Nonatomic Solvent-Driven Voronoi Tessellation of Proteins:An Open Tool to Analyze Protein Folds

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JANUARY 2003

Impact Factor: 2.63 · DOI: 10.1002/prot.10220 · Source: PubMed

6 AUTHORS, INCLUDING:

Borislav Angelov
Academy of Sciences of the Czech Republic
51 PUBLICATIONS **900** CITATIONS
SEE PROFILE

Jean-Francois Sadoc
Université Paris-Sud 11
123 PUBLICATIONS **1,865** CITATIONS
SEE PROFILE

Jacques Chomilier
French National Centre for Scientific Resea…
22 PUBLICATIONS **263** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Genetics 49:446–456 (2002)

Nonatomic Solvent-Driven Voronoi Tessellation of Proteins: An Open Tool to Analyze Protein Folds

Borislav Angelov,[1] Jean-François Sadoc,[1*] Rémi Jullien,[2] Alain Soyer,[3] Jean-Paul Mornon,[3] and Jacques Chomilier[3]

[1]*Laboratoire de Physique des Solides, Université Paris 11, Orsay, France*
[2]*Laboratoire des Verres,[†] Université Montpellier 2, Montpellier, France*
[3]*Laboratoire de Minéralogie Cristallographie, Universités Paris 6 et 7, case 115, Paris, France*

ABSTRACT **A three-dimensional Voronoi tessellation of folded proteins is used to analyze geometrical and topological properties of a set of proteins. To each amino acid is associated a central point surrounded by a Voronoi cell. Voronoi cells describe the packing of the amino acids. Special attention is given to reproduction of the protein surface. Once the Voronoi cells are built, a lot of tools from geometrical analysis can be applied to investigate the protein structure; volume of cells, number of faces per cell, and number of sides per face are the usual signatures of the protein structure. A distinct difference between faces related to primary, secondary, and tertiary structures has been observed. Faces threaded by the main-chain have on average more than six edges, whereas those related to helical packing of the amino acid chain have less than five edges. The faces on the protein surface have on average five edges within 1% error. The average number of faces on the protein surface for a given type of amino acid brings a new point of view in the characterization of the exposition to the solvent and the classification of amino acid as hydrophilic or hydrophobic. It may be a convenient tool for model validation. Proteins 2002;49:446–456.**

Key words: Voronoi tessellation; protein folding; hydrophilic/hydrophobic properties

INTRODUCTION

The folding of an amino acid chain to a protein of a well-defined structure is still an enigma. Tremendous amounts of experimental work have been done in the field of molecular biology, biochemistry, and biological physics to understand this complex phenomenon.[1–4]. In this work, we present the ground for development of a geometrical theory of protein folding. As an adequate theoretical description, it has been first accepted that the folding of a protein is ruled by the common principle of minimal free energy. This is commonly referred to as the "old view" of protein folding. More recently, a new view[5–7] was introduced, which has admitted a funnel-like energy surface[8,9] consistent with multiple folding pathways. In addition, it has been assumed that topology determines protein folding mechanisms.[10] Statistical analysis of contacting residues has shown that their localization is not randomly distributed but highly favors particular lengths of peptides between them. The literature that has been devoted to this subject is not normalized for the moment, because one can see different terms such as: contact order,[11] closed loops,[12] or tightened end fragments.[13]

How to predict the native state structure of a protein from its sequence[14,15] remains unclear. One possible way to overcome this failure of predictability of the molecular structure is not only to look at the energy landscape but also to examine in more details the information that comes from coordinates (i.e., from pure geometry of the protein structure). In the field of liquids, liquid crystals, crystalline, and amorphous solids, the geometrical approach yielded many fruitful results.[16–20] To analyze the structure of folded proteins, it was proposed by some of us[21] to use a very sensitive geometrical method based on the so-called Voronoi tessellation (VT).[22] A tessellation is a mean to describe the space filled by a packing of solid polyhedra connected by their faces without empty space between them. Giving a set of discrete points in space, a Voronoi tessellation associates to each point a polyhedral domain, called a Voronoi cell, containing all the neighborhood closer to the considered points than to others. There are several examples of VT methods applied to proteins in the literature,[23–28] but only a few of them[26–28] concern directly the packing of amino acids (AA) or fold recognition.[29] Moreover, in Refs. 26–28, the investigators used a Delaunay tessellation, which can be viewed as a first step before VT, and considered the α-carbon locations as the starting set of points. Because an α-carbon is almost

Abbreviations: AA, amino acid; PDB, Protein Data Bank; RRPS, relaxed random packing of spheres; RSA, random sequential aggregation; VT, Voronoi tessellation; VC, Voronoi cell.

Grant sponsor: Marie Curie Program of the European Union; Grant sponsor: Centre National de la Recherche Scientifique, France.

B. Angelov's permanent address is Institute of Biophysics, Bulgarian Academy of Science, Acad. G. Bonchev Str. Bl. 21, Sofia 1113, Bulgaria.

*Correspondence to: Jean-François Sadoc, Laboratoire de Physique des Solides, Université Paris 11, Centre d'Orsay, 91405, Orsay, France. E-mail: sadoc@lps.u-psud.fr

[†]Laboratoire associé au Centre National de la Recherche Scientifique (CNRS, France).

Received 9 January 2002; Accepted 7 June 2002

systematically located on the border of the volume occupied by its corresponding AA, we have preferred, in this study, to consider the geometrical centers of individual AA. We think that, with this choice, the cells are representing topologically better the true volume occupied by the AA and give more homogeneous distributions of distance (e.g., avoiding a peak corresponding to C_α–C_α first distances). In recent studies on AA packing in proteins on a coarse-grained scale, the calculations were performed by using C_β to represent each residue.[30,31]

The purpose of this article is to go further in investigating the protein folding in terms of VT. To build the Voronoi cell (VC) of a given amino acid, it is necessary to know the position of its neighbors. This condition is easy to satisfy for AA inside the protein volume, but for AA that are on the surface or inside a cavity, the VC definition could be ambiguous. Here we solve this problem by surrounding the protein with a model of solvent, often called environment in this article, whose characteristics are similar to generic proteins considered as dense packing of spheres whose volumes are the average AA volume. The cell characteristics provide essential information on the local geometrical properties of the considered packing. This tool is widely and currently used to study random sphere packings, granular materials, foams, froths, and glasses.

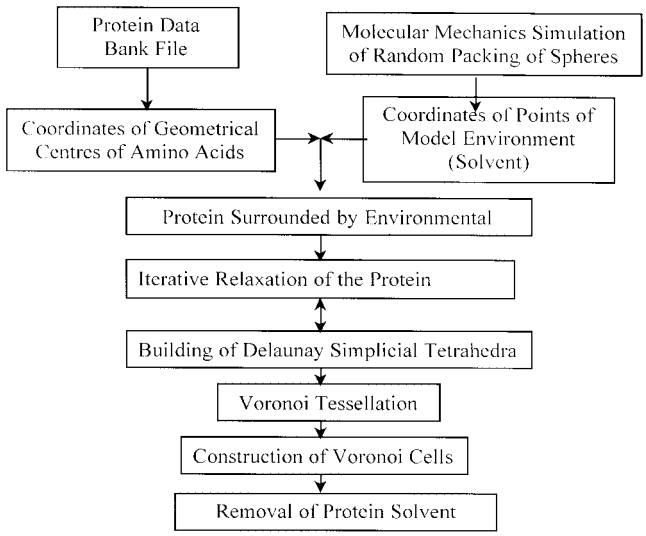
Voronoi cells are highly characterized by the number of faces and edges. Statistics of edges and faces of VCs corresponding to a given type of amino acid are presented. In these statistical distributions, the solvent effect was investigated, and the relaxed random packing of spheres (RRPS) appeared as the most suitable environment. The analysis of edge and face distribution per amino acid type (i.e., 20 different distributions) led to their "absolute" classification from only geometrical and compactness data, giving a tool for structure prediction based on burying of AA. An attempt is made to collect statistics for faces between different neighbors.

