Eur. Phys. J. B **33**, 355–363 (2003)
DOI: 10.1140/epjb/e2003-00176-5

THE EUROPEAN PHYSICAL JOURNAL B

The Laguerre polyhedral decomposition: application to protein folds

J.F. Sadoc[1,a], R. Jullien[2], and N. Rivier[3]

[1] Laboratoire de Physique des Solides, Université Paris Sud (associé au CNRS), bâtiment 510, Centre d'Orsay, 91405 Orsay, France
[2] Laboratoire des Verres, Université Montpellier 2 (associé au CNRS), Place E. Bataillon Case 003, 34095 Montpellier, France
[3] Laboratoire de Dynamique des Fluides Complexes, Université Louis Pasteur (associé au CNRS), 3 rue de l'Université, 67084 Strasbourg, France

Received 20 November 2002 / Received in final form 26 March 2003
Published online 20 June 2003 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2003

Abstract. An extension of the Voronoi tessellation, the Laguerre polyhedral decomposition, is introduced and applied to the analysis of the packing geometry of amino-acids in folded proteins. This method considers an ensemble of points with different weights and therefore it is well suited for a geometrical analysis of a set of objects with a wide size distribution. With this method it is shown that the true volumes occupied by the amino-acids inside a protein is better described than with the standard Voronoi procedure. This method allows defining unambiguously (without cut-off distance) the neighborhood for each amino-acid in a given protein and contact matrices can be established which contain all topological informations on the internal structure. Finally, a statistical analysis of the geometrical characteristics of the polyhedra attached to each amino-acid is done over a collection of 35 proteins.

PACS. 87.10.+e General theory and mathematical aspects – 36.20.-r Macromolecules and polymer molecules

1 Introduction

Understanding the folding of an amino acid chain into a well defined globular protein structure is a challenge. Tremendous amount of experimental work has been done in the fields of molecular biology, biochemistry and biological physics to try to understand this complex phenomenon. A folded protein is a very compact and reproducible packing of different amino-acid residues (AA) with lateral chains attached to a backbone chain of peptide bonds *via* α-carbon atoms [1]. The knowledge of the precise structure of a folded protein is very essential. Up to now the only efficient tools are X-rays which requires crystallized samples and NMR. Quite a lot of structures are presently known [2], at least enough to recognize some general trends but, anyway, they represent only a few fraction of the total number of proteins.

What still remains unclear is how to predict the native state structure of a protein from its sequence (*i.e,* given the ordered list of AAs along the chain). One possible way to overcome this failure of predictability of the molecular structure is to analyze the energy landscape, as currently done by many authors [3]. But another complementary approach is to examine in details all the informations that come from pure geometry of the known protein structures. In the field of liquids, liquid crystals, crystalline and amorphous solids, such a geometrical approach was very fruitful [4].

Several authors have emphasised the importance of geometry and topology in protein folds (structure and folding). In his review, Baker [5] illustrates the role of topology in folding mechanism. In this paper he quotes several analyses, which make the same point [6] and [7]. Other authors [8] have proposed to consider a protein as a packing of a tube (garden hose), surrounding its backbone. They use this approach to predict possible phase transitions between different states of compactness. The importance of the backbone has also been emphasised by the result of Tifonov *et al.* [9], of a larger probability of contact for two AA, 27 steps apart along the polypeptide chain. In a more mathematical approach, two of us have focused on the relation between packing, chains, helices and their chirality [10]. Thus, various new geometrical and topological concepts are now introduced in protein studies.

To analyze the structure of folded proteins, we have recently proposed a very sensitive geometrical approach based on the so-called Voronoi tessellation (VT) [11]. A tessellation is a way to describe the filling of space by packing solid polyhedra connected by common faces without empty space between them. Given a set of discrete points in space, the Voronoi tessellation consists in associating to each point a polyhedral domain, called the Voronoi cell

[a] e-mail: sadoc@lps.u-psud.fr

(VC), containing the entire neighborhood closer to a given point than all the others. The cell characteristics provide essential information on the local geometry of the considered packing. This tool is widely and currently used to study random sphere packings, granular materials, foams, froths or glasses [12].

There are several examples of VT methods applied to proteins in the literature [13] but only a few of them concern directly the packing of amino-acids. In order to associate a site to each AA the α-carbons could be chosen as the starting set of points, but we have preferred, in our recent study [11], to consider the geometrical centers of individual AAs. With this choice the VCs are representing topologically better the true volumes occupied by the AAs and lead to a more homogeneous distribution of distances. As a result the lateral chains are well located inside their cells.

In order to build the VC associated with a given AA, it is necessary to have a precise knowledge of its neighborhood. This is easy for AAs which are located well inside a dense region of the bulk of the protein. In that case it is enough to know the positions of its neighboring AA centers. But for an AA located close to the external surface or in a cavity, this becomes more difficult as, in principle, a detailed knowledge of the nature and location of the surrounding molecules is needed. We have recently proposed [14] to resolve this difficulty by surrounding the protein with a model of solvent, that we will often call "environment" in this paper, whose characteristics are similar to generic proteins considered as random dense packing of equal sized spheres of average AA volume.

In our preceding works [11,14] we have always used a standard VT method based on a previous so-called "Delaunay tetrahedrization" which treats all points with the same weight and which is known to work quite well for components of almost the same size [12]. However for systems containing small and large components, like SiO_2 glasses for instance, with such a standard method the VCs for small atoms appear too large and, reciprocally, the VCs for large atoms appear too small and therefore the VCs do not give a physically correct image of the volumes occupied by the atoms. This is a trivial consequence of the standard method in which the cell faces are located at equal distance of closer atoms. In the case of proteins, considered as a packing of AAs, it is known that the size distribution of the AA volumes is quite wide as there is a very large difference of volume between the smallest AA (Gly) and the largest one (Trp) and it is worth improving the method to take care of this. In the literature, it exists another cellular decomposition, called "Laguerre froth" [15,16], which can be viewed as a weighted Voronoi decomposition in which the VC faces remain planar, but no longer equidistant to the two points that separates: it is closer to the point with smaller weight. This method has been nicely applied to polyhedral patterns in physics, metallurgy and applied mathematics, by a group in Lausanne [15], and to the study of foams by one of us [16]. In this paper we intend to present this Laguerre decomposition method in order to apply it to globular proteins.

