

Compositional Preferences in Quadruplets of Nearest Neighbor Residues in Protein Structures: Statistical Geometry Analysis

Iosif I. Vaisman, Alexander Tropsha, and Weifan Zheng
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599
E-mail: iosif_vaisman@unc.edu

Abstract

Three-dimensional structure and amino acid sequence of proteins are related by an unknown set of rules that is often referred to as the folding code. This code is believed to be significantly influenced by nonlocal interactions between the residues. A quantitative description of nonlocal contacts requires the identification of neighboring residues. We applied statistical geometry approach to analyze the patterns of spatial proximity of residues in known protein structures. Structures from a dataset of well resolved nonhomologous proteins with a single point representation of residues by C_α atoms were tessellated using Delaunay algorithm. The Delaunay tessellation generates an aggregate of space-filling irregular tetrahedra, or Delaunay simplices. The vertices of each simplex objectively define four nearest neighbor C_α atoms and therefore four nearest neighbor residues. Compositional analysis of Delaunay simplices reveals highly nonrandom clustering of amino acid residues in protein structures. Relative abundance or deficiency of residue quadruplets with certain compositions reflects propensities of different types of amino acids to be associated or disassociated in folded proteins. The likelihood of occurrence of four residues in one simplex displays strong nonrandom signal also in the case of a reduced amino acid alphabet. We used several different three-letter alphabets based on the residue chemical and structural properties and on

the complementarity of the corresponding codons. In all cases the clustering of residues correlates with their properties or genetic origin. The results of this analysis are being implemented in algorithms for protein structure classification and prediction.

Introduction

Revolutionary developments in genomics and computational structural biology lead to the rapidly increasing amount of data on biomolecular sequences and structures. The deposition rate for both sequence and structure databases continues to grow exponentially. The efficient utilization of this data depends on the availability of the methods and tools for protein structure analysis and prediction. Accurate prediction of protein three-dimensional structure from its primary sequence represents one of the greatest challenges of modern theoretical biology. Detailed knowledge of protein structure is essential for understanding the mechanisms of biological processes at molecular, cellular, and evolutionary levels. The structures of only a fraction of all known primary sequences have been determined experimentally. Several approaches to protein structure prediction have been developed in recent years. Many of these approaches rely on the knowledge derived from the analysis of spatial and compositional patterns in known protein structures. Such an analysis require an objective definition of nearest

neighbor residues, that can be provided by the statistical geometry methods. In the statistical geometry methods the nearest neighbors and identified by statistical analysis of irregular polyhedra obtained as a result of a specific tessellation in three-dimensional space. Voronoi tessellation which partitions the space into convex polytopes called Voronoi polyhedra was used for the analysis of geometry and packing density of a number of molecular systems including protein crystals [1-5]. A group of four atoms, whose Voronoi polyhedra meet at one vertex, forms another basic topological object, the Delaunay simplex. The Delaunay tessellation was used for structural analysis of various disordered systems and showed a good potential as a structure description method [6,7]. Unlike many of the existing definitions of nearest neighbors that are often based on some selected distance criteria (e.g. separation of C_α atoms or other pairs of atoms by a certain distance), Delaunay tessellation provides the robust definition of nearest neighbors, independent of any arbitrary criteria. Statistical analysis of the compositional propensities in adjacent residues can be used for protein structure classification and prediction.

Delaunay tessellation of protein structures

By studying patterns of spatial arrangement of individual amino acids in known protein structures, valuable information about specific interactions between particular residues types can be obtained. It is likely that multibody interactions make significant contribution to the potential energy landscape of folded proteins, and therefore it is important to identify and correctly estimate these contributions. To define groups of nearest neighbors, we propose to use Delaunay tessellation [8,9] in the analysis of protein structures, where each amino-acid residue is represented by a single point. The geometry of Delaunay simplices and Voronoi polytopes is illustrated in Fig. 1. Two-dimensional Delaunay simplices shown in Fig. 1 are triangles, Delaunay simplices in three dimensions are tetrahedra. Since the Delaunay simplex in three-dimensional space always has four vertices, the Delaunay tessellation of a protein

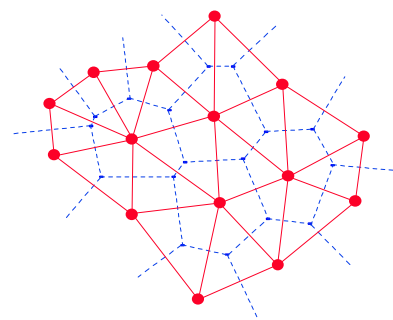


Fig. 1. Voronoi/Delaunay tessellation in 2D space

(Voronoi polyhedra - dashed line, Delaunay simplices - solid line).

structure generates an aggregate of space-filling, non-overlapping, irregular tetrahedra.

