1pga, 1pho, 1php, 1phy, 1pii, 1pkn, 1plq, 1pls, 1pmy, 1pnt, 1poa, 1poc, 1poxa, 1ppbl, 1ppi, 1ppn, 1ppt, 1prcc, 1prs, 1prta, 1prtc, 1prtd, 1prtf, 1psm, 1pspa, 1pte, 1ptx, 1pxtb, 1pyab, 1pyda, 1pyp, 1qora, 1rblm, 1rcb, 1rec, 1ret, 1rgd, 1riba, 1ropa, 1rpa, 1rsy, 1rtm1, 1rtp1, 1rvaa, 1s01, 1saca, 1scma, 1scmc, 1scua, 1scub, 1scy, 1slta, 1spf, 1srga, 1srya, 1sto, 1tabi, 1tadc, 1tap, 1tca, 1tfi, 1tgsi, 1thta, 1thv, 1tib, 1tie, 1tlca, 1tlk, 1tml, 1tnra, 1tnrr, 1tph1, 1tpla, 1tpn, 1tpt, 1trb, 1trka, 1trt, 1trzb, 1tssa, 1tvt, 1ula, 1urk, 1vaaa, 1vil, 1was, 1wfba, 1whta, 1whtb, 1wsya, 1wsyb, 1xnb, 1xsoa, 1xys, 1yptb, 1ytba, 1zaac, 2acg, 2achb, 2ak3b, 2alp, 2atcb, 2ayh, 2azaa, 2bbkh, 2bbvc, 2bds, 2bopa, 2bpa1, 2bpa2, 2bpa3, 2btfp, 2cas, 2cba, 2cbh, 2cdv, 2chsa, 2cnd, 2cp4, 2cpl, 2crd, 2cro, 2ctc, 2dkb, 2dnja, 2dri, 2drpa, 2ebn, 2ech, 2end, 2er7e, 2fcr, 2gsta, 2hbg, 2hhma, 2hipa, 2hmza, 2hnq, 2hnte, 2hpda, 2hsp, 2ihl, 2ila, 2kaib, 2kaub, 2kauc, 2lgsa, 2liv, 2madl, 2mev1, 2mge, 2mhu, 2mnr, 2mtac, 2ohxa, 2pac, 2pcda, 2pde, 2pf1, 2pfkd, 2pgd, 2pia, 2pmga, 2por, 2reb, 2rn2, 2rslb, 2rspb, 2sas, 2scpa, 2sh1, 2sil, 2sn3, 2snv, 2stv, 2tbva, 2tgi, 2tmda, 2tmvp, 2tpra, 2ts1, 2ztaa, 3aaha, 3aahb, 3cd4, 3chy, 3cla, 3cox, 3dfr, 3ebx, 3egf, 3gapb, 3gly, 3hhrc, 3hsc, 3il8, 3mdda, 3mona, 3sdha, 3sgbi, 4blma, 4cpai, 4enl, 4fxn, 4gcr, 4rhv1, 4rhv3, 4rhv4, 4sbva, 4sgbi, 4tgf, 4xiaa, 4znf, 5p21, 5ruba, 5znf, 6fabl, 6taa, 7apib, 7ccp, 7pti, 7rsa, 8abp, 8acn, 8atca, 8cata, 8fabb, 8rxna, 8tlne, 9rnt, 9wgaa, 1aba, 1ack, 1acp, 1adr, 1amy, 1ang, 1apa, 1aps, 1avda, 1ayaa, 1bbre, 1bct, 1bgh, 1bhb, 1bmta, 1bn21, 1bnh, 1bova, 1brsd, 1byb, 1c53, 1cbs, 1cd8, 1chc, 1cis, 1clh, 1croa, 1cxa, 1dlha, 1dlhb, 1dpi, 1drf, 1eaf, 1ego, 1esl, 1exg, 1ezm, 1fas, 1flp, 1fna, 1frd, 1gmpa, 1gsra, 1hdp, 1hip, 1hma, 1hmx, 1hrha, 1hsta, 1inp, 1itha, 1lac, 1leb, 1lfb, 1lid, 1lmb3, 1lpba, 1lpt, 1mdka, 1mdyb, 1mmod, 1mpp, 1mypa, 1ncia, 1nrca, 1nsca, 1ntr, 1onc, 1pba, 1pcrl, 1pda, 1pgs, 1pkp, 1pkt, 1plc, 1poh, 1pou, 1ppa, 1pse, 1put, 1pvua, 1pyaa, 1rfba, 1rhd, 1rip, 1ris, 1sbp, 1sema, 1shfa, 1shg, 1spha, 1srp, 1sso, 1stfi, 1svr, 1sxl, 1taha, 1ten, 1tin, 1tnn, 1tnt, 1ubi, 1udpa, 1ukz, 1utg, 1vmoa, 1vsga, 1wapb, 256ba, 2at2a, 2blta, 2ccya, 2cmd, 2cy3, 2fx2, 2kaua, 2lhb, 2mcm, 2mev2, 2mev3, 2nada, 2pcdm, 2plea, 2pola, 2ptl, 2sblb, 2sga, 2spca, 2tldi, 2trxa, 2wrpr, 3b5c, 3dpa, 3lada, 3rubl, 3tgl, 4dfra, 4icb, 4rhv2, 7icd, 8i1b