2 Laguerre polyhedral decomposition

2.1 Power of a point with respect to a weighted point

Let us define the power of a point $\mathbf{x}$ (of Cartesian coordinates x_1, x_2, x_3) with respect to another point (site) $\mathbf{s}$ (of Cartesian coordinates s_1, s_2, s_3) of weight $w(\mathbf{s})$ by:

$$p(\mathbf{x},\mathbf{s}) = d^2(\mathbf{x},\mathbf{s}) - w(\mathbf{s}) \qquad (1)$$

where $d^2 = (x_1-s_1)^2+(x_2-s_2)^2+(x_3-s_3)^2$ is the square of the Cartesian distance between $\mathbf{x}$ and $\mathbf{s}$.

This is the same as the power of a point relatively to a sphere of center $\mathbf{s}$ and radius ρ when one sets $w = \rho^2$, but the above definition can be generalized to negative weights. Recall that the power of a point relatively to a sphere is the square of its tangent to the sphere, when it is located outside the sphere.

Consider two points $\mathbf{s}_1$ and $\mathbf{s}_2$ with weights w_1 and w_2, it can be easily shown that the set of points with equal power relatively to $\mathbf{s}_1$ and $\mathbf{s}_2$ is a plane perpendicular to the straight line joining $\mathbf{s}_1$ and $\mathbf{s}_2$. This generalizes, to the case of a set of weighted points, the standard Voronoi construction where the VC face is the median plane equidistant to $\mathbf{s}_1$ and $\mathbf{s}_2$.

2.2 Weighted Delaunay tetrahedral decomposition

Considering a set of sites in space, a first step prior to a Voronoi tessellation consists in performing a Delaunay tetrahedral decomposition. This is done by determining the "simplicial tetrahedra" formed with four sites such that any other point of the set cannot fall inside the sphere passing through these four sites (we assume that, in the random systems considered here, four points are never in the same plane, so that such a sphere always exists). In practice, to determine the simplicial tetrahedra, we consider all sets of four points and, for each set, we make a loop over all the other points to test if they are inside the sphere. The test for a point $\mathbf{x}$ to be inside the sphere defined by the positively oriented tetrahedron $(\mathbf{r},\mathbf{s},\mathbf{t},\mathbf{u})$ can be performed by considering the following determinant involving the coordinates:

$$\delta = \begin{vmatrix} x_1 & x_2 & x_3 & x_1^2+x_2^2+x_3^2 & 1 \\ r_1 & r_2 & r_3 & r_1^2+r_2^2+r_3^2 & 1 \\ s_1 & s_2 & s_3 & s_1^2+s_2^2+s_3^2 & 1 \\ t_1 & t_2 & t_3 & t_1^2+t_2^2+t_3^2 & 1 \\ u_1 & u_2 & u_3 & u_1^2+u_2^2+u_3^2 & 1 \end{vmatrix}. \qquad (2)$$

If δ is positive (resp. negative) the point $\mathbf{x}$ is inside (resp. outside) the sphere.

This procedure can be straightforwardly extended to the weighted Delaunay tetrahedrization, by considering instead the following determinant in which the weights $(w_x, w_r, w_s, w_t, w_u)$ of the five points $(\mathbf{x},\mathbf{r},\mathbf{s},\mathbf{t},\mathbf{u})$ appear:

$$\delta_w = \begin{vmatrix} x_1 & x_2 & x_3 & x_1^2+x_2^2+x_3^2-w_x & 1 \\ r_1 & r_2 & r_3 & r_1^2+r_2^2+r_3^2-w_r & 1 \\ s_1 & s_2 & s_3 & s_1^2+s_2^2+s_3^2-w_s & 1 \\ t_1 & t_2 & t_3 & t_1^2+t_2^2+t_3^2-w_t & 1 \\ u_1 & u_2 & u_3 & u_1^2+u_2^2+u_3^2-w_u & 1 \end{vmatrix}. \qquad (3)$$

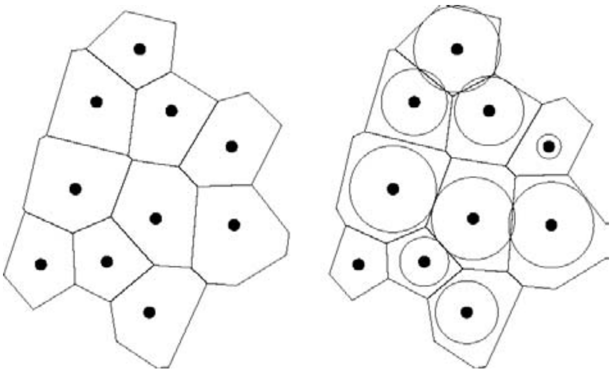

Fig. 1. Left: standard Voronoï decomposition in two dimensions, all points have the same weights. Right: Laguerre decomposition for the same set of points but with different weights represented by the circles whose radii are the square roots of the weights. Note that points with smaller (larger) weights have smaller (larger) cells compare to the standard Voronoi decomposition.

In the standard case, a Voronoi tessellation can easily be obtained from a Delaunay tetrahedrization by duality: the vertices of the VC are the centers of circumsphere of the simplicial tetrahedra. The equation $\delta(\mathbf{x}) = 0$ is the equation of the circumsphere and consequently the coordinates of its center (equidistant to the four vertices of the corresponding tetrahedron) can easily be obtained. The vertices of the Laguerre polyhedra can be obtained in a similar way, replacing δ by δ_w. Now such a vertex is no more equidistant to the four vertices of the corresponding simplicial tetrahedra but, it has the same power with respect to these four weighted points. (It is at equal "tangential distances" to these four points, when it lies outside them.)