For the analysis of correlations between the structure and sequence of proteins, we introduced a classification of simplices based on the relative positions of vertex residues in the primary sequence. The following five nonoverlapping classes are considered (Fig. 2): class {4}, where all four residues of the simplex are consecutive in the protein sequence; class {3,1}, where three residues are consecutive, and the fourth is distant in the sequence; class {2,2}, where two pairs of consecutive residues are distant in sequence; class {2,1,1}, where two residues are consecutive, and two other residues are distant from the first two and from each other; and class {1,1,1,1} where all four residues are distant from each other. The two residues were defined as distant if they were separated by one or more residues in protein

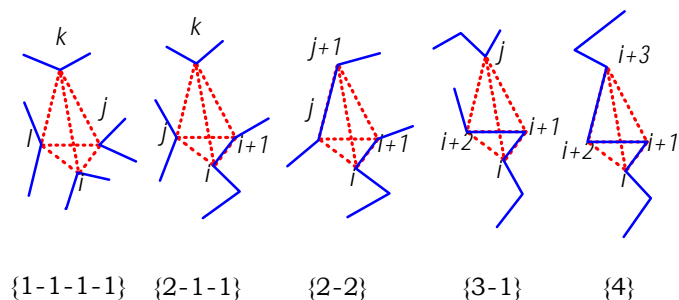


Fig. 2. Five classes of Delaunay simplices.

sequence. The representatives of each class are usually found in every protein structure. Simplices belonging to each of the five classes have characteristic geometrical properties such as edge length distribution, volume, and tetrahedrality. Tetrahedrality is a quantitative measure of the degree of distortion of the Delaunay simplices from the ideal tetrahedron [10].

$$T = \sum_{i>j} (l_i - l_j)^2 / 15 \bar{l}^2 \quad (1)$$

where l_i is the length of the i -th edge, and \bar{l} is the mean length of the edges of a given simplex. Distribution of tetrahedrality and volumes for all five classes of tetrahedra is shown in Fig. 3. The figure shows significant differences between tetrahedra in different classes. Simplices of class {4} (solid line) and class {2,2} (dots) have pronounced, sharp peaks in both tetrahedrality and volume distributions.

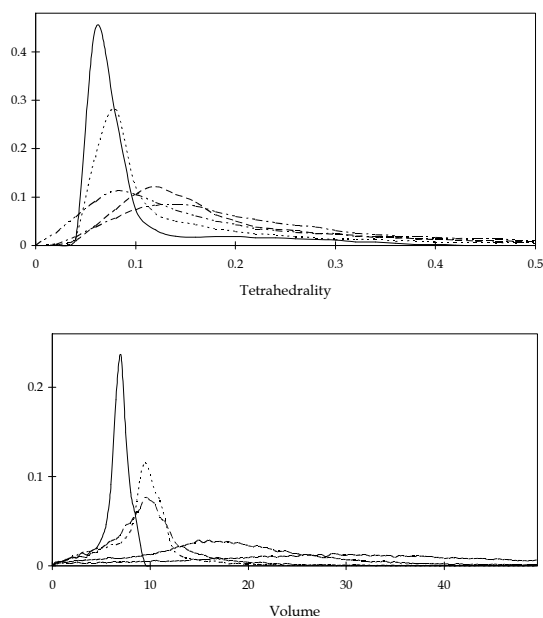


Figure 3. Distribution of tetrahedrality and volume (in Å³) of Delaunay simplices

Compositional analysis of Delaunay simplices

Statistical analysis of amino acid composition of the Delaunay simplices provides information about spatial propensities of all quadruplets of amino acid residues to be clustered together in folded protein structures. We shall analyze the results of Delaunay tessellation of some known proteins in terms of statistical likelihood of occurrence of four nearest neighbor amino acid residues for all observed quadruplet combinations of 20 natural amino acids. The log-likelihood factor, q , for each quadruplet is defined as

