Voronoï Tessellation Reveals the Condensed Matter Character of Folded Proteins

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

[1]*Laboratoire de Minéralogie Cristallographie,* Universités Paris 6 et 7, 4 place Jussieu, 75252 Paris, France*

[2]*Laboratoire des Verres,* Université Montpellier 2, place E. Bataillon, 34095 Montpellier, France*

[3]*Laboratoire de Physique des Solides,* Université Paris 11, Centre d'Orsay, 91405 Orsay, France*

(Received 18 February 2000)

The packing geometry of amino acids in folded proteins is analyzed via a modified Voronoï tessellation method which distinguishes bulk and surface. From a statistical analysis of the Voronoï cells over 40 representative proteins, it appears that the packings are in average similar to random packings of hard spheres encountered in condensed matter physics, with a quite strong fivefold local symmetry. Moreover, the statistics permits one to establish a classification of amino acids in terms of increasing propensity to be buried in agreement with what is known from chemical considerations.

PACS numbers: 87.15.By, 02.70.Lq, 61.43.–j

A folded protein is a very compact and reproductible packing of different amino-acid residues (a.a.) 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 for many purposes. Up to now the only efficient tools are x-rays which requires crystallized samples and NMR. Both techniques necessitate a large amount of materials, as well as sufficient accumulation of data to resolve a structure. Quite a lot of structures are presently known [2], at least enough to recognize some general trends but, anyway, they represent only a small fraction of the total number of proteins. Of course, it would be very interesting to be able to have an idea of the structure of a folded protein given its set of constitutive a.a. Therefore, any statistical analysis of the peculiar positions of each a.a. in the already known packings is welcome and efforts have already been made in this direction [3].

In this Letter we analyze the structure of folded proteins by means of a quite common tool used in condensed matter physics, namely, the Voronoï tessellation (VT) [4]. Given a set of points, VT proceeds by determining for each point the polyhedron, called "Voronoï cell," containing the portion of space closer to that point than to all others. 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 [5]. There are several examples of VT methods applied to proteins in the literature [6,7] but only a few of them [7] concern directly the packing of a.a. Moreover, in [7], the authors 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. Since an α carbon is almost systematically located on the border of the volume occupied by its corresponding a.a., we have preferred, in this study, to consider the geometrical centers of the lateral chain of the a.a. We think that, with this choice, the cells are representing topologically better the true volume occupied by the a.a. [8]. Moreover, since proteins are finite objects with a quite large surface over volume ratio (larger than most of the systems considered in condensed matter physics), we have modified VT to distinguish between surface and bulk a.a. After performing a statistics over the bulk cell characteristics of a wide set of different globular proteins [9] it turns out that they are quite typical of arrangements encountered in disordered condensed matter systems such as frustrated packings of spheres [10,11] with a quite strong fivefold local symmetry. Furthermore, when restricting the statistics to a given a.a., we find that the average cell characteristics are typical of this a.a. For instance, the percentage p of occurrence in the bulk of proteins permits one to classify the a.a. according to their burying propensity (strongly related to hydrophobicity) in agreement with what is already known [3,12].

We have considered the known structures of 40 globular folded proteins chosen to belong to independent folds and representative of the different classes. The number of a.a. per protein varies from 54 (for the 1enh [2] protein) to 686 (for the 1cdg [2] one). For each protein, we have determined the geometrical centers of all the a.a. from which it is made. This has been done by calculating, for each a.a., the barycenter of the atoms of its lateral chain [8] considering the same weight for each atom and discarding the hydrogens. This barycenter is very close to the true center of mass as all the atoms, except hydrogens, have almost the same mass. Then for each protein, given the set of centers (called points in the following) we have performed a VT by using our own code which has proven to be very efficient when applied to large assemblies of spheres [13] and to glasses [14]. Moreover, this code has been modified to take care of surface effects. We have first determined the Delaunay simplicial tetrahedra (ST) [5] which are, among all the tetrahedra formed with four points, the ones such that no other point lies inside their circumscribed sphere. Note that the edges of ST define unambiguously nearest-neighboring pairs of points. For a quite compact arrangement such as those studied here the ST located in the bulk exhibit quite regular shapes (relatively close to the one of a regular tetrahedron), while those close to the