There is no difficulty to implement an algorithm based on these results. Figure 1 shows a 2D example in which Voronoi and Laguerre decompositions are contrasted.

2.3 Properties of Laguerre polyhedral decomposition

The Laguerre decomposition depends on the distribution of weights. Of course, for a given set of points, if all the w_i's are zero the Laguerre decomposition is identical to the standard Voronoi decomposition, but this is also true if all the w_i's have the same value. More generally, the decomposition obtained with a given set of $\{w_i\}$ is identical to the one obtained with $\{w_i+c\}$ where c is any positive or negative constant. In practice, it is always possible to add a sufficiently large constant to have all weights positive and to represent them by (mostly overlapping) spheres of radii $r_i = \sqrt{w_i}$. Conversely, it is clear that the set $\{cw_i\}$ would not give the same decomposition: this is important as it means that there is a local scale for the effect of the weights related to the mean distances between sites.

Another tricky property is typical of Laguerre decomposition and cannot occur in the standard Voronoi decomposition: in some particular case it may happen that a site is located outside its own cell. This happens if a sphere (1) of weight w_1 is at a distance d from a site (2) of weight w_2 such that $d^2 < w_2 - w_1$. In that case the face associated with these two sites cut the line (1-2) on a point located *outside* the segment (1-2), on the side of site (1), and consequently the site (1), located on the wrong side of the face is pushed outside its own cell. Such a pathological situation should be considered as unphysical and if it happens for some sites of a given set this would mean that either the site locations or/and the weights have not been correctly chosen.

3 Proteins as polyhedral packings of amino-acids

3.1 Proteins and their environment

A folded protein is a very compact packing of different AA residues which can be represented by the set of their geometrical centers. We have considered the known structures of 35 globular folded proteins chosen to belong to the different folds and representative of the different classes. For each protein we have determined the geometrical centers of all the AAs from which it is made [11]. The number N_{AA} of AAs in a protein is not very large: from $N_{AA} \simeq 60$ for small proteins to $N_{AA} \simeq 600$ for the largest ones, but for most of them the number of N_{AA} lies between 100 and 300. Thus, about two third of the AAs lie on the outer surface of the fold. There is a problem in defining the Voronoi cells for these AAs since they have no neighbors outside. If one restricts the polyhedral decomposition to the set of AAs only, the cells on the surface cannot be closed or they are dramatically elongated.

In our first work [11] we have used an ad hoc method to define and eliminate amino-acids at the surface. Here we have preferred to use a more physically justified method that we have introduced in a more recent paper [14] and which consists in adding an artificial environment around the protein. This environment is chosen to have packing properties close to those of AAs in a protein. It can be considered as a solvent made of molecules similar to a generic AA. In practice we have considered random packing of spheres generated with an algorithm proposed by one of us [17] after improving the well-known Jodrey-Tory algorithm [18].

In order to simulate the solvent, the environment is spread around the protein as a shell of constant thickness. If there were a cavity inside the protein, it would also be filled with the environment. The volume of an individual sphere used to simulate the solvent is taken close the mean volume of the AAs. In practice we have considered a radius of 6.5 Å and the total number of points, *i.e.* N_{AA} plus the number solvent spheres was kept to be close to $10N_{AA}$ to insure closed cells for all AAs.

In practice we have built a large sphere packing and we have immersed the protein inside it. Then we have removed the environmental spheres superimposed to AA spheres. Doing so it results some discrepancy between the environment and the protein at its surface. To resolve this mismatch we have used an iterative relaxation of the environment to obtain the best possible tessellation. We proceed by considering the protein and its environment as a random dense ensemble of spheres. During the relaxation, the spheres representing the AAs are kept fixed and only the centers of the spheres of the environment are moved, assuming a constant volume. At each iteration step, the Voronoi cells are built and the sphere centers are shifted towards the geometrical centers of the VC. This relaxation allows regularizing the shape of the cells of the environment: elongated cells become more isotropic with a narrower distribution for their volumes. After about nine iterations the environment centers do not move anymore and the relaxation is deemed to be achieved.

3.2 Volumes of the amino-acids

In our previous works [11,14], a linear relation was observed between the volume of an AA and the volume of its corresponding VC: the larger (smaller) AAs have larger (smaller) cells. But owing to the large disparity of the volumes, the slope of this linear dependence is certainly underestimated, since a regular Voronoi decomposition overestimates the size of the smaller AAs and *vice versa.* One reason for using the Laguerre decomposition is that it yields more realistic volumes. However, to perform a Laguerre decomposition, we need to know the weights attributed to each AA. In practice we have chosen the weights which give volumes for the AAs close to the ones already known.

Volumes of AAs have been estimated by different authors [19,20].

Here we have used the more recent work of Pontius *et al.* [21], which is based on a Voronoi decomposition based on the individual atoms of the protein. Each cell corresponds then to an atom, and the volume of an AA is the sum of the volumes of the cells of its constituting atoms. The volumes that we have used are listed in Table 1. We call them Pontius volumes and note them v_P in the following.

Since the atoms involved, C, N, O, H and S, have roughly the same size, a Laguerre decomposition for the individual atoms is not necessary. As seen in table I, the Pontius volume of each AA depends on the architecture of its side-group; it is not simply proportional to the number of its constituting atoms. (Compare Ile 19 ($= 5 + 1 + 13$) with His 17 ($= 5 + 1 + 11$) which has nearly the same Pontius volumes, or the different volumes of Arg and Trp, which have both 24 ($= 5 + 1 + 18$) atoms.)