Details on the performed calculations of VT, the distributions of edges and faces of VC, and concluding remarks are given in the following three sections.

VORONOI TESSELLATION OF A PROTEIN WITH A MODEL ENVIRONMENT

This work follows the method previously[21] presented; in its technical aspects, it can be represented as a procedure. In a first step, it is necessary to transform the original atomic coordinates file from the Protein Data Bank[32] to a file containing only the coordinates of amino acid geometrical centers, including backbone.

To determine Voronoi cells in a set of points, the following step consists in defining dense packing of tetrahedra joining the points defining the AA (the Delaunay tetrahedral decomposition). Then from this tetrahedral packing, Voronoi polyhedral cells are built, with vertices on centers of tetrahedra. An important property of Voronoi decomposition results from the fact that it is unique: it defines neighbors of a point in an absolute topological way. Nevertheless, in a finite set of points, a difficulty appears for points close to the surface. Resulting VC on the protein surface appear as elongated or open cells that are not realistic images of the volume associated to corresponding AA. To overcome this problem of surface VC, the protein is embedded inside a model environment, which has packing properties close to those of proteins. This environment can be considered as a solvent made of a generic AA. But further extensions of this study could concern more realistic solvents as water or even lipid for membrane protein studies. VC of this model solvent are not considered in the analysis, but the environment is involved in the definition of VC faces belonging to the surface of the protein. The code used for VT is the same as in Ref. 21, but other investigators[30] use, for instance, the delaunay3.m function of the Matlab 6.0. Using our personal code will allow having further developments and generalizations of the Voronoi procedure. Block Schema 1 depicts the complete procedure whose different steps are developed below.



Schema 1. Schematic representation of the algorithm used in the present study.

The data set used is a representative bank of 39 proteins made of 46 chains used in the previous study.[21] The PDB codes are "1arb, 1bdm, 1bp2, 1cdg, 1csn, 1cus, 1eca, 1enh, 1esl, 1fas, 1fds, 1frd, 1fxd, 1gse, 1hle, 1hvc, 1knt, 1lec, 1nhs, 1phb, 1phm, 1pk4, 1plf, 1pmy, 1ptf, 1rro, 1sha, 1tml, 1tud, 1xnb, 2act, 2apr, 2mcm, 2mhr, 3chy, 3pte, 4gcr, 5p21, 8abp". The total number of AA used to collect statistics is 8570.

Criteria for Selection of RRPS as a Model Environment

To simulate the solvent, the environment was spread around the protein as a shell of constant thickness. A point representation of an example protein with its environment is given in Figure 1. Bold dark points correspond to amino acid geometrical centers, whereas small points are centers of spheres packed in the solvent. If there were a cavity inside the protein, it would also be filled with environment. The volumes of the spheres used to simulate the solvent are taken close to the mean volume occupied by AA. This

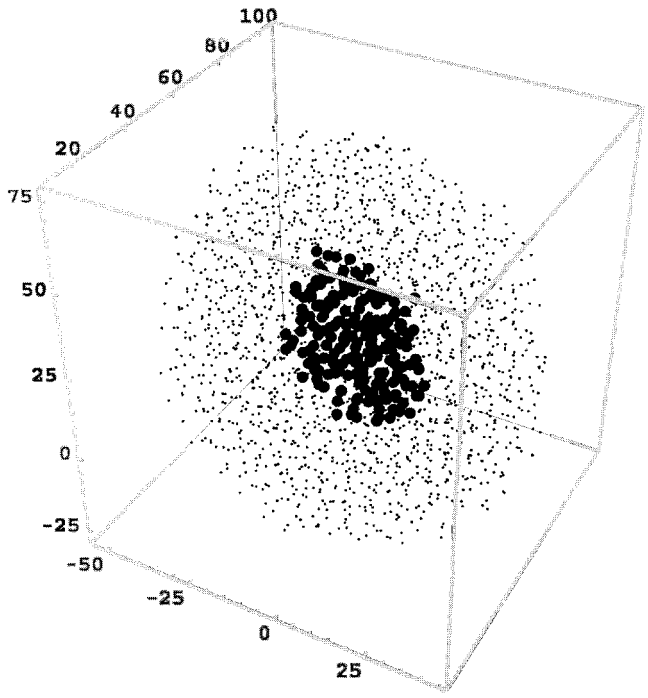


Fig. 1. Example of a protein (PDB code: 1cus) surrounded by an environment of relaxed random packing of spheres. Bold points are the geometrical centers of AA. Small points are the centers of spheres. Axes of the box represent distances in Angstroms.

leads to a diameter for environment spheres of 7 Å. For each protein from the data set, whose mean number of AA is close to 200, the total number of points, i.e. the number of AA plus the number of solvent spheres, was kept to be close to 2000 points. These spheres have been selected among a larger set of 8000 spheres packed in a cubic box. The thickness of the environment, corresponding to at least three shells of solvent spheres around the protein has been tested to be sufficient. More than 2000 solvent points just make the procedure time consuming without expected improvement of the accuracy of the results.

We have checked that the VC of the environment spheres that neighbor the protein are perfect VC not influenced by the outside surface of the aggregate formed by the protein and its surrounding environment.

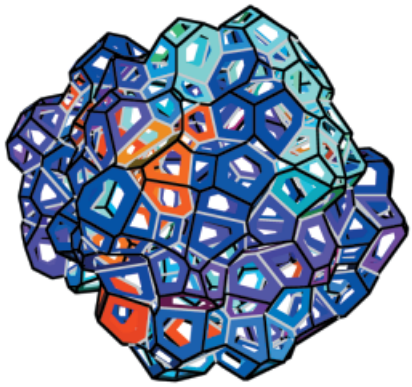
In fact, the VC for the environment could be classified as three types: (a) those in contact with the protein, and so the more important because they are used to define the protein surface; (b) those that are at the surface of the aggregate formed by the protein and all its environment; they are imperfect cells often unclosed, with elongated faces; and (c) the other remaining cells in volume of the environment. During the relaxation, the protein is rigid and the cells on the aggregate surface are also fixed, so the environment is confined between two surfaces. A too small number of environment cells in between these two surfaces leads some of the environment cells sharing faces with AA cells to also share faces with cells on the aggregate surface. This has never been observed in any used aggregate; it warrants a good coverage of the protein surface with a sufficient thickness.