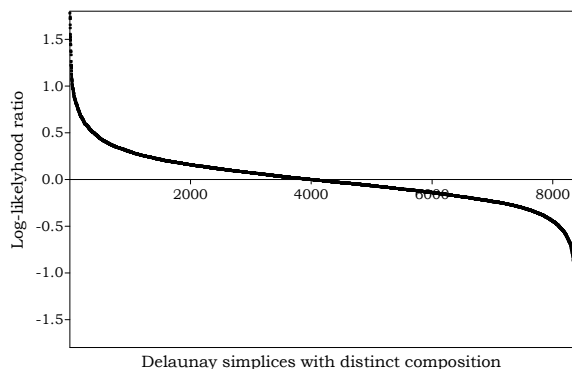
$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}} \quad (2)$$

where i,j,k,l are amino acid residues, f_{ijkl} is the observed frequency of occurrence of a given quadruplet, and p_{ijkl} is the frequency of random occurrence of a given quadruplet. The f_{ijkl} is calculated by dividing the total number of occurrences of each quadruplet type by the total number of observed quadruplets of all types. The p_{ijkl} is calculated as $p_{ijkl} = ca_i a_j a_k a_l$, where \mathbf{a}_i , \mathbf{a}_j , \mathbf{a}_k , and \mathbf{a}_l are the frequencies of individual amino acid residues, and c is the permutation factor, defined as

$$c = \frac{4!}{n \prod_i (t_i!)} \quad (3)$$

where n is the number of distinct residue types in a quadruplet and t_i is the number of amino acids of type i . The factor c accounts for permutability of replicated residue types in a quadruplet. The q_{ijkl} shows the likelihood of finding four particular residues in one simplex.

Theoretically, the maximum number of all possible quadruplets of 20 natural amino acid residues is 8,855. The distribution of likelihood factor for all possible compositions is shown in Fig. 4. The plot reveals highly non-random distribution: for some quadruplets observed frequencies are orders of magnitude higher (or lower) than expected from random model. Some



1	CCCC	3.081003
2	CCCY	2.13004
3	CCHH	1.960814
4	CCCG	1.782267
5	CCCH	1.742759
6	CCCW	1.724275
7	CCCS	1.724275
8	CCCQ	1.657329
9	CCCF	1.621613
...
...
8343	CDDL	-0.90166
8344	IRRV	-0.90217
8345	AEYY	-0.90535
8346	KKRV	-0.95081
8347	CKRS	-0.96133
8348	CEKP	-0.98433
8349	HKKS	-0.98472
8350	CGLR	-1.14737
8351	ACKN	-1.16297

Figure 4. Log-likelihood ratio for the Delaunay simplices (20-letter alphabet).

of the top scoring compositions, as well as several compositions with the lowest scores are listed in the table. It is worth to note the presence of cysteins in all of the quadruplets with the highest scores. High likelihood of the all-cystein cluster suggests that when two sulfur bridges are present in protein they tend to be together. While 20-letter

log-likelihood distribution contain significant amount of useful information and can be used as a statistical potential function, for some other types of analysis it may be more convenient to use reduced alphabets.

Reduced amino acid alphabets

More general trends in the clustering of amino acid residues in proteins can be identified easier if the analysis uses classes of residues grouped according to their properties. Such a classification may be based on various properties, e.g. on the residue chemical and structural properties, hydrophobicity of the side chains, genetic properties, etc. In a chemically based classification [11], the 20 amino acids are divided into three groups: (1) hydrophobic (F), (2) charged (L), and (3) polar (P) types; the hydrophobic amino acids include Ala, Val, Phe, Ile, Leu, Pro, Met, the charged amino acids include Asp, Glu, Lys, Arg, and the polar amino acids include Ser, Thr, Tyr, Cys, Asn, Gln, His, Trp.

Figure 5 shows the log-likelihood ratio for the fifteen possible quadruplets of the three types of amino acids among all simplices in tessellated proteins of the dataset. The quadruplets containing four or three residues of types F and P are much more likely to occur than the ones with four or three type L residues. Figure 5 illustrates also the differences in log-likelihood ratios for the same compositions belonging to the different sequential classes. E.g., the likelihood of quadruplet FFFF is very high in all classes except class {4}. It means that hydrophobic residues are likely to be found closely together in three-dimensional structure, unless they are neighbors in primary sequence. Opposite is true for the FFLL composition of the quadruplets. This type of analysis can provide an important insight into structural peculiarities of folded proteins.