Table 1. The list of the amino-acids with their corresponding volumes v_P in Å^3 from Pontius work [21]. The standard deviations σ for the v_P estimates are also listed. We have distinguished two Cystines, Cyss and Cysh depending whether they are involved in a disulfide bridge or not.

AA	v_P	σ
Ala	91.5	5.32
Arg	196.1	4.25
Asn	138.3	5.60
Asp	135.2	5.21
Cysh	114.4	6.64
Cyss	102.4	6.13
Gln	156.4	4.32
Glu	154.6	5.75
Gly	67.5	5.72
His	163.2	4.04
Ile	162.6	3.64
Leu	163.4	4.19
Lys	162.5	2.24
Met	165.9	5.31
Phe	198.8	3.99
Pro	123.4	6.26
Ser	102.0	6.72
Thr	126.0	4.92
Trp	237.2	3.65
Tyr	209.8	5.29
Val	138.4	3.87

3.3 How to choose the weights for the amino-acids

We want to choose the weights in order to have cell volumes close to Pontius volumes. This is not trivial as the decomposition is invariant under uniform translation of the weights. To determine the best weights for the amino-acids we have used an iterative procedure in order to obtain the volumes of the cells proportional to their corresponding Pontius volumes. Initially, we set all the weights (for the AAs as well as for the spheres of the environment) arbitrarily equal to zero (leading to a Voronoi decomposition). Then knowing all the weights w_i at some step of the iteration, we perform a Laguerre decomposition and we calculate the volumes V_i of the cells for all the AAs (the label i refers to the position of the AA along the peptide chain). To proceed to the next iteration, the weights of the AAs are modified according to the formula:

$$w'_i = w_i + K(u\mathrm{v}_{Pi} - V_i) \tag{4}$$

where v_{Pi} is the Pontius volume of the AA labelled i, u is calculated by

$$u = \frac{\sum_i V_i}{\sum_i \mathrm{v}_{Pi}} \tag{5}$$

(the sums run over the AAs and do not consider the environment) and the constant K is chosen positive and sufficiently small to insure a good convergence for this iterative process. The weights of the surrounding spheres (environment) are not modified during this iterative scheme. A typical example of such a calculation is depicted in Figure 2 in which we have used Myohemerythrin (whose PDB code is 2mhr), a fairly small protein containing 118 AAs. On this figure the mean cell volume $\langle V \rangle$ for each AA (average made over the same AAs of the protein) has been plotted as a function of the Pontius volume before starting the iterations (open circle) and after 330 iterations (full

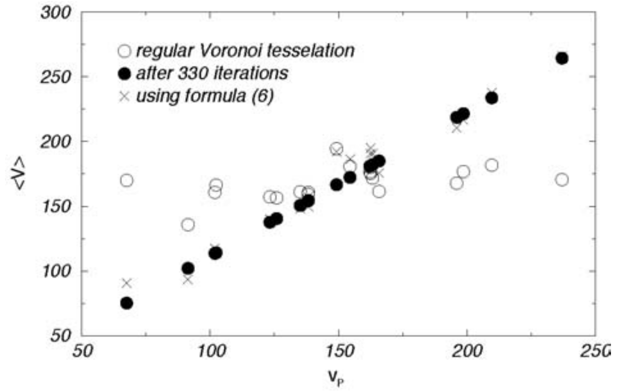

Fig. 2. Mean volume $\langle V \rangle$ for the cell of a given protein as a function of its Pontius volume v_P in the case of the protein 2mhr. The open circles correspond to the results obtained with the regular Voronoi tessellation, the full circles correspond to the results obtained after 330 iterations with the iterative Laguerre scheme explained in the text and the crosses correspond to the results obtained with a unique Laguerre decomposition using formula (6) for the weights.

Table 2. The list of the proteins on which we have applied an iterative scheme to get a mean cell volume proportional to the Pontius volume for each AA together with their numbers of AAs and the numerical results for the ratio of proportionality.

Protein	N_{AA}	u
1utg	70	1.125
2mhr	118	1.116
3chy	128	1.237
1hvc	203	1.145
1phm	305	1.222
1hle	362	1.080

circles), using $K = 0.01$. At the end of this iteration process $\langle V \rangle$ becomes proportional to v_P within less than 10^{-5} error. The slope is slightly larger than, but roughly equal to 1. This should be compared to the very small slope of about 0.2 obtained with the regular Voronoi tessellation method (open circles in Fig. 2).

The same iterative scheme has been applied successfully to some other proteins of various sizes and the results for the slope u have been reported in Table 2.

In practice we have observed that this slope can vary slightly with the protein and, for a given protein, with the choice of the environment but it is always of order 1.12 ± 0.12, a quite reasonable result.

At the end of the iteration scheme the weights converge to asymptotic values which has been plotted as a function of $v_P^{2/3}$ in Figure 3 for all the proteins listed in Table 2. One observes a quite nice common linearity with error bars which are perfectly consistent with the standard deviations listed in Table 1 for the v_P values.

A linear regression made through all these data gives:

$$w = -14.325 + 0.5282 v_P^{2/3} \tag{6}$$

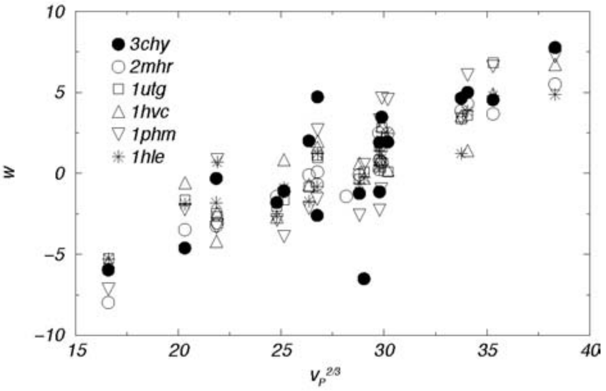

Fig. 3. Plot of the weight w as a function of $v_P^{2/3}$ obtained at the end of the iteration scheme applied to the six proteins listed in Table 2.

where w and v_P are expressed in Å^2 and Å^3, respectively. This is the relation that we have systematically used and injected into the Laguerre polyhedral decomposition method to get most of the results presented below.