The random packing of spheres with relaxation was chosen as a model solvent after testing another four alternative environments. The following criteria were applied to select the best environment:

1. Cells associated to AA at the protein surface are cells in contact with those associated to solvent. The ratio of cells at the protein surface over the total number of cells has to be close to 2:3, which is realistic for the size of proteins used in the bank.[21] Some environments, with a low density, can artificially decrease this ratio; for example, a small protein surrounded by a few big spheres and a big protein surrounded by many small spheres would have a different percent of surface cells.
2. The distribution of faces for cells in the volume and the distribution of faces for cells at the surface should be similarly shaped. Although these distributions have different average values, their shapes should be smooth with single peak and similar width.
3. The distribution of faces on the surface per cells at the surface should be without strong singularity compared with other distributions to exclude elongated faces.
4. The number of neighbor cells on the AA chain that do not contact each other should be close to zero, to confirm the choice of the radius of environment spheres.

We have tested different close packing of spheres as environments, including crystalline or disordered. The best choice according to the previous criteria is the RRPS, which is a disordered dense structure obtained by closely packing spheres. It is the structure formed by a set of balls in a bag. An efficient algorithm for generating random packings of spheres was proposed by Jullien et al.[33] after some modifications of the well-known Jodrey-Tory algorithm.[34]

Putting the protein in the environment and removing solvent spheres superimposed to AA spheres leads to a discrepancy on the surface between the protein and the environment. Then an iterative relaxation of the environment was performed to adjust it to the protein and to give the best possible tessellation. We proceeded by considering the protein and its environment as a random dense structure. So, during this relaxation, spheres representing AA are kept fixed; only centers of spheres of the environment are moved, assuming a constant volume during the relaxation. At every step of the relaxation the VC are built, and then the sphere centers are shifted toward the geometric center of VC for environment sites. This relaxation tends to regularize VC of the environment: elongated cells become more isotropic, with a narrow distribution of their volume. Consequently, the environment is well adjusted to the protein, avoiding a density singularity at the protein surface. After six cycles of relaxation the environment remains fixed, but we performed nine cycles. In the following statistics, the VCs related to the solvent were removed after completing the VT procedure. At this stage of modeling of the protein solvent, RRPS gives statistically reproducible results concerning the protein surface. For instance, as presented in section "±nth double face," VC



PHE, CYS, ILE, LEU, VAL, TYR, MET, TRP, HIS, ALA,
GLY, THR, ARG, SER, ASN, GLN, PRO, ASP, LYS, GLU

Fig. 2. Voronoi tessellation of the signal transduction protein Che Y from *Escherichia coli* (PDB code: 3chy). Colors of cells correspond to hydrophilic/hydrophobic properties of AA (respectively, blue to red). Black lines border the faces of one cell, whereas white lines separate the faces that belong to different cells. Only the protein surface is shown.

faces on the surfaces have the same mean number of edges (5.03) than other faces that are not between two neighbor sites along the chain.

The Protein 3chy as an Example of VT

A complete VT of the signal transduction protein Che Y from *Escherichia coli* (PDB code 3chy; 128 AA; 1 chain) is shown on Figure 2. We have used the atomic coordinates from crystallographic X-ray structure[35] obtained at 1.7 Å resolution. The geometrical centers of every AA have been calculated. The color of cells in Figure 2 corresponds to hydrophilic (in blue) and hydrophobic (in red) properties of AA (see Table II). The environment of random packing of spheres is simulated under periodic boundary conditions. A box of about 1000 randomly packed spheres is translated eight times to construct a bigger box of 8000 spheres that completely embeds the protein. Spheres that overlap with the AA are removed, as well as spheres too far from the protein. Finally, the total number of AA and spheres is about 2000. The complete procedure on a larger number of points can take a too long time. The time-consuming subprograms are written in high performance FORTRAN 95 and executed on a Compaq DEC Alpha workstation. These Fortran subprograms were managed as a function from a Wolfram Mathematica 4 notebook. Total time to generate and relax the environment and then to tessellate is about 20 min for a protein as 3chy on a standard PC computer and a few minutes on a workstation. Figure 2 and the other figures in this article are made with Mathematica.

STATISTICAL DISTRIBUTION OF EDGES AND FACES OF VORONOI CELLS

Surface and Volume Cells and Faces

Types of Voronoi cells: surface and volume

Because the environment has been removed, the protein surface is perfectly defined by all uncovered faces of VC. VC can be split into the family of cells on the protein surface and the family of cells in the protein volume. If a given VC has at least one face on the protein surface, it is considered as a surface cell. The cells with zero face on the protein surface are considered as cells in the volume. The distribution of number of faces per cell $h(f)$ and edges per face $h(e)$ are compared for these two types of cells in Figure 3. Volume and surface distributions are similar in shape. The average number of edges per face is almost similar for surface and volume with values of 5.15 and 5.18, respectively. But the number of faces per cell is slightly greater in the volume (14.65) than on the surface (14.13). They obey the following relation, which is a theorem deduced from the Euler relation (see page 81 in Ref. 16):

$$<f> = \frac{12}{6 - <e>} \tag{1}$$

Notice that this relation leads to large variations for $<f>$ even for small changes of $<e>$. The statistical distribution of edges per face $h(e)$ and faces per cell $h(f)$ for all VC and each type of AA is given in the left plot of Figure 4. The mean values for all VC are $<e> = 5.16$, $<f> = 14.27$, and they also obey the relation in Eq. 1. Remember that the $<e>$ value is related to the ratio $2\pi/\theta$ where θ is the regular tetrahedron dihedral angle.[16] In usual dense structures this number remains in the range of 5.1–5.2. It was found[21] that $h(f)$ distribution for proteins is very well modeled by $h(f)$ distribution for finite clusters made of random sequential packing of spheres [random sequential aggregation (RSA)].[36] The number $h(e)$ for volume cells given here is higher than the one given in Ref. 21. Actually, some close cells, considered as volume, have fewer faces because they are very elongated because of the absence of solvent in the previous article. The present value agrees very well with the RSA model value. It does not mean that a particular protein structure should be considered as a random one, and it must be noticed that protein distribution $h(f)$ represents an average over many structures. Nevertheless, it could indicate that there are properties of folding that are related to sequential aggregation. To test the sensitivity of these values we have reduced the protein bank to only five proteins: this changes the $<f>$ values only by ± 0.02; the $<e>$ values are still less sensitive but are related exactly to $<f>$ by Eq. 1. Convergence toward these values goes very quickly as the number of used proteins increases.

In Table I, the average values $<f>$ and $<e>$ are given for each type of AA. They are sorted in decreasing order of $<f>$. It is clear that each AA has a specific histogram related to the interaction with its neighbors. The number of faces of a particular VC directly gives the number of topological neighbors of the corresponding AA as long as one considers the solvent as neighbors. Therefore, the histogram of the number of faces per cell $h(f)$ represents the distribution of neighbors.

