

Multi-body Interactions within the Graph of Protein Structure

Peter J. Munson¹ and Raj K. Singh²

¹Analytical Biostatistics Section, LSB, DCRT
National Institutes of Health
Bldg 12A, Room 2041, Bethesda, MD 20892-5626
munson@helix.nih.gov

and

²Dept. Computer Science, U. North Carolina, Chapel Hill, NC
singh@cs.unc.edu

Abstract

We construct a graphical representation of protein structure based on the 3D C-alpha carbon point set, using the Delaunay tessellation to define interacting quadruples of amino acid residues. The tessellation is filtered by two criteria: interaction distance less than 9.5 angstroms and circumsphere radius less than 8.0 angstroms using dataset of 608 protein structures of low mutual sequence identity and a likelihood ratio test, we show that 3-body and 4-body interactions are indeed significant. We identify particular significant three-body interactions by first reducing the dataset to interacting triples, and classifying amino acid residues in a reduced alphabet. Although cystein was previously shown to be a dominant source of 3-body interactions, we now identify additional significant 3-body interactions of charged, hydrophobic and small residues.

Introduction

The internal organization of peptide residues within a protein is a reflection of both the primary amino acid sequence and the folding process determining its three-dimensional structure. Understanding this organization is key to recognizing the correct "fold" for a given sequence, using threading techniques. Current threading techniques generally use empirical potentials with at most pairwise interactions to describe this organization. The performance of such techniques appears to be bounded by the lack of specificity of the potential. We sought to extend empirical potentials with 3-body and 4-body interactions (Singh, Tropsha et al. 1996; Munson and Singh 1997). Here we refine our definition of relevant multi-body interactions and categorize the significant three-body interactions according to the physical properties of the participating amino acids.

The geometric organization of the peptide residues within a protein structure may be represented by the Delaunay tessellation. This tessellation divides the interior volume of the protein into a set of nonoverlapping tetrahedra with vertices at the C^α carbons, thereby uniquely defining interacting quadruples of residues. Several advantages have been noted for this representation: nearest-neighbors are defined without reference to a

distance cutoff; each portion of interior volume is bounded by exactly four vertices thus facilitating statistical calculations of multi-body interactions and the necessary computational geometry algorithms are generally available. We and others have shown this representation to be useful for fold recognition using threading approaches (Munson and Singh 1997; Zheng, Cho et al. 1997).

Previously, we investigated the statistical significance of three- and four-body terms using three methods: likelihood-based log-linear statistical models, potential energy differences in fold-recognition and sequence-recognition tests, and graphical display of the complete four-body potential. The most dramatic feature of the four-body potential arises from interactions of cystein residues in pairs, triples and quadruples. This is not surprising, as the covalent disulfide bridges found between cystein pairs imply a strongly "attractive" pairwise term which must be compensated at the three- and four-body level.

A 4-body potential is actually made up of many terms including 4 one-body, 6 two-body, 4 three-body and only 1 pure four-body term. While the pure four-body interactions as a group were clearly significant, only interactions involving four cysteins (CCCC) could be judged statistically significant by itself. Given the limitations of the dataset, we now explore in detail the individual pure three-body terms. We find that even in the non-cystein bearing triples there are significant higher-order interactions. The strongly interacting triples of residues fall into patterns involving interactions of charged and hydrophobic residues.

Methods

Delaunay tessellation is a technique to establish the spatial neighbors of a 3-D point set. It is the mathematical dual of the Voronoi diagram for that set. Basically, the tessellation divides the convex hull of the point set into a (nearly always) unique set of non-overlapping tetrahedra whose vertices are the original points. A feature of Delaunay tessellation is that the circumsphere of each tetrahedron (sphere with the four vertices on its surface) does not

contain any other points in the set. The edges of the resulting tetrahedra connect pairs of vertices which are "nearest-neighbors". Groups of four vertices in a tetrahedron are considered a *clique* in the connection graph for the point set and are available for analysis of the four-body and lower-order interactions. The tessellation can be efficiently calculated with available software (Barber, Dobkin et al. 1995).

In applying this technique to protein structures, we represent the protein as simply a 3-D point set with one point placed at the center of each C α carbon. More complex representations (points at the C β carbon, or other representative atoms, etc.) have been tried, but do not make a substantial difference for the current purposes. Clearly, a more detailed representation of the protein can be useful for protein threading-fold recognition. To make the tessellation more representative of actual interactions within proteins, we employed a filtering technique described elsewhere (Munson and Singh 1997). Basically tetrahedra with any edges are greater than 9.5Å or with circumsphere radius greater than 8.0Å are rejected. This eliminates about half of the original tetrahedra, which generally lie outside the water-accessible surface of the protein.