To check the self-consistency of this choice we have indicated by crosses in Figure 2 the mean volume $\langle V \rangle$ as a function of the Pontius volume v_P for the 2mhr protein when using equation (6) for the weights of the AAs. Of course we now obtain some dispersion for the cell volumes but the general linear dependence with the Pontius volume is perfectly well recovered.

4 Contact matrices

4.1 Definition and methods of calculation

The arrangement of amino-acids in a given protein can usefully be represented by a contact matrix. This is a symmetric N_{AA}^2 matrix α_{ij} with element $\alpha_{ij} = 1$ if AA i is in contact with the AA j, $= 0$ otherwise. The contact matrix has already been introduced by several authors, which assert that two AAs are in contact if their distance is smaller than a given cut-off distance [22–25]. (Usually, the distance between C_α atoms are used, but if one views proteins as a packing of AAs, it seems preferable to use the distance between the geometrical AA centers.) The Voronoi and Laguerre decomposition define neighborhood unambiguously, without relying to any cut-off distance. This is why we built the contact matrix by stating that two AAs are in contact if their cells share a face.

From a mathematical point of view, contact matrices obtained using Voronoi or Laguerre decompositions are 2D representations that define the topological 3D structure uniquely (a unique topological graph). In physical language, the entropy of all possible 3D structures described by one contact matrix is zero, or very small. There remains to ascribe specific amino-acids to the position i along the polypeptide chain. But the combinatorics are severely limited, geometrically because the volume of the amino-acids must fit in the hole left by its contacts, and chemically for the fold to be realisable.

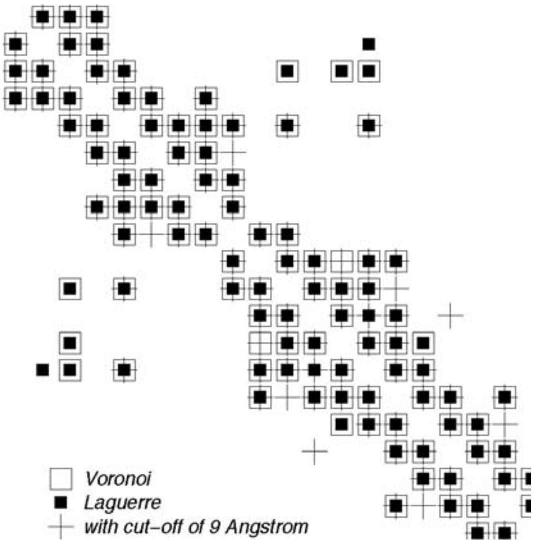

Fig. 4. 20 × 20 top-left corner of the contact matrix for the 2mhr protein. The non-zero elements are represented by open squares, filled squares and crosses if the contacts are defined by the Voronoi decomposition, the Laguerre decomposition and by using a cut-off distance of 9 Å respectively.

As an illustration of the differences obtained by using these methods we have enlarged the top left corner of the contact matrix for the 2mhr protein (Fig. 4). One can observe differences in several places. The differences are largest between the cut-off and the tessellation methods. Here we have chosen the cut-off distance in order to obtain roughly the same total number of non-zero elements and one can see that on several places contacts obtained with the tessellation methods are not recovered (especially among the elements far from the diagonal) and some other artificial contacts are added (especially on the border of the diagonal stripe). This can be understood since by using a cut-off distance, one can miss contacts between two large AAs and introduce artificial contact between two small AAs. Differences between the Voronoi and Laguerre methods are less frequent (they often occur when a small AA is close to a large one) but we think they can be relevant and in the following we will report two examples in which we have used the Laguerre method.

4.2 Two typical examples

Since geometry and metric are highly correlated in a dense packing, the contact matrix encodes most of the geometry of a protein. Nevertheless we should add information on the chirality of proteins because two proteins with the same chemical composition but with opposite chiralities have the same contact matrices. We show how to read folds in contact matrices on the two examples below. The first one is Myohemerythrin (2mhr), already introduced above. This is a globular protein made of four helices. The second one is the signal transduction protein chey from escherichia coli (3chy). This is a protein formed from a β-sheet sandwiched between five α helices. The two matrices are shown in Figures 5 and 6. On these two matrices some secondary structures show up on the border of the matrix diagonal, as a rather broad stripe for an α-helix and a narrow one for a β-strand. This can be understood since, in an α-helix, the AA number i is always connected to the AAs $i\pm1$, $i\pm3$, $i\pm4$ and sometimes to the AAs $i\pm2$. In a β-strand, the AA number i is always connected to AAs $i \pm 2$ only. Coils are less visible. It is important to remark that other contacts are not at all randomly spread within the matrix but rather concentrated along lines either perpendicular or parallel to the diagonal. A line perpendicular to the diagonal appear if two secondary structures are antiparallel. Consider the two α-helices in 1mhr (which is roughly the half of 2mhr). They run from Ala 41 to Ala 64 and from Val 71 to Ile 84. They are linked together by a coil between 67 and 70. As two helices are refolded one onto the other, it occurs very often that if AAs i and j

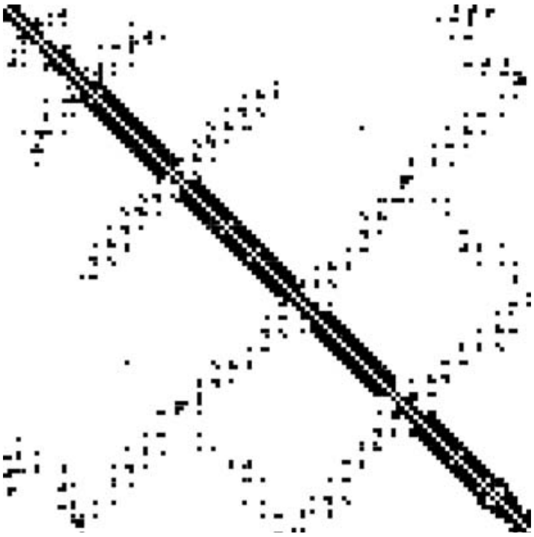

Fig. 5. Contact matrix (118 × 118) for the 2mhr protein.