After computing the tessellations and filtering, the 608 proteins provided $N = 386{,}425$ tetrahedra, which implies an average effective cell size of 44.6 in the symmetrized four-way table. Using the purged calibration dataset of 581 proteins, $N = 374{,}791$ tetrahedra were obtained.

Acknowledgments

We would like to acknowledge Dr. V. DiFrancesco for a critical reading of our manuscript, and Dr. D. Carr who suggested appropriate references for the four-dimensional co-plots. Vernon Chi provided a helpful reading of the draft manuscript.

References

Barber CB, Dobkin DP, Huhdanpaa HT. 1995. The Quickhull algorithm for convex hulls. *ACM: Trans on Mathematical Software 22(4)*:469–483.

Bishop YMM, Fienberg SE, Holland PW. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, Massachusetts: The MIT Press.

Bryant SH, Lawrence CE. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins 16*:92–112.

Chellappa R, Jain A, eds. 1993. *Markov random fields theory and application*. San Diego: Academic Press.

Cleveland W, Grosse E, Shyu W. 1992. Local regression models. In: Chambers J, Hastie T, eds. *Statistical models in S*. New York: Chapman and Hall. pp 309–376.

Cleveland WS. 1993. *Visualizing data*. Summit, New Jersey: Hobart Press.

Godzik A, Kolinski A, Skolnick J. 1992. Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol 227*:227–283.

Godzik A, Kolinski A, Skolnick J. 1993. De novo and inverse folding predictions of protein structure and dynamics. *J Comput Aided Mol Des 7*:397–438.

Godzik A, Skolnick J. 1992. Sequence–structure matching in globular proteins: Application to supersecondary and tertiary structure determination. *Proc Natl Acad Sci USA 89*:12098–12102.

Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci 3*:522–524.

Holm L, Sander C. 1992. Evaluation of protein models by atomic solvation preference. *J Mol Biol 225*:93–105.

Kindermann R, Snell JL. 1980. *Markov random fields and their applications*. Providence, Rhode Island: American Mathematical Society.

Lee B. 1993. Estimation of the maximum change in stability of globular proteins upon mutation of a hydrophobic residue to another of smaller size. *Protein Sci 2*:733–738.

Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. 1996. Analytical shape computation of macromolecules. I. Molecular area and volume through alpha shape. Urbana-Champaign, Illinois: Department of Computer Science, Beckman Institute, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign.

Lim W, Sauer R. 1991. The role of internal packing interactions in determining the structure and stability of a protein. *J Mol Biol 219*:359–376.

Miyazawa S, Jernigan RL. 1983. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules 18*:534–552.

Ptitsyn O. 1995. Molten globule and protein folding. *Adv Protein Chem 47*:83–229.

Singh RK, Tropsha A, Vaisman I. 1996. Delaunay tessellation of proteins: Four-body nearest-neighbor propensities of amino-acid residues. *J Comp Bio 3*:213–221.

Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol 5*:229–235.

Srinivasan R, Rose G. 1995. LINUS: A hierarchic procedure to predict the fold of a protein. *Proteins 22*:81–99.

Tukey JW, Tukey PA. 1983. *Some graphics for studying four-dimensional data*. Computer science and statistics: Proceedings of the 14th Symposium on the Interface. New York: Springer Verlag.

Wodak SJ, Rooman MJ. 1993. Generating and testing protein folds. *Curr Opin Struct Biol 3*:247–259.

Zheng W, Cho S, Vaisman I, Tropsha A. 1997. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. In: Altman R, Dunker K, Hunter L, Klein T, eds. *Pacific Symposium on Biocomputing '97*. Maui, Hawaii: World Scientific Publishing Co. Pte. Ltd. pp 486–497.