For example, it is possible to distinguish between AA according to the value of $<f>$. There are AA with a large number of neighbours ($<f> > 14.8$); others have this

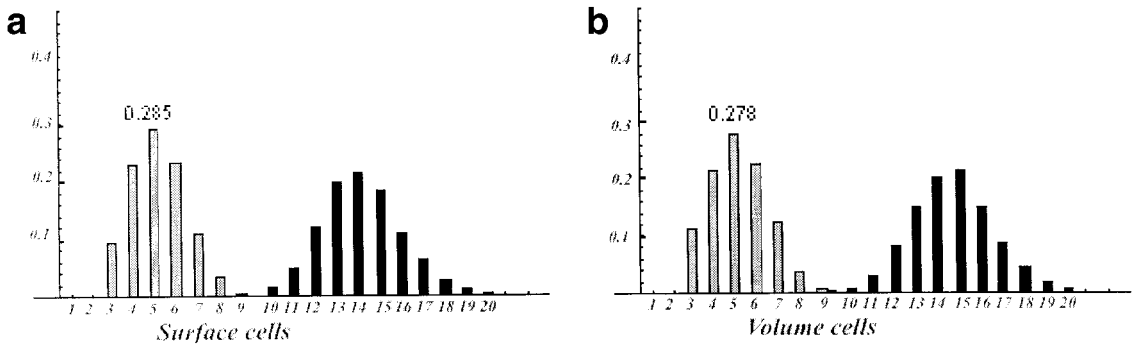


Fig. 3. Distribution of edges per face *h(e)* (gray histograms) and faces per cell *h(f)* (black histograms) for VC on the protein surface (**a**) and in the protein volume (**b**). The values above histogram maximums are given as a reference. These distributions are calibrated to 1, which correspond to 100%.

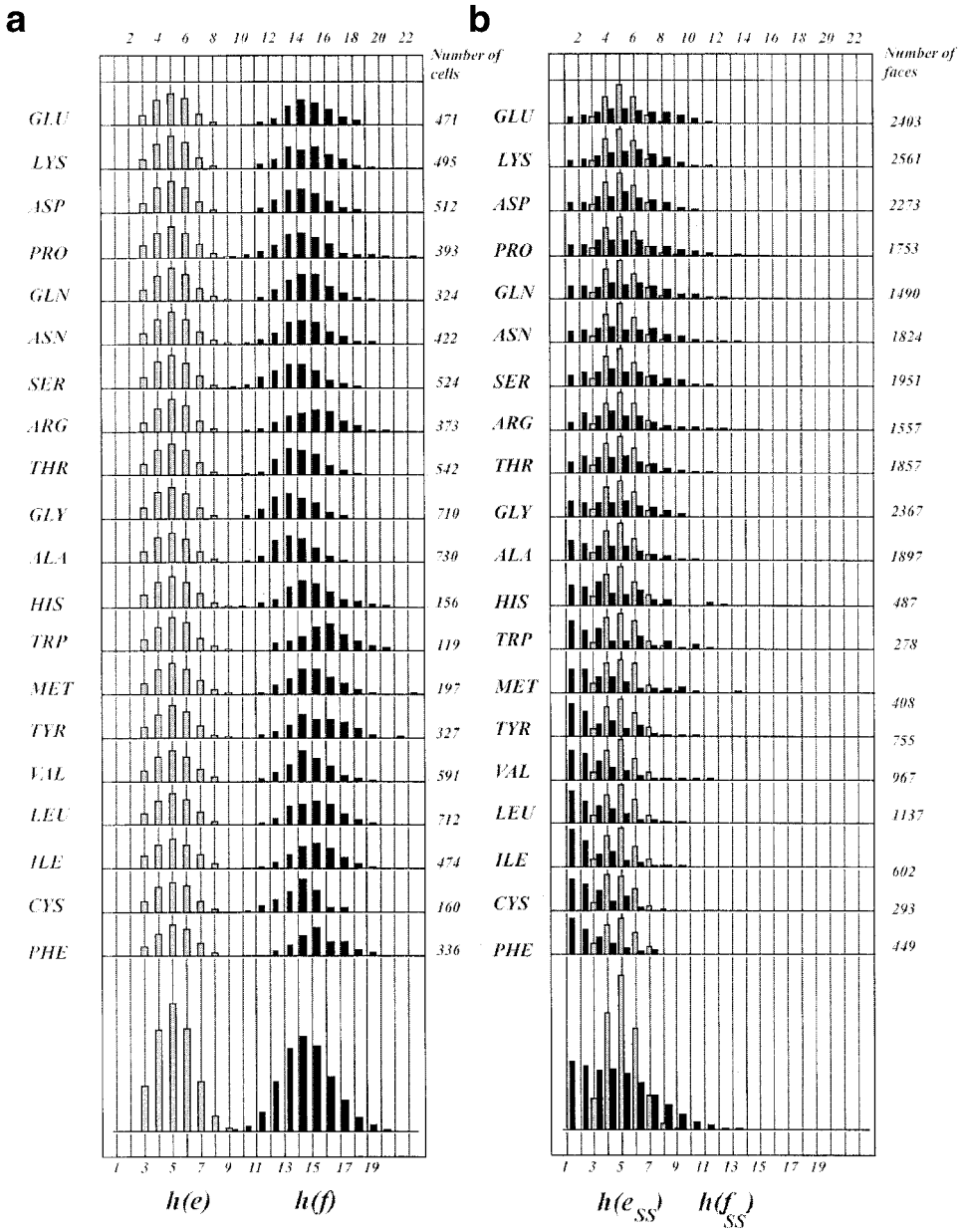


Fig. 4. Distribution for each AA type of edges per face *h(e)* (gray histograms) and faces per cell *h(f)* (black histograms) for all VC (**a**). The number of cells per AA used to collect these data is given in (**b**). The plot on the right is the distribution of edges per surface face $h(e_{SS})$ (gray histograms) and of surface faces per surface cell $h(f_{SS})$ (black histogram). The number of surface faces per AA is given on the right axis. For both plots the histograms at the bottom are weighted sum of all AA. The average number of $h(e_{SS})$ is close to 5 e/f (see Table II).

TABLE I. Average Number of Faces per Cell $\langle f \rangle$, Edges per Face $\langle e \rangle$, and Edges per Chain Face $\langle e \rangle_{+1}$ for Every Type of Amino Acid and for All Amino Acids

	Faces per cell $\langle f \rangle$	Edges per face $\langle e \rangle$	Edges per chain face $\langle e \rangle_{+1}$
ALL	14.27	5.16	6.33
TRP	15.52	5.23	5.89
PHE	15.22	5.21	6.19
TYR	15.12	5.21	6.11
ILE	14.88	5.20	6.51
ARG	14.82	5.19	5.91
LEU	14.82	5.19	6.39
MET	14.73	5.19	6.31
HIS	14.65	5.18	6.29
GLU	14.61	5.18	6.33
GLN	14.45	5.17	6.27
LYS	14.45	5.17	6.16
VAL	14.40	5.17	6.42
ASN	14.22	5.16	6.31
PRO	14.09	5.15	6.53
ASP	14.04	5.15	6.31
THR	13.89	5.14	6.38
CYS	13.71	5.13	6.37
SER	13.51	5.11	6.37
ALA	13.42	5.11	6.49
GLY	13.35	5.10	6.33

Amino acids are sorted by $\langle f \rangle$ in decreasing order. Corresponding histograms $h(f)$ and $h(e)$ are given on Figure 4.