 0031-9007/00/85(16)/3532(4)$15.00

surface can be very flat and/or strongly elongated (due to the lack of neighbors). We have decided to discard such tetrahedra by introducing a cutoff r_c: any tetrahedron which has a circumscribed sphere with radius larger than r_c is not considered. Then, given the ST, the VT cells can be built knowing that their vertices are the centers of ST's circumscribed spheres. The number of edges e for a polygonal face of a cell is the number of distinct ST sharing a given edge and the number of faces f is the number of distinct ST sharing a given vertex. Moreover, the face area and cell volume V of a VT cell can be calculated by picking up their constitutive elements within the corresponding ST as illustrated in Fig. 1. The total cell surface area S can be calculated by summing up the face surface area. When building the cells it turns out that for peculiar points, the cells are not closed (the sum of the vertex solid angles of the ST does not fit 4π): this defines the surface points by opposition to bulk points. It should be noticed that this definition depends on the choice for r_c: if r_c is too large, only the rising points (with strong local positive curvature) are considered, while if r_c is too small some internal points can be artificially considered as surface points. In practice r_c gives an idea of the smallest local radius of curvature authorized for deeps in the external surface. We have checked, on different examples of finite sphere packings without holes, that a reasonable choice is $r_c = 1.5d_m$, where d_m is the mean distance between neighboring bulk points. In practice in our case of protein structures we have taken $r_c = 10$ Å to be compared with the mean neighboring distance between volumic a.a. which was found to be of about 6.6 Å. Anyway, we have noticed that, even with our reasonable choice for r_c, some cells close to the surface remain elongated. This means that we cannot avoid some finite-size effects due to the presence of an external surface; in particular, some bulk cells close to the surface have fewer faces than those located well inside the bulk.

To have an idea of the overall geometry of the packing of a.a. in proteins, two quantities of interest are the mean number of faces per cell $\langle f \rangle$, which is the mean coordination number between a.a. and the mean number of edges per face $\langle e \rangle$, which is related to the symmetry around bonds of neighboring a.a. When averaging these quantities over all the closed cells (corresponding to bulk points) of our whole collection of 40 proteins, we find $\langle e \rangle = 5.14$ and $\langle f \rangle = 13.97$, values remarkably close to compact structures often encountered in condensed matter physics [5]. A value of $\langle e \rangle$ close to 5 is characteristic of compact structures built with local rules. Such fivefold local symmetry cannot be extended in the regular three dimensional space due to geometrical frustration [15]. To be more precise, we give in Fig. 2 the histograms $h(e)$ and $h(f)$, fraction of faces with e edges and fraction of cells with f faces, respectively. For comparison, we have shown the corresponding histograms for the most compact random packing of equal spheres, the so-called Bernal packing [10] of packing fraction 0.645 here built with the Jodrey-Tory algorithm [16] considering 10 000 spheres in a cube with periodic boundary conditions (CPBC) [13]. We have also

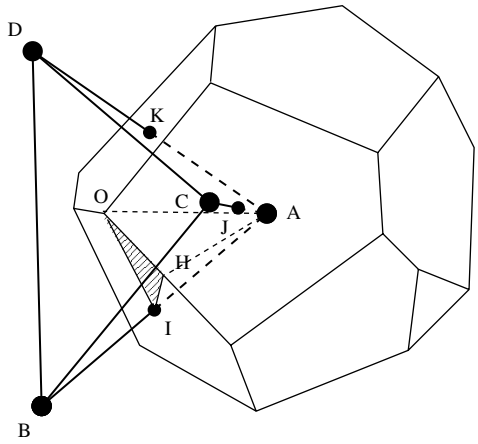

FIG. 1. A VT cell (corresponding to a Leu of the 3chy protein), whose center is called A, is depicted together with one of the ST involving A (thick lines). Three faces of this cell are perpendicular to AB, AC, and AD in their middle, I, J, K, respectively, and meet in O, center of the sphere circumscribed to $ABCD$. Elementary tetrahedra, such as $OAIH$ (where H is the orthogonal projection of O on the plane ABC), are used to collect the characteristics of this VT cell. Summing up algebraically their volumes and the surface area (such as the dashed one) allow one to calculate the volume of the cell and the area of its faces.