For this study we tessellated a dataset of 608 protein chains of known structure having less than 35% pairwise sequence identity and a resolution of less than 3.0 Å (Hobohm and Sander 1994).

The frequencies of observed tetrahedra, labeled by the standard 20-letter amino acid residue code, are arranged into a 20x20x20x20 table and form the basis for our analyses. We use a log-linear statistical model to represent the natural logarithm of the frequencies as a sum of zero-, first-, second, third and fourth order terms. The full 4-body model is:

$$\begin{aligned} \ln m_{ijkl}^4 &= u_i + u_j + u_k + u_l && \text{1-body effects} \\ &+ u_{ij} + u_{ik} + u_{il} + u_{jk} + u_{jl} + u_{kl} && \text{2-body interaction} \\ &+ u_{ijk} + u_{ijl} + u_{ikl} + u_{jkl} && \text{3-body interaction} \\ &+ u_{ijkl} && \text{4-body interaction} \end{aligned}$$

where the predicted frequencies m_{ijkl} , are subject to the symmetry constraint ($m_{ijkl} = m_{\sigma(ijkl)}$ for all 24 permutations σ). This model and all hierarchical submodels are estimated using the maximum likelihood iterative proportional fitting algorithm (Bishop, Fienberg et al. 1975) programmed in MATLAB (The Mathworks, Inc., Natick, Mass., USA). Log likelihood differences (Δ 's) between models are compared to the difference in number of parameters for two models, and the result is given as a Z-score, $Z = (2\Delta \log \text{likelihood} - \Delta df) / \sqrt{2\Delta df}$, where df is the degrees of freedom or number of

parameters associated with each model. This statistic is asymptotically distributed as a standard normal variate and can be used to judge the significance of a high-order model compared to a lower-order model. Absolute values greater than 3.3 correspond to P-values of about 0.001 or less.

The potential energy associated with any particular assignment of residues to the four tetrahedral vertices is estimated from the modeled frequencies as $E_{ijkl} = -\ln(m_{ijkl}^4 / Np_i p_j p_k p_l)$ where the p_i are the proportion of residue type i in the database. Full graphical displays of the 20x20x20x20=160,000 terms in the potential have been presented elsewhere (Munson and Singh 1997). Components of the potential energy can be obtained by referring to the appropriate u term of the log-linear model.

Because many of the observed or expected frequencies are quite low (average observed frequency is about 50 per cell), some of the energy components are unreliably estimated. Thus, judging the true significance of individual components may be problematic. To account for this sparse data, we have also calculated the Freeman-Tukey residual ((Bishop, Fienberg et al. 1975), p 136) which combines observed, x , and predicted, m , frequencies into a statistic of approximately unit variance regardless of the small frequencies. It is $z = \sqrt{x} + \sqrt{x+1} - \sqrt{4m+1}$.

Invoking the permutability assumption, we sum over all distinct permutations of the subscripts for the observed and expected frequencies, yielding a smaller number (8,855 rather than 160,000) of categories and larger average frequencies.

To find patterns in the three-body terms, a 3-way table of the frequencies of triangles in the original tessellation is required. A three-way table is obtained by summing one of the factors of the four-way table. Thus, the expected number m_{ijk}^3 of triangles with vertices i,j,k is given by

$$m_{ijk}^3 \approx \frac{4}{2} * \sum_{l=1}^{20} m_{ijkl}^4,$$

where the factor 4 arises since there are four triangular faces on each tetrahedron, and the factor 2 since each triangle is present in approximately 2 tetrahedra; those on the surface are present in only one. Invoking permutability, we can again condense 8,000 categories to 1330. The observed and expected frequencies for triangles can be studied with the methods given above. Still further reduction is obtained by recoding the standard 20 letter amino acid residue code as follows: *hydrophobic* "h"={A,F,I,L,M,V,W,Y}, *positively charged* "+"={K,R}, *negatively charged* "-"={D,E,N,Q}, *small* "s"={P,S,T}. Other residues C, G, H were not recoded.