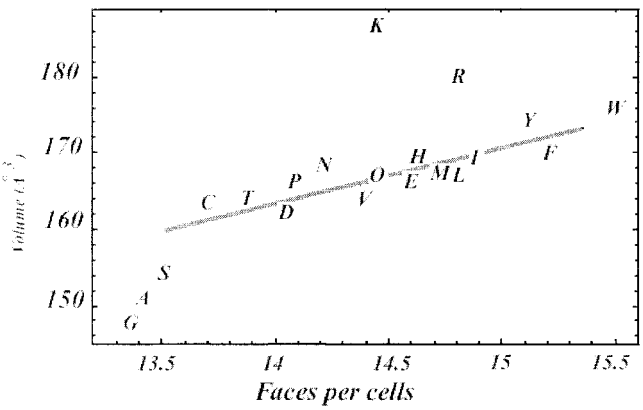


Fig. 5. Relation between average volume of VC and mean number of faces per cell for all AA. The twenty AA are labeled by using the standard one letter code.[33] Average volume of VCs of AA was taken from Ref. 21.

number close to the average number, and it remains a group of AA with $\langle f \rangle < 14.0$. For example, we can compare glycine ($\langle f \rangle = 13.35$) and isoleucine ($\langle f \rangle = 14.88$), which show a clearly shifted histogram for the $h(f)$ distribution (Fig. 4). The possible relation between the number of faces per cell and their average VC volume is given on Figure 5. This relation seems to be linear with a notable deviation only for glycine, alanine, lysine, and arginine. Large AA have a large number of neighbors: the larger the number of faces per cell, the larger the volume for most of the AA.

TABLE II. Average Number of Single Faces per Surface Cell $\langle f \rangle_{ss}$, *Faces per Surface Cell* $\langle f \rangle_s$, and Edges per Single Face $\langle e \rangle_{ss}$ for Every Type of Amino Acid (AA) and for All AA Together

	Single faces per surface cell $\langle f \rangle_{ss}$	Faces per surface cell $\langle f \rangle_s$	Edges per single face $\langle e \rangle_{ss}$
ALL	4.32	14.13	5.03
GLU	5.59	14.58	5.11
LYS	5.34	14.40	5.07
ASP	5.03	13.97	5.04
PRO	5.00	14.14	5.07
GLN	4.98	14.44	5.08
ASN	4.96	14.20	5.06
SER	4.69	13.47	5.00
ARG	4.65	14.75	5.12
THR	4.33	13.81	5.00
GLY	4.30	13.36	4.98
ALA	4.00	13.40	4.96
HIS	3.84	14.65	5.01
TRP	3.81	14.86	5.08
MET	3.71	14.39	5.02
TYR	3.12	15.02	5.02
VAL	3.07	14.19	4.96
LEU	3.02	14.44	4.99
ILE	2.81	14.55	4.99
CYS	2.71	13.62	4.84
PHE	2.67	14.71	4.93

Amino acids are sorted by descending order of $\langle f \rangle_{ss}$. Histograms $h(f)_{ss}$ and $h(e)_{ss}$ are given on Figure 4. Hydrophilic/hydrophobic properties of AA closely follow $\langle f \rangle_{ss}$. Most hydrophilic A.A are at the top of the table, whereas most hydrophobic AA are at the bottom of the table.

The number of edges per face is related to the symmetry and type of interactions between neighboring AA. Faces with many edges are more likely to have larger areas and a stronger contact between related AA. The number of edges ranges from three to eight. Nine and ten edges occur negligibly rarely. For proteins, the distribution $h(e)$ is not symmetric, and the maximum is at five edges as it is shown in Figure 4. If we consider faces with five edges as a reference, then the faces with three and four edges would correspond to unfavorable interactions among neighbors, whereas the faces with six, seven, and eight edges would be related to favorable interaction.

Types of faces: single and double

If we consider faces of VC with respect to the number of cells they belong to, there are two kinds of faces: single and double. Faces on the protein surface only contact the solvent, and they do not share any face with other cells from the protein, so they are single. The other faces are double, because they belong to two neighbor cells of the protein. Single faces are suitable for the analysis of surface properties of AA. The mean numbers of faces per surface cell $\langle f_S \rangle$ and single faces per surface cell $\langle f_{SS} \rangle$ for every type of AA are given in Table II. $\langle f_{SS} \rangle$ are given in this table in decreasing order and show a smooth regularity. It is interesting to notice that the average number of sides for the surface (single) faces is very close to five: $\langle e \rangle_{ss} \approx 5.03$,

lower than the average value for all faces (5.16) and lower than the value in a random close packing of spheres ($<e>$ = 5.15), and so, for the surface of any cluster taken in such a packing.

The average numbers of single faces per surface cell can be used to define AA exposure to the solvent. The most hydrophilic AA tend to have more than five single faces per surface cell versus less than three single faces per surface cell for the most hydrophobic ones. It is easy to distinguish the contrast between the hydrophilic group of Glu, Lys, Asp, Arg, well exposed to the solvent, and the hydrophobic group of "hidden" AA Val, Leu, Ile, Phe. It is worth noting that the hydrophobic Pro[37,38] has 5.0 single faces per surface cell as for a typical hydrophilic AA. It is not due to uncertainty, because the statistics for proline is good enough, but marks again the peculiar properties of proline. The sorting using the number of single faces per surface cells almost exactly corresponds to the sorting done with the percentage of occurrence of a given AA in the bulk of a protein (see Ref. 21 for a definition). The most solvent exposed AA has smaller percentages of occurrence in the bulk of the protein. In Ref. 39, it is shown that hydrophobic AA Trp, Met, Tyr, Val, Leu, Ile, and Phe have higher propensities to form either an α-helix or a β-sheet instead of a coil (group I in Fig. 3 in Ref. 39). This group of AA has average numbers of single faces per surface cell in the range of 2.67–3.81 and occupies the bottom of Table II. The rest of AA, which form equally regular secondary structures and coils for group II or preferentially coils for group III, have average numbers of single faces per surface cell which are more dispersed.

Faces Between Different Neighbor Cells $\pm n^{th}$ double face

Double faces could be used to analyze local structures that are not exposed to the solvent. First, the faces threaded by the main-chain are situated between every pair of neighbor cells i and $i + 1$ consecutive along the AA chain. When the index i is allowed to go from the first to the last residue of the chain, it will pick up all such faces. The mean numbers of edges for the faces common to the i and $i + 1$ cells, denoted here as $<e>_{+1}$, have range interval from 5.89 (Trp) to 6.53 (Pro), and the mean number of edges for all types of AA for this kind of faces is $<e>_{+1}$ = 6.33. This is significantly higher than $<e>$ = 5.16 and clearly related to the type of AA (Table I). In such protein tessellation, this results from the covalent bonds relating consecutive α-carbon atoms. When the distance between the geometrical centers of two AA is shorter than the average one, the resulting face after VT has a larger area than the average area per face. Because of the conservation of the volume of the individual AA, it follows that faces with larger area will have more edges than the average number of edges per face $<e>$. Covalent bonds along the main-chain are among the strongest bonds in proteins. The resulting distances are shorter, and the corresponding faces of VC have more edges representing a better contact.