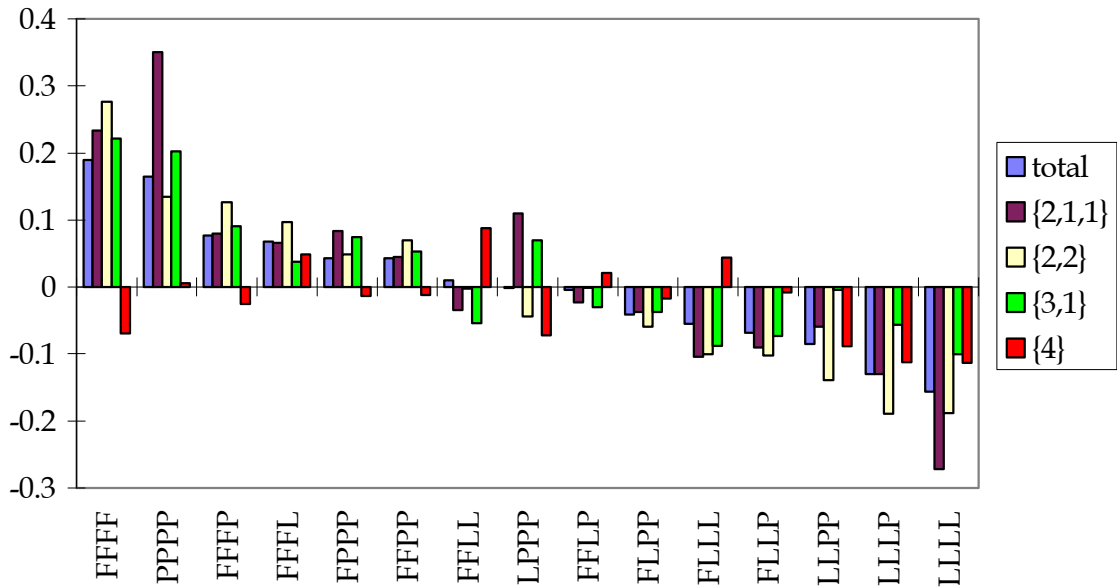


Fig. 5. Log-likelihood ratio for the Delaunay simplices (3-letter alphabet based on chemical properties).