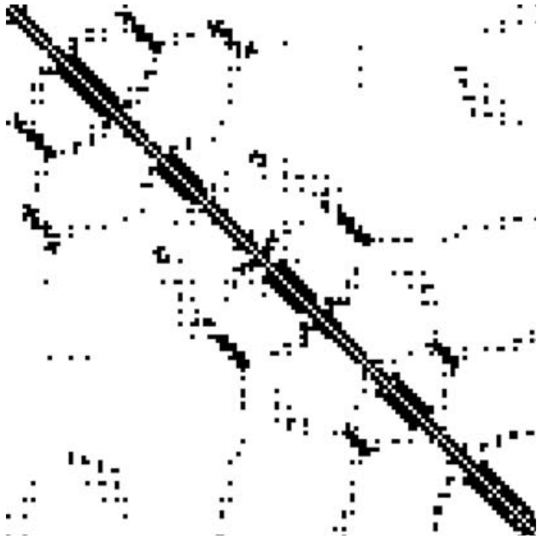

Fig. 6. Contact matrix (128 × 128) for the 3chy protein.

are neighbors, then several AAs $i + n$ and $j - n$ are also neighbors leading to non-zero matrix elements lying on a segment perpendicular to the diagonal. Moreover, if the chain folds again onto two already folded parts, this may lead to a concentration of contacts along a segment parallel to the diagonal. A lot of these segments can be seen in the case of 3chy.

Another intriguing observation can be made: in both cases there is a concentration of contacts in the bottom-left and top-right corners. This means that the two ends of the protein chain are close together. This situation, which occurs very often in protein chains, is worth investigating.

5 Cell statistics

5.1 Overal statistics

As already noticed in our previous work [11] two quantities of interest that give an idea of the overall geometry of the packings of AAs in proteins are the mean number of faces per cell $\langle f \rangle$, which is the mean coordination number between AAs, and the mean number of edges per faces $\langle e \rangle$, which is related to the local symmetry around bonds between neighboring AAs. Using the Laguerre method and averaging over the 35 proteins we find $\langle e \rangle = 5.15$ and $\langle f \rangle = 14.17$. The slight differences with the values reported in reference [11] comes more from the realistic consideration of the surface than from the difference from Voronoi and Laguerre methods.

More significant are the differences with values reported in reference [14] which were obtained exactly under the same conditions, but with the Voronoi method. In [14] $\langle f \rangle = 14.27 > 14.17$ was obtained. Theoretical works (see [4] for references) on polydisperse sphere packings show that large fluctuations in cell sizes decrease the coordination number $\langle f \rangle$. By contrast, cell shape anisotropy, increases $\langle f \rangle$. Since disorder cause both cell fluctuations and disorder, small values of $\langle f \rangle$ are usually observed only in crystalline structures. Its observation here indicates that the Laguerre cells for AAs in proteins are not too anisotropic and adjusted to some extend. Anyhow these values are still close to those of compact structures often encountered in condensed matter physics. A value of $\langle e \rangle \simeq 5$ is characteristics of compact structures built using local rules for packing, like hard sphere random close packings. Such strong five fold local symmetry cannot be extended too far in the regular three dimensional space due to geometrical frustration [4].

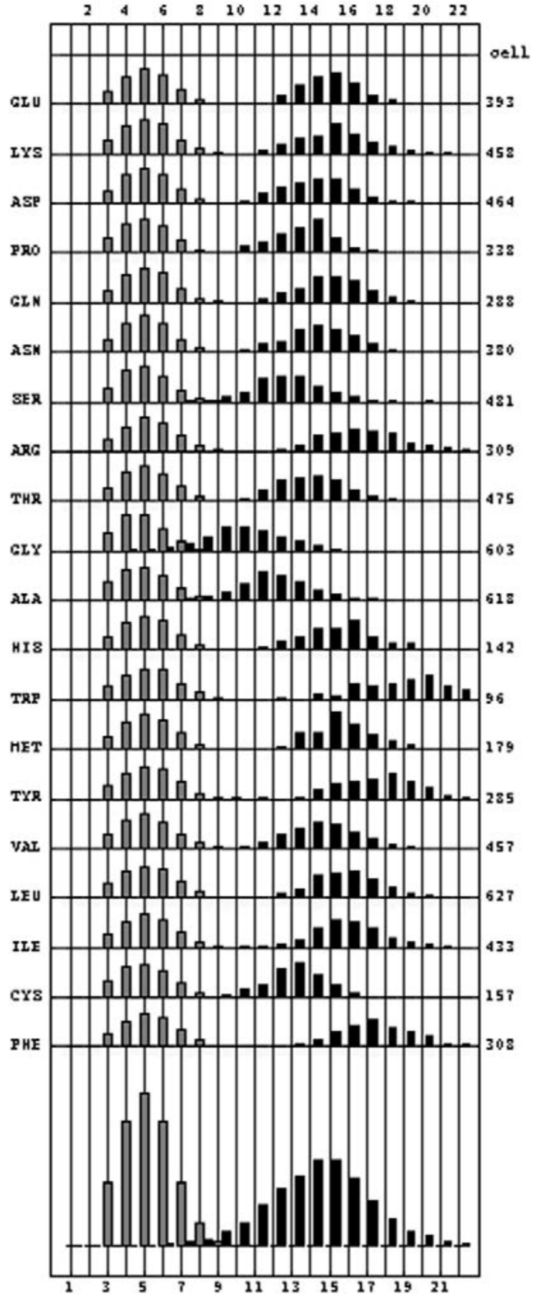

Fig. 7. Histograms $h(e)$, in grey, and $h(f)$, in black (fraction of faces with e edges and fraction of cells with f faces) for each amino-acid. The numbers indicated on the right are the numbers of cells involved in the statistics. The amino-acids are classified from hydrophilic (top) to hydrophobic (bottom). The full histograms (regardless the amino-acid) are given at the bottom of the figure.