An attempt is made to collect statistics for faces between contacting cells. If there is a face between cells i and j, where $j > i$ and $j = i + n$, it is denoted here as the $+n$th face for convenience (instead of face ij). Conceptually, this can be related to the contact order relating the sequence interval between a pair of neighbor residues in the three-dimensional space.[11,40,41] When $j < i$ and $j = i - n$, the corresponding face is denoted as the $-n$th face. The range interval for n depends on the length of the chain and on the existence of contacts between cells. For example, if the chain has 120 residues, the contact face with maximal value of n occurs for $n = +119$ (i.e., between first cell ($i = 1$) and last cell ($j = 120$). Obviously, this face will coincide with face -119. Figure 6 shows the number of nth neighbors denoted as $N_f(n)$ in the present bank of 39 proteins. The number of contacts between i and $i+n$ neighbors found in our analysis rapidly decreases and becomes insignificant after n = 120 and zero for n = 361. The last observed neighbor is at $n_{\max}$ = 549. The number of neighbors in the range n = 361–549 fluctuates between 0 and 5, and it is not plotted in Figure 6. Figure 6(A) presents the histogram of contacts between VC as a function of n on a log-log scale. Figure 6(B) is an onset on a linear scale to focus on the irregular decrease of this distribution. Actually, a local minimum can be seen around 18th neighbor and a local maximum occurs near 27 as expected from recent observations.[13,42] This bump has been shown to be due to closed loops of mean standard size 27 AA, that have been called Tightened End Fragments.[13] It means that, on average, nature favored some particular fragment length that was able to turn back to the core of the protein with ends close in the three-dimensional space.

Average numbers of edges per faces shared by nth neighbors are given on Figure 7. The first 120 neighbors that have representative statistics are shown. It appears that the first five neighbors constitute the short-range order on the AA chain. The average numbers of edges per face alternate with increasing n; for n = 1, 3, 5 they are above 5, whereas for n = 2, 4 they are below 5. After the 5th neighbor, the pattern is not so clearly alternating (see above: average numbers at nth faces per cell decrease very rapidly). The statistics for neighbors with n bigger than 35 is not very credible because of the rare occurrence of such neighbors.

Another way to represent the average numbers of edges per nth double face consists of considering cumulative averages. To a large extent, this approach overcomes the problem of low statistics for rare neighbors. The cumulative average number for edges per nth face denoted here as $E(n)$ is shown in Figure 8. It was calculated by using the formula:

$$E(n) = \frac{\sum_{i=n+1}^{n_{\max}} N_e(i)}{\sum_{i=n+1}^{n_{\max}} N_f(i)} \qquad (2)$$

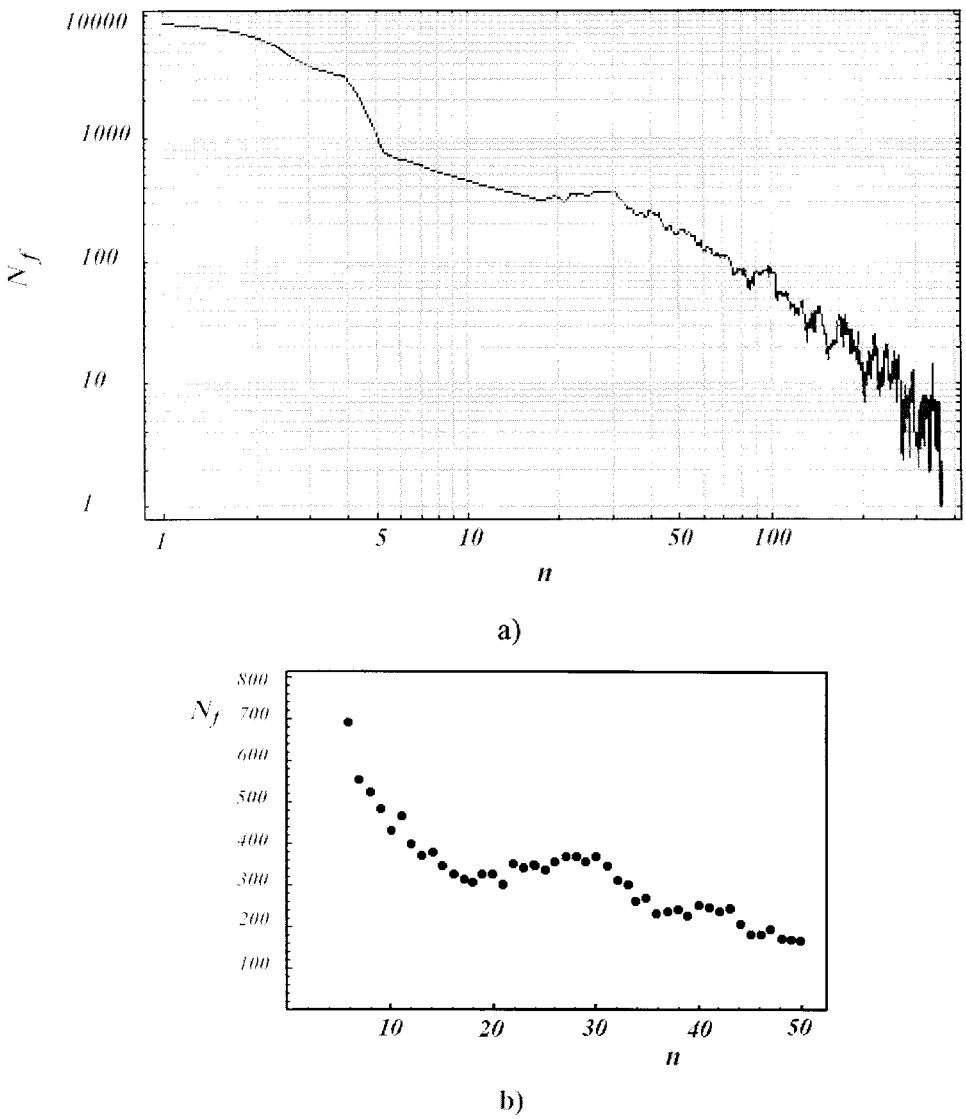


Fig. 6. **a:** The number of *n*th neighbors $N_f(n)$ as a function of sequence distance n between contacting cells. Both axes are logarithmic. **b:** The same function $N_f(n)$ in linear scale: the bump corresponding to tightened end fragments at n = 27 is clearly visible.

where $N_e(i)$ is the total number of edges of all *i*th faces. *Nf(i)* is the number of all *i*th faces, and it is the same as the numbers of *i*th neighbors shown in Figure 6. Both sums in Eq. 2 are running from $n + 1$ to $n_{\max}$. $E(0)$ represents the edges per all double faces. For $n = 1$, the edges and faces of 1st neighbors along the sequence ($N_f(1) = 8507$ faces with $N_e(1) = 53804$ edges) are not included in the sums. For $n = 2$, the edges and faces of 1st and 2nd neighbors are not included, and so on. After removing the first five neighbors, the cumulative average number $E(n)$ practically stops to evolve and becomes equal to the average of $<e>_{ss}$ for single faces. Thus,

$$E(5) \approx E(6) \approx E(7) \approx \ldots \approx <e>_{ss} \approx 5.03 \qquad (3)$$