Results & Discussion

We had previously established the presence of higher-order (greater than 2-body) interactions in the frequency of tetrahedra in tessellated proteins (Munson and Singh 1997). One line of evidence is repeated in Table 1 for a refined dataset (more stringent filtering of the original tessellation, see *Methods*). The hierarchical comparison of the three-body model to the two-body model yields a Z-score of 54, which is extremely significant. The comparison of the 4- to the 3-body model yields a Z-score which is significant (more so than previously reported), but still uncomfortably close to the usual cutoff ($|Z| > 3.1$). Nevertheless, with 32% of the total log-likelihood or information from all the multibody terms (2,3, and 4), the 4-body term seems to make an important contribution.

Table 1. Hierarchical Comparison of Models.

Comparison	$\Delta \log\text{Likelihood}$	Δdf	Z-score
1 vs 0	-18676	19	6056
2 vs 1	-12904	190	1314
3 vs 2	-2059	1330	54
4 vs 3	-3947	7315	4.8

We previously sought important pure 4-body interactions responsible for the significance of the 4-body vs 3-body comparison. (Munson and Singh 1997). Only the 4-body term associated with the quadruple CCCC was clearly significant. Multibody interactions involving cysteine arise from the ability of C to form covalently linked pairs. Such pairs imply a very strong pairwise term, which contribute to all six pairs of edges in a tetrahedra and thereby over-predict the occurrences of CCC and CCCC.

To look beyond this group of effects, we removed all tetrahedra with C at any of their vertices. Of the original 385,161 tetrahedra, 355,500 remained. Repeating the analysis of significance of the high-order models shows (Table 2) that both the three- and four-body components remain significant (Z-score = 38 and 4, resp.), although the Z-score values are noticeably reduced. Thus, even after C is removed from the analysis, there appears to be high-order interactions.

Table 2. Hierarchical Comparison of Models after Removing Cystein

Comparison	$\Delta \log\text{Likelihood}$	Δdf	Z-score
1 vs 0	-17053	18	5681
2 vs 1	-7077	171	756
3 vs 2	-1476	1140	38
4 vs 3	-3190	5985	4

To investigate for significant three-body interactions, we reduced our dataset from a tetrahedral representation to one embodying only triangles of residues. Accordingly, we computed a three-way table of the frequencies of all triples (triangles in the original tessellation) of residues (see *Methods*). From the original filtered tessellation, we found about 770,300 triangles. The three-way table showed a strongly significant three-body component ($Z=18.6$). When cysteine-containing triangles were dropped (yielding 711,000 triangles), the three-body Z-score dropped to 8.5, but was still highly significant.

After dropping the C residues, the set of Freeman-Tukey residuals (versus two-body predictions) were ranked, yielding a list of 1,140 values (all distinct combinations of 19 letters). The most over-represented triples were DRV, EKL, EKV, AER, DFK, all examples of an oppositely charged pair with a hydrophobic residue. At the bottom of the list we find a more heterogeneous group: GMR, DER, ELV, NQR, EEK. To find common patterns in this list, we re-expressed the 20 letter amino acid code into a reduced alphabet (see *Methods*). We then looked for consistent patterns of residue types in the top 40 and the bottom 40 Freeman-Tukey residuals.

Six patterns emerged from this analysis (Table 3). The first pattern (+ - h) involves oppositely charged pairs with hydrophobic residues. This pattern produced the largest positive value of any residual (rank 1), and had a total of 13 members in the top 40. Taking all 64 members of the group together, the average tendency was significantly positive (Student's *t* statistic value = 5.8), indicating that as a group the (+ - h) pattern was significant. Three-body interactions between oppositely charged residues and a hydrophobic have been observed (Godzik, Kolinski et al. 1992; Godzik and Skolnick 1992; Munson and Singh 1997).

Interestingly, the pattern (- - h) also appeared with a significant positive three-body interaction. However, the pattern (+ + h) did not appear in any of the top 40 residuals. The pattern (+/- h s) appeared 5 times in the top 40 and had a significant *t*-value overall.

For the underrepresented patterns, the most striking pattern involves three charged residues (+ - -) or (+ + -). Almost as important are the two groups (- h h) and (+ h h), involving a single charge with two hydrophobics. It should be noted that the hierarchical statistical model creates the situation where these various patterns may be linked: a large positive interaction in the model for one cell, say, for a (+ - h) pattern, implies that other interactions must be negative, in order that the sum of three-body interactions be zero. Thus, the patterns involving hydrophobic and charged residues are obviously interrelated.