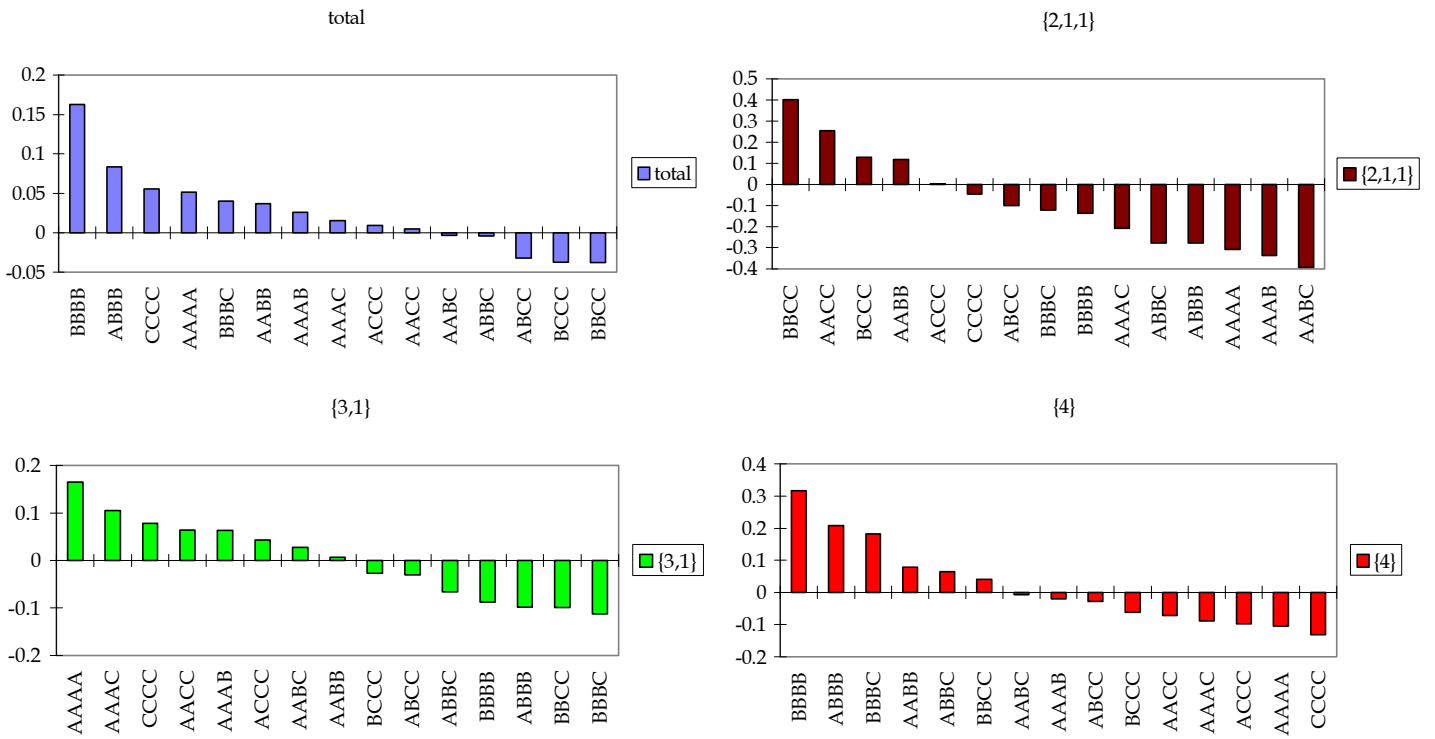


Fig. 6. Log-likelihood ratio for the Delaunay simplices (3-letter alphabet based on genetic code)

Amino acid residues could be classified according to their genetic code. Molecular recognition theory suggests that residues encoded by complementary codons may interact specifically [12]. This phenomenon of antisense peptide recognition may play an important role in determining the structure of folded proteins by influencing spatial proximity of matching (i.e. encoded by complementary codons) residues. In the classification based on the genetic code 20 natural amino acid residues can be divided into three independent groups: (A) Met, His, Val, Tyr, Asn, Asp, Ile, (B) Gln, Leu, Glu, Lys, Phe, (C) Trp, Pro, Arg, Gly, Ser, Ala, Thr, Cys. The residues in each group are related to each other by sense-antisense relationships. If the assumptions of the molecular recognition theory are correct, residues belonging to one group are likely to be found in the proximal positions in folded proteins more often than in random models. This hypothesis is in good agreement with the results of our analysis (Figure 6). Quadruplet compositions AAAA, BBBB, and CCCC occur much more frequently than expected from the random model (Figure 6a). As in the previous example, the likelihood distribution strongly depend on the sequential proximity of participating residues (Figures 6b-d). For example, compositions AAAA and CCCC that are highly scored in overall distribution, have the lowest scores for class {4} quadruplets.

Conclusions

This work demonstrates the applicability of the statistical geometry methods for protein structure analysis. The examples discussed in this paper deal with the objective and robust identification of nearest neighbors and analysis of clustering patterns in three-dimensional protein structures. The results of this analysis can be used in various algorithms for protein structure classification and prediction.

Bibliography

1. Bernal, J.D. A geometrical approach to the structure of liquids, *Nature*, 1959, **183**, 141-147.
2. Finney, J.L., Random packing and the structure of simple liquids, *Proc. R. Soc.*, 1970, **A319**, 479-493; Finney, J.L., Volume occupation, environment and accessibility in proteins, *J. Mol. Biol.*, 1975, **96**, 721-732.
3. Richards, F.M., The interpretation of protein structures: total volume, group volume distribution and packing density. *J. Mol. Biol.*, 1974, **82**, 1-14.
4. Chothia, C., Structural invariants in protein folding, *Nature*, 1975, **254**, 304-308.
5. Gerstein, M., Tsai, J., and Levitt, M. The volume of atoms on the protein surface: calculated from simulation using Voronoi polyhedra, *J. Mol. Biol.*, 1995, **249**, 955-966.
6. Voloshin, V.P., Naberukhin, Y.I. and Medvedev, N.N., Can various classes of atomic configurations (Delaunay simplices) be distinguished in random dense packings of spherical particles? *Mol. Simul.*, 1989, **4**, 209-227.
7. Vaisman, I. I., Brown, F.K., and Tropsha A., Distance dependence of water structure around model solutes, *J. Phys. Chem.*, 1994, **98**, 5559-5564.
8. Singh, R. K., Tropsha, A and Vaisman, I. I., Delaunay Tessellation of Proteins: four-body nearest-neighbor propensities of amino-acid residues", *J. Comp. Biol.* 1996, **3**, 2, 213-222
9. Tropsha, A., Singh, R. K., Vaisman, I. I., and Zheng, W., Statistical geometry analysis of proteins: implications for inverted structure prediction, Proc. Pacific Symp. on Biocomp., 1996.
10. Medvedev, N. N., Voloshin, V.P., and Naberukhin, Y. I., *J. Phys. A: Math. Gen.*, 1988, **21**, L247-L252.
11. Branden, C and Tooze, J. Introduction to Protein Structure, Garland Publishing, New York and London, 1991.
12. Mekler, L. B., Specific Selective Interaction Between Amino Acid Residues of the Polypeptide Chains. *Biophys. USSR*, 1970, **14**, 613-617.