The fact that these values are clearly apart from the average number currently found in disordered cell packings[16] ($<e> \approx 5.15$) μ but also for the VT of proteins is a topological evidence of the hierarchy between secondary and tertiary structures. In secondary structures there are two kinds of faces. The first type concerns those with a large number of edges ($<e> > 5.2$) corresponding to contacts between AA separated along the chain by $n = \pm 1$, ± 3, ± 5. The second type corresponds to other faces with a small number of edges separated by even *n* steps. But if we consider contacts between two AA belonging to different secondary structures, so corresponding to tertiary folding, the $<e>$ mean value is very close to 5, a value significantly below the average number of edge for all faces. From Figures 7 and 8, it could be concluded that first four (and even the fifth) neighbors contribute to the local structure. Another nontrivial observation is the alternating way of change of $E(n)$. Remember that small numbers of edges

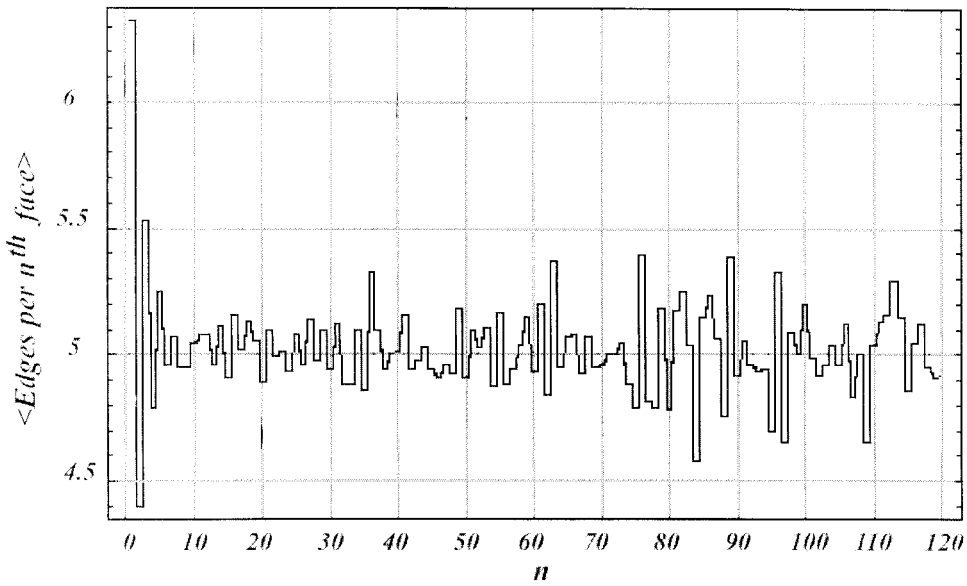


Fig. 7. Average numbers of edges per face $<e>_n$ for the faces of *n*th neighbors. These values are significant up to n = 35; for largest *n*, the statistic is poor. It is clear, at least for small *n* values that $<e>$ values alternate.

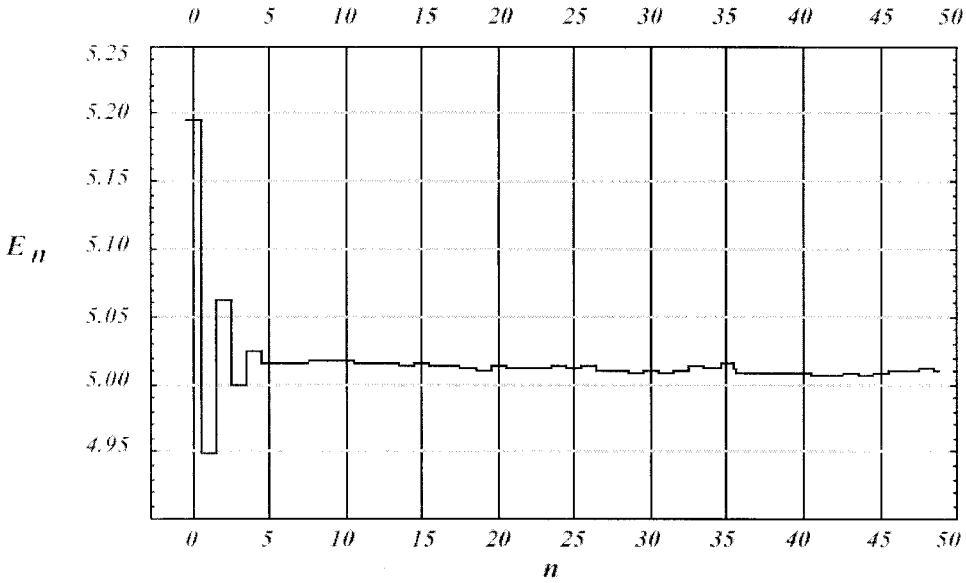


Fig. 8. Cumulative average numbers of edges per double face *E(n)* as a function of *n*th neighbor. *E(n)* are calculated according to Eq. 2.

correspond to large distance between neighbors and vice versa. So the alternating variation of *E(n)* means that, on average, 1st, 3rd, and 5th neighbors are closer and with better contacts than the average one. For 2nd and 4th neighbors, in opposite, their contacts are weaker than the average one. Such combination of alternating neighbors mainly concerns secondary structures. There are no evidences that $<e>ss$ depends on the number of AA of the protein chain, at least for investigated proteins composed of 55–550 AA. Such property means that addition or removal of an amino acid at the terminal ends of the chain does not change the value of $<e>ss$ (i.e., it is a property of those faces of the single amino acid contacting the solvent or other AA in different secondary structures). This also shows that the RRPS environment realistically models the external faces of the secondary structures toward the solvent, because these faces appear similar to faces separating two secondary structures.

Alpha-helices: ±3rd and ±4th double faces

The *h(e)* and *h(f)* distributions for right-handed α-helices are given in Figure 9(A). Corresponding averages are $<e>$ = 5.16 and $<f>$ = 14.34. To collect statistics for the α-helical surface, the α-helices are considered as small

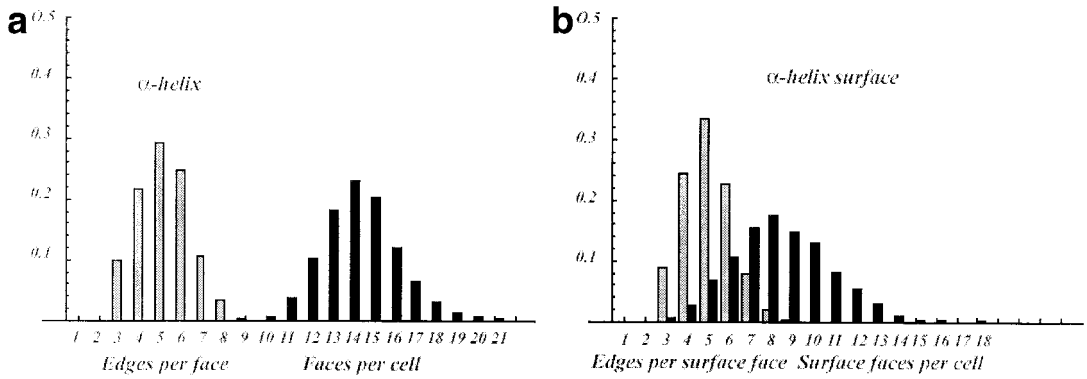


Fig. 9. Distribution of edges per face (gray histograms) and faces per cell (black histograms) for VC of α-helices (**a**). The α-helices are considered as small proteins surrounded by an environment. The edges per "surface" face and "surface" faces per cell are shown in (**b**).

proteins surrounded by the native environment formed by other parts of proteins, not the RRPS environment, so single faces in this case are simply external faces of the α-helix. The number of edges per single face and the number of single faces per cell are shown in Figure 9(B). They have averages $<e_{SS}> = 5.04$ and $<f_{SS}> = 8.44$. The number of double faces of an α-helix can be obtained by subtracting the single faces from all faces. It is not surprising that it gives a value close to six double faces per cell. Every AA of the α-helix has only six neighbors from the same α-helix. Namely, these neighbors are ± 1, ± 3, and ± 4. The carbon chain is presented by ± 1 neighbors, and the number of edges for $n = +1$ faces is 6.38, a value close to the one for the whole proteins. This analysis of α-helices is a preliminary step of a more complete subsequent study, by VT, devoted to helices in proteins.