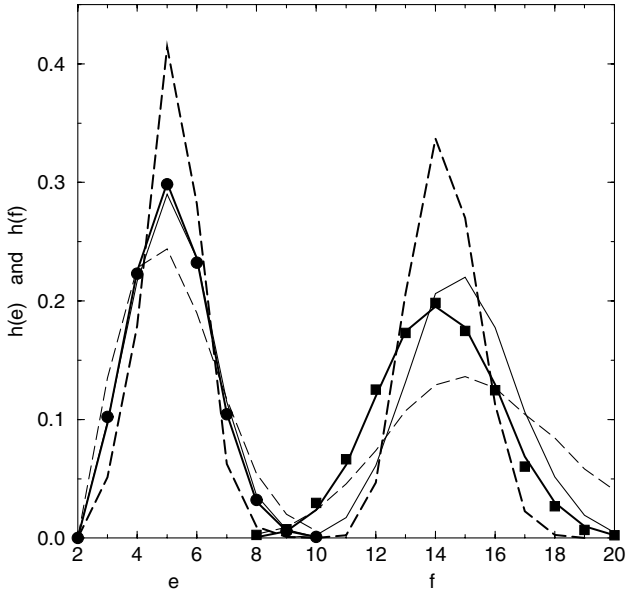

FIG. 2. Histograms $h(e)$ (dots) and $h(f)$ (squares) for the numbers of edges e per face and the number of faces f per cell, obtained from an average over all the closed cells (corresponding to bulk a.a.) of the whole collection of proteins. Thin and thick dashed lines correspond to random points and Bernal packing, respectively. Thin and thick continuous lines correspond to the RSA packing at jamming and an average over finite clusters made out of it, respectively (see text).

shown the histograms obtained by considering a uniformly random set of points (10 000 points in a CPBC). It is interesting to notice that the protein histograms are intermediate between those of a random set of points and those of the most compact random packing of spheres. We can obtain some reasonable fits by considering less compact sphere packings than the Bernal one. In the figure we have shown the histograms for the so-called random addition model (RSA) [17] close to its jamming threshold (packing fraction of about 0.38), made of 10 000 spheres in a CPBC. In this model a packing is built sequentially with the simple rule that the next sphere, whose center is chosen at random, should not overlap the previous ones. The jamming limit is obtained when no extra sphere can be added. While $h(e)$ fits the corresponding protein histogram quite well, $h(f)$ have a similar shape as the experimental one but are horizontally shifted as if they had systematically fewer faces in the case of proteins. This is simply due to finite-size effects. When considering instead an average over finite clusters, made of RSA sphere centers contained in a large sphere of 8 times longer diameter, containing about 200 points (well representative of the number of a.a. per protein), we get excellent fits for both $h(e)$ and $h(f)$ as seen in Fig. 2. Other fits could have been obtained by considering other building rules or polydisperse sphere packings. Anyway, all these fits should not be taken too seriously. They demonstrate only that the packing of a.a. in proteins resembles, on average, the ones commonly found in disordered systems of condensed matter physics.

To try to go further, for each a.a., we have counted the percentage p of occurrence for which it appears in the bulk of a protein and, when it is so, we have averaged the cell characteristics, such as f, e, S, and V, over the whole set of available proteins. In Fig. 3 we have chosen to give $\langle V\rangle$, as well as the dimensionless surface over volume ratio $R = 0.434\langle S\rangle^{1/2}/\langle V\rangle^{1/3}$ as a function of p. Such a ratio is here defined to be equal to 1 for a regular dodecahedron, which is the ideal cell corresponding to the locally most compact arrangement of spheres [15]. This ratio is larger for less spherical (anisotropic) cells. The 20 a.a. have been labeled using the standard [18] one-letter code. Here, when increasing p, they appear in the following order [19]: K = Lys, E = Glu, D = Asp, Q = Gln, R = Arg, P = Pro, N = Asn, T = Thr, S = Ser, H = His, G = Gly, A = Ala, Y = Tyr, M = Met, W = Trp, V = Val, L = Leu, F = Phe, C = Cys, I = Ile (the notation C′ will be introduced later). It is remarkable that this order roughly corresponds to the order of increasing hydrophobicity as already known from chemical considerations [3,12]. Furthermore, there is definitely a gap around 50% separating the hydrophilic a.a. ($p < 50\%$) from the hydrophobic ones ($p > 50\%$). Although the curves giving $\langle V\rangle$ and R are quite scattered, some general trends can be observed: in the hydrophilic case $\langle V\rangle$ decreases when p increases while R stays roughly constant (about 1.045) and in the hydrophobic case R decreases