5.2 Statistics for each amino-acid

In the spirit of finding *ab initio* methods to obtain the protein structure knowing the sequence of its AAs, it is very useful to perform cell statistics separately for each amino-acid to try to recognize them from the geometry of their cells. In Figure 7 one provides the histograms $h(e)$ and $h(f)$, fraction of faces with e edges and fraction of cells with f faces, for the twenty amino-acids (here Cyss and Cysh are not distinguished). There are particular faces which define the surface of the protein, *i.e.* those at the interface between an AA and a sphere of the environment. Figure 8 gives the histograms $h_S(e)$ for the number of edges of surface faces, and $h_S(f)$ for the number of surface faces per surface cells. The full histograms (regardless the AA) are given at the bottom of the figure.

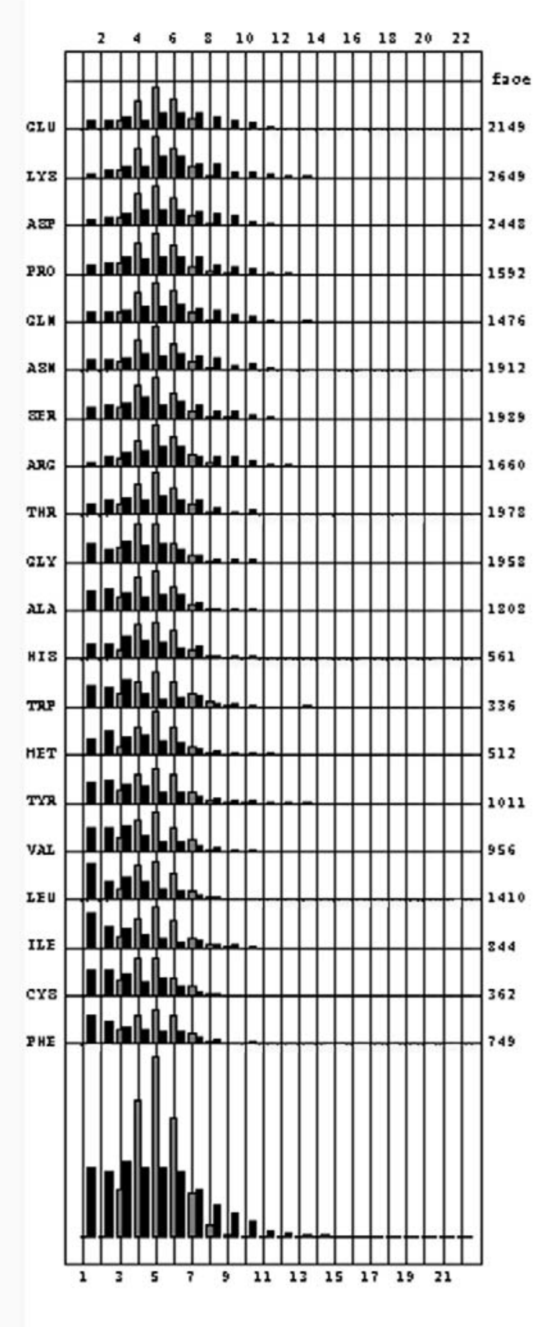

Fig. 8. Histograms $h_S(e)$ for faces corresponding to the surface of the protein, in grey, and $h_S(f)$, in black, of the number of surface faces for the cells at the surface of the protein. Same as in Figure 7 save for the numbers on the right which are now the numbers of faces involved in the statistics.

In these figures it appears clearly that large AAs have a larger number of faces than small AAs. Furthermore, for surface cells, hydrophilic AAs, have a larger number of surface faces (5.75 for Glu) than hydrophobic AAs (3.13 for Leu).

The surface faces are in contact with the environment. Their statistics are especially interesting. The mean number of edges of surface faces is almost exactly $\langle e_s \rangle = 5.00$. Any finite cell packing, extracted from a disordered infinite packing should have the same statistic for its internal faces and for its surface faces, thus $\langle e_s \rangle = \langle e \rangle \simeq 5.15$. This is not the case here, indicating again some topological order in the AA packing.

There is another point appearing on these histograms which remains unexplained, but which could be fruitful in fold prediction, as it seems related to some of the AAs only. The width of the $h(f)$ distribution is very different from one AA to another; it is neither correlated to its size, nor to its hydrophobicity, as can be seen by comparing Glu and Lys or Arg and His, for instance. This probably indicates that some AAs have a local neighborhood more uniform than others.

5.3 Proteins versus random close packed structure

At first sight, if one considers only global statistics for the number of edges per faces or for the number of faces per cells, a protein seems very similar to random close packing. If less constrained than hard sphere packing, it can be viewed as some kind of froth with cells of different sizes. Nevertheless there are interesting details which indicate that a protein has a backbone.

In a previous study [14] analysing Voronoi cells in proteins, two of us have observed that faces shared by two successive AA along the chain have, on average, a larger number of edges. Laguerre cells exhibit the same effect, with a number of edges for these faces $\langle e_{chain} \rangle \simeq 6.5$. For any face, the average number of edges per face is close to $\langle e \rangle = 5.1$ for isotropic cells of equal sizes. This implies that interfaces other than those separating two successive AA along the chain, are smaller than average: they have approximately 5 edges per face. So we can view a protein as a chain of closely packed cells, slightly compressed along the chain, thus making a kind of tube tiled by faces with an average number of edges close to 5. This can be seen as a justification of the "tube" model mentioned in the introduction [8].

