Protein Secondary Structure Assignment Through Voronoï Tessellation

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · MAY 2004

Impact Factor: 2.63 · DOI: 10.1002/prot.10566 · Source: PubMed

3 AUTHORS, INCLUDING:

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

PROTEINS: Structure, Function, and Bioinformatics 55:519–528 (2004)

Protein Secondary Structure Assignment Through Voronoï Tessellation

Franck Dupuis,[1] Jean-François Sadoc,[2] and Jean-Paul Mornon[1]*
[1]*Laboratoire de Minéralogie Cristallographie Paris, CNRS UMR 7590, Universités Paris 6 et 7, Paris, France*
[2]*Laboratoire de Physique des Solides, CNRS UMR 8502, Université Paris 11, Orsay, France*

ABSTRACT **We present a new automatic algorithm