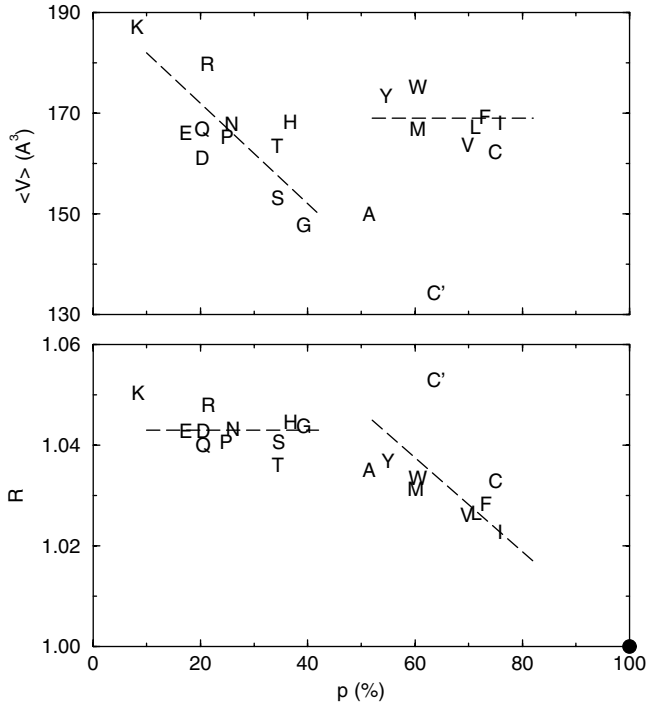

FIG. 3. Mean cell volume $\langle V\rangle$ (top) and dimensionless ratio $R = 0.434\langle S\rangle^{1/2}/\langle V\rangle^{1/3}$ (bottom), as a function of p, percentage of appearance of a given a.a. in the bulk of a protein. The letter labels of the a.a. are centered on the data. The dashed lines are guides for the eyes. The black dot corresponds to a regular dodecahedron.

when p increases while $\langle V\rangle$ stays roughly constant (about 170 Å^3). This means that the reasons to be more buried are different in the two cases: smaller size in the hydrophilic case and better spherical shape in the hydrophobic case. It is also worth noticing that a linear fit of R through the hydrophobic data extrapolates nicely to one as p tends to 100%. This means that the more hydrophobic the a.a., the more they try to pack as in an ideal icosahedral structure with dodecahedral cells. One can even go further, e.g., among the largest p values is a subgroup of four a.a., I, F, L, V, which are known to be highly hydrophobic and to be the main constituents of regular secondary structures [3,20,21]. The C's are quite peculiar as they can be linked together through a covalent disulfide bridges [18]. Two states can be distinguished, free cysteine and a covalently linked one (i.e., half-cystine), named C and C′, respectively, in Fig. 3. Note that C′ is, as expected, clearly apart from the other a.a., while C joins the top buried a.a., coherently with previous observations [22]. Moreover, the covalently linked cysteine data align quite well with those of hydrophilic a.a., as if the high value of p is due only to an artificially smaller size (due to bonding).

Although the above analysis is enlightening, it remains that it is based on small variations of the cell characteristics over the different a.a. As seen in Fig. 3 the data for $\langle V\rangle$ vary by about 10% around the mean. The same conclusion is reached when looking at other quantities not

reported here. The mean distances between neighboring a.a. (which are the edge lengths of ST) stay in the range 6.8 to 7.5 Å, varying roughly like $\langle V \rangle^{1/3}$. Moveover, when performing partial histograms $h(e)$ and $h(f)$ over hydrophobic and hydrophilic a.a., separately, they look remarkably identical to those reported in Fig. 2 [apart from a trivial shift of $h(f)$ due to the fact that hydrophilic a.a. are more likely located near the external surface]. All this justifies *a posteriori* the comparison with simple models, like hard sphere packings, based on geometrical rules independent on the chemical details, and suggest that essential ingredients to understanding protein folding are compaction and hard core repulsions. Some efforts have recently been made in this direction [23]. Of course, the chemical details are very important as they govern both the biological properties of each protein and the final differences between their folded structures.

In conclusion, using a very simple geometrical tool such as VT, we have shown that the packings of a.a. in folded proteins resemble those found in condensed matter physics. Moreover, by determining their probability to be in the bulk, we have established an ordering of the a.a. which constitutes a new scale to estimate burying from sequence data alone and likely opens new ways to investigate and predict protein features. This calculation is a first step and we intend to go further by enlarging the statistics and by analyzing the correlations between a.a., directly in the packing or along the peptide chain. We hope the tools of condensed matter physics will more often be used to study proteins as we think they provide a novel and powerful approach to this promising field of research.