CONCLUSIONS

This study shows how the structure of proteins can be analyzed by means of VT. Relaxed random packing of spheres is used as a suitable model environment to take into account the solvent. The complete tessellation procedure is shown on the protein 3chy. Although only 39 proteins are used in this study, there is no obstacle to tessellate any known protein. The analysis of faces on the protein surface is in agreement with the previous research[21] and restates the hydrophobic/hydrophilic classification of AA from a pure geometrical viewpoint. It could be indispensable in the context of hydrophobic cluster analysis of proteins.[39] The analysis of internal (double) faces shows in a very simple way that the AA chain tends to form a local secondary structure, which certainly plays an important topological role. Further studies of the protein structure by VT could be considered in different directions. For instance, VT analysis could provide efficient tools in the attribution of secondary structures as in this case, the attribution is purely topological and does not need to use numerical values for angles or distances.

ACKNOWLEDGMENTS

B. A. acknowledges a visiting fellowship supported by Marie Curie Program of the European Union. A. S., J. C., and J-P. M. acknowledge support from "Programme Génome" (Centre National de la Recherche Scientifique, France).

REFERENCES

1. Pande VS, Grosberg A, Tanaka T. Heteropolymer freezing and design: towards physical models of protein folding. Rev Mod Phys 2000;72:259–314.
2. Brockwell DJ, Smith DA, Radford SE. Protein folding mechanisms: new methods and emerging ideas. Curr Opin Struct Biol 2000;10:16–25.
3. Dobson CM, Karplus M. The fundamental of protein folding: bringing together theory and experiment. Curr Opin Struct Biol 1999;9:92–101.
4. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. Principles of protein folding: a perspective from simple exact models. Protein Sci 1995;4:561–602.
5. Baldwin RL. The nature of protein folding pathways: the classical versus the new view. J Biomol NMR 1995;5:103–109.
6. Lazaridis T, Karplus M. "New view" of protein folding reconciled with the old through multiple unfolding simulations. Nature 1997;278:1928–1931.
7. Matagne A, Dobson C. The folding process of hen lysozyme: a perspective from the "new view." Cell Mol Life Sci 1998;54:363–371.
8. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways and the energy landscape of protein folding; a synthesis. Proteins 1995;21:167–195.
9. Tsai C-J, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels, and protein function. Protein Sci 1999;8:1181–1190.
10. Baker D. A surprising simplicity to protein folding. Nature 2000;405:39–42.
11. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
12. Berezovsky IN, Grosberg AY, Trifonov EI. Closed loops of nearly standard size: common basic element of protein structure. FEBS Lett 2000;466:283–286.
13. Lamarine M, Berezovsky I, Mornon JP. Chomilier J. Distribution of tightened end fragments of globular proteins statistically mathes that of topohydrophobic positions: towards an efficient punctuation of protein folding. CMLS 2001;58:492–498.
14. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
15. Finkelstein AV. Protein structure: what is it possible to predict now? Curr Opin Struct Biol 1997;7:60–71.
16. Sadoc JF, Mosseri R. Geometrical frustration. Cambridge, England: Cambridge University Press; 1999.
17. Sadoc JF, editor. Geometry in condensed matter physics. Singapore: World Scientific; 1990.
18. Sadoc JF, Rivier N, editors. Foams and emulsions. Dordrecht, The Netherlands: Kluwer Academic; 1999.
19. Jund P, Jullien R, editors. Physics of glasses. New York: American Institute of Physics; 1999.

20. Bideau D, Hansen JP, editors. Disorder and granular media. New York/Amsterdam: Elsevier; 1993.
21. Soyer A, Chomilier J, Mornon JP, Jullien R, Sadoc JF. Voronoi tessellation reveals the condensed matter character of folded proteins. Phys Rev Lett 2000;85:3532–3535.
22. Voronoi G. Recherches sur les paralleloedres primitifs. J Reine Angew Math 1908;134:198–287.
23. Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 1974;82:1–14.
24. Gerstein M, Tsai J, Levitt M. The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. J Mol Biol 1995;249:955–966.
25. Tsai J, Taylor R, Chothia C, Gerstein M. The packing density in proteins: standard radii and volumes. J Mol Biol 1999;290:253–266.
26. Singh RK, Tropsha A, Vaisman II. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol 1996;3:213–221.
27. Wako H, Yamato T. Novel method to detect a motif of local structures in different protein conformations. Protein Eng 1998;11: 981–990.
28. Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Sci 1997;6: 1467–1481.
29. Zimmer R, Wöhler M, Thiele R. New scoring schemes for protein fold recognition based on Voronoi contacts. Bioinformatics 1998;14: 295–308.
30. Bagci Z, Jernigan R, Bahar I. Residue packing in proteins: uniform distribution on a coarse-grained scale. J Chem Phys 2002;116:2269–2276.
31. Bagci Z, Jernigan RL, Bahar I. Residue coordination in proteins conforms to the closest packing of spheres. Polymer 2002;43;451–459.
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242. See also on the web site: http://beta.rcsb.org/pdb/.
33. Jullien R, Jund F, Caprion D. Computer investigation of long-range correlations and local order in random packings of spheres. Phys Rev E 1996;54:6035–6041.
34. Jodrey W, Tory E. Computer simulation of close random packing of equal spheres. Phys Rev A 1985;32:2347–2351.
35. Volz K, Matsumura P. Crystal structure of Escherichia Coli CHE*Y refined at 1.7-Å resolution. J Biol Chem 1991;266:15511–15519.
36. Cooper DW. Random-sequential-packing simulations in three dimensions for spheres. Phys Rev A 1988;38:522–524.
37. Branden C, Tooze J. Introduction to protein structure, 2nd ed. New York and London: Garland Publishing; 1999.
38. Liang J, Edelsbrunner H, Fu P, Sudhakar P, Subramaniam S. Analytical shape computation of macromolecules. I. Molecular area volume through alpha shape. Proteins 1998;33:1–17.
39. Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. CMLS Cell Mol Life Sci 1997;53:621–645.
40. Fersht AR. Transition state structure as a unifying basis in protein folding mechanisms: contact order, chain topology, stability and the extended nucleus mechanism. Proc Natl Acad Sci USA 2000;97:1525–1529.
41. Chiti F, Taddei N, White PM, Bucciantini M, Magherini M, Stefani M, Dobson CM. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. Nat Struct Biol 1989;6:1005–1009.
42. Berezovsky I, Trifonov EN. Van der Walls locks: loop-n-lock structure of globular proteins. J Mol Biol 2001;307:1419–1426.