What is clearly new in this analysis is the following. Significant three-body interactions remain after removing the masking effect of cysteine interactions. Not only do the

previously identified oppositely charge pair-hydrophobic interactions emerge as significant, but a strong tendency is identified for similarly-charged pair and hydrophobic triples to cluster. These interactions are balanced by significant *anti*-clustering for charged residue triples, or single charged residues with two hydrophobic residues.

Within each of the general patterns, there is clearly heterogeneity as some (+ - h) members (DRV for example) produce a much more positive residual than others. Some members actually underrepresented (for example ERF). Precisely how these residue combinations interact in protein structures remains to be explained fully.

Finally, we offer an explanation for the interaction terms involving charged and hydrophobic residues. The paradoxical clustering of a charged pair with a hydrophobic residue must be interpreted in the context of the pairwise components acting on this same set. There are three pairwise components (one negative, two positive in this instance) and three one-body components to consider as well. Thus, the pure three body term is really a correction added to the otherwise largely repulsive pairwise terms between the hydrophobic and the charged residues.

Table 3. Prominent 3-body Interactions not Involving Cystein

Pattern* and Members in Top or Bottom 40	Top rank	Group size	Group t-value
<u>Over-represented (positive interaction)</u>			
(+ - h) DRV,EKL,EKV,AER,DFK,AQR, EIK,ERV,DLR,EIR,DIK,ELR,LNR	1	64	5.8
(+/- h s) ADP, MRT, QSV, KLP, EFT	8	96	2.8
(- - h) DEI,NNV,NQV,DEV,EIN,DDM	9	80	4.2
<u>Under-represented (negative interaction)</u>			
(+ . .) or (+ + -) DER,NQR,EEK,DDK,KNN,DEK	2	20	-4.4
(- h h) ELV, LNV, IQV, EIV, AEV	3	144	-4.8
(+ h h) ILR,FIK,ARV,KLV,KLM,IKL,MRW	11	72	-2.4

*See *Methods* for reduced alphabet in patterns.

Three-body terms may arise partly as a consequence of the permutability assumption. A logical consequence of that assumption is that the model does not distinguish any individual vertex position in the graph; all positions are assumed to have the identical distribution on the set of 20 residues. In reality, the presence of a surface boundary

implies that some vertices are special; they lie on the solvent accessible surface and are much more likely to be occupied by charged residues, while other vertices are more likely to be occupied by hydrophobic residues. Depending on the geometric arrangement of residues near the surface, pairs of charged residues, likely to lie on the surface, may also be likely to interact with a buried hydrophobic neighbor. Multi-body interactions also arise when special relationships (covalent bonding of C-C pairs but not of triples or quadruples), are present. Advancing the statistical description of the protein structure clearly requires inclusion of this detailed information.

For now, it is clear that many threading methods which look primarily at two-body interactions are missing about 50% of the available information in these higher-order interactions. Here, we have confirmed the existence of high-order interactions on a large dataset, and identified two major sources of these interactions: cystein interactions and hydrophobic-charge combinations. The power of protein threading methods which use pairwise pseudopotentials would surely be enhanced if such multibody interactions were incorporated.

References

- Barber, C. B.; Dobkin, D. P. and Huhdanpaa, H. T. 1995. The Quickhull Algorithm for Convex Hulls. *ACM: Trans. on Mathematical Software*, Forthcoming.
- Bishop, Y. M. M.; Fienberg, S. E. and Holland, P. W. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: The MIT Press.
- Godzik, A.; Kolinski, A. and Skolnick, J. 1992. Topology Fingerprint Approach to the Inverse Protein Folding Problem. *J Mol Biol* 227: 227-283.
- Godzik, A. and 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. and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Science* 3: 522-524.
- Munson, P. and Singh, R. 1997. Statistical Significance of Hierarchical Multi-body Potentials Based on Delaunay Tessellation and their Application in Sequence-Structure Alignment. *Protein Science* Forthcoming.
- Singh, R. K.; Tropsha, A. and Vaisman, I. I. 1996. Delaunay Tessellation of Proteins: Four-body Nearest-neighbor Propensities of Amino-acid Residues. *J Comp Bio* 3: 213-221.
- Zheng, W. et al. 1997. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. In Pacific Symposium on Biocomputing '97, Maui, Hawaii, USA, World Scientific Publishing Co. Pte. Ltd.