6 Conclusion

In this paper we have presented a numerical method, the Laguerre polyhedral decomposition, a very useful tool to investigate the packing geometry of objects with a wide size polydispersity, and we have applied it to the study of the arrangement of amino-acid residues in protein folds. Not only this method shares with the traditional Voronoi method the advantage of defining the neighborhood of amino-acids in absolute way, without introducing any artificial cut-off, but also, by introducing some weights depending on the volumes of the amino-acids, it permits to better treat very small and very large amino-acids, especially when they are in contact.

What is the potential of our method? The ultimate aim is to obtain a typical cell for each AA. This could be the case if all AAs of a given type always have cells with the same number of faces and with faces of the same type. This is probably too ambitious, but fluctuations in the types of cell faces could be limited. According to topological properties of cell packings, by displacing slightly their centers one can obtain faces with 4, 5 and 6 edges for almost all the cells. This being achieved, all AAs of a given type would have some kind of signature associate with the number of such faces. So, it would be possible to construct a contact matrix *a priori* from the sequence and obtain the topological folding.

There are other possibilities that appear more accessible with this method. For instance, structures obtained from X-ray diffraction or NMR analysis could have slight imperfections. Today, their detection and correction are time consuming. An algorithm based on Voronoi or Laguerre decomposition could be very efficient for this purpose.

Very often the biological activity of a protein is associated with configuration changes. These changes could affect only a few numbers of cells, even if they correspond to a noticeable modification of the AA coordinates. The comparison of the contact matrices obtained by Voronoi or Laguerre decomposition reveals to be a very sensitive method to follow such modifications. The Voronoi and Laguerre decomposition can be used also in secondary structure attributions [26].

References

1. C. Chotia, Proc. Natl. Acad. Sci., USA **74**, 4130 (1977)
2. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acid Research **28**, 235 (2000); see also the Protein Data Bank on the web site `http://www.rcsb.org/pdb/`
3. R.L. Baldwin, J. Biomol. NMR **5**, 103 (1995); T. Lazaridis, M. Karplus, Nature **278**, 1928 (1997); A. Matagne, C. Dobson, Cell. Mol. Life Sci. **54**, 363 (1998)
4. J.F. Sadoc, R. Mosseri, in *Geometrical Frustration* (Cambridge Univ. Press, 1999)
5. D. Baker, Nature **405**, 39 (2000)
6. F. Chiti, N. Taddei, P.M. White, M. Bucciantini, F. Magherini, M. Stefani, C.M. Dobson, Nature Struct. Biol. **6**, 1005 (1999)
7. C. Micheletti, J.R. Banavar, A. Maritan, F. Seno, Phys. Rev. Lett. **82**, 3372 (1999)
8. J.R. Banavar, A. Maritan, Rev. Mod. Phys. **75**, 23 (2003)
9. E.N. Trifonov, I. Berezovsky, Molecular Biology **36**, 239 (2002)
10. J.F. Sadoc, N. Rivier, Eur. Phys. J. B **12**, 309 (1999)
11. A. Soyer, J. Chomilier, J. P. Mornon, R. Jullien, J.F. Sadoc, Phys. Rev. Lett. **85**, 3532 (2000)
12. see for example *Foams and Emulsions*, edited by J.F. Sadoc, N. Rivier (Kluwer Academic, 1999); *Physics of Glasses*, edited by P. Jund, R. Jullien (American Institute of Physics, 1999); *Disorder and Granular Media*, edited by D. Bideau, J.P. Hansen (Elsevier, 1993)
13. F.M. Richards, J. Mol. Biol. **82**, 1 (1974); M. Gerstein, J. Tsai, M. Levitt, J. Mol. Biol. **249**, 955 (1995); J. Tsai, R. Taylor, C. Chothia, M. Gerstein, J. Mol. Biol. **290**, 253 (1999); R.K. Singh, A. Tropsha, I. Vaisman, J. Comput. Biol. **3**, 213 (1996); H. Wako, T. Yamoto, Protein Eng. **11**, 981 (1998); P.J. Munson, R.K. Singh, Protein Sci. **6**, 1467 (1997); R. Zimmer, M. Wehler, R. Thiele, Bioinformatics **14**, 295 (1998)
14. B. Angelov, J.F. Sadoc, R. Jullien, A. Soyer, J.P. Mornon, J. Chomilier, Proteins: structure, function and genetics **49**, 446 (2002)
15. H. Telley, T.M. Liebling, A. Mocelin, Phil. Mag. B **73**, 395 and *ibid.* **73**, 409 (1996)
16. N. Rivier, J. Phys. Colloq. France **23**, C7-309 (1990)
17. R. Jullien, P. Jund, D. Caprion, D. Quitmann, Phys. Rev. E **54**, 6035 (1996)
18. W. Jodrey, E. Tory, Phys. Rev. A **32**, 2347 (1985)
19. C. Chothia, Nature **254**, 304 (1975)
20. Y. Harpaz, M. Gerstein, C. Chothia, Proteins **2**, 611 (1994)
21. J. Pontius, J. Richelle, S.J. Wodak, J. Mol. Biol. **264**, 121 (1996)
22. S. Saitô, T. Nakai, K. Nishikawa, Proteins **15**, 191 (1993)
23. N. Gö, Nature **291**, 90 (1982)
24. J. Selbig, Protein Engineering **8**, 339 (1995)
25. N. Saitô, Y. Kobayashi, Int. J. Mod. Phys. B **13**, 2431 (1999)
26. F. Dupuis, J.P. Mornon, J.F. Sadoc, to appear in Proteins (2003)