A. S., J. C., and J.-P. M. acknowledge support from "Programme Genome" (Centre National de la Recherche Scientifique, France). R. J. thanks Magali Jullien for stimulative discussions.

*Laboratoire associé au Centre National de la Recherche Scientifique (CNRS, France).

[1] C. Chothia, M. Levitt, and D. Richardson, Proc. Natl. Acad. Sci. U.S.A. **74**, 4130 (1977); A. Adzhubei and M. Stenberg, Protein Sci. **3**, 2395 (1994).

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000); see also the Protein Data Bank on the web site: http://www.rcsb.org/pdb/

[3] See, for example, I. Callebaut, G. Labesse, P. Durand, A. Poupon, L. Canard, J. Chomilier, B. Henrissat, and J. P. Mornon, Cell Mol. Life Sci. **53**, 621 (1997); C. Dobson, A. Sali, and M. Karplus, Angew. Chem., Int. Ed. Engl. **37**, 868 (1998).

[4] J. L. Finney, Proc. R. Soc. London A **319**, 479 (1970); B. N. Boots, Metallography **15**, 53 (1982); M. R. Hoare, J. Non-Cryst. Solids **31**, 157 (1978).

[5] See, for example, *Foams and Emulsions,* edited by J. F. Sadoc and N. Rivier (Kluwer Academic, Dordrecht, The Netherlands, 1999); *Physics of Glasses,* edited by P. Jund and R. Jullien (American Institute of Physics, New York, 1999); *Disorder and Granular Media,* edited by D. Bideau and J. P. Hansen (Elsevier, New York/Amsterdam, 1993).

[6] F. M. Richards, J. Mol. Biol. **82**, 1 (1974); M. Gerstein, J. Tsai, and M. Levitt, J. Mol. Biol. **249**, 955 (1995); J. Liang, H. Edelsbrunner, P. Fu, P. Sudhakar, and S. Subramaniam, Proteins **33**, 1 (1998); J. Tsai, R. Taylor, C. Chothia, and M. Gerstein, J. Mol. Biol. **290**, 253 (1999).

[7] R. K. Singh, A. Tropsha, and I. I. Vaisman, J. Comput. Biol. **3**, 213 (1996); H. Wako and T. Yamato, Protein Eng. **11**, 981 (1998); P. J. Munson and R. K. Singh, Protein Sci. **6**, 1467 (1997).

[8] Including the backbone atoms together with the lateral chain atoms to determine the a.a. centers does not change the results significantly: the data of Fig. 2 remain within their error bar and Fig. 3 exhibits only one inversion, M and W. But, anyway, these a.a. remain extremely close to each other.

[9] In this study the few membrane proteins for which 3D structures are presently available have not been considered, but they will constitute an attractive point for further investigations.

[10] J. D. Bernal, Proc. R. Soc. London A **280**, 299 (1964).

[11] R. Jullien, P. Meakin, and A. Pavlovitch, in *Disorder and Granular Media* (Ref. [5]), p. 103.

[12] Y. Nozaki and C. Tanford, J. Biol. Chem. **246**, 2211 (1971); R. M. Sweet and D. Eisenberg, J. Mol. Biol. **171**, 479 (1983).

[13] R. Jullien, P. Jund, D. Caprion, and D. Quitmann, Phys. Rev. E **54**, 6035 (1996).

[14] P. Jund, D. Caprion, and R. Jullien, Phys. Rev. Lett. **79**, 91 (1997).

[15] J. F. Sadoc and R. Mosseri, *Geometrical Frustration* (Cambridge University Press, Cambridge, England, 1999).

[16] W. S. Jodrey and E. M. Tory, Phys. Rev. A **32**, 2347 (1985).

[17] D. W. Cooper, Phys. Rev. A **38**, 522 (1988).

[18] C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland Publishing, New York and London, 1999), 2nd ed.

[19] We have checked that this order stays independent of r_c in a large range of r_c values.

[20] S. Woodcock, J. P. Mornon, and B. Henrissat, Protein Eng. **5**, 629 (1992).

[21] A. Poupon and J. P. Mornon (to be published).

[22] L. Canard, Ph.D. thesis, University of Paris 6, 1997.

[23] See, for example, J. F. Sadoc and N. Rivier, Eur. Phys. J. B **12**, 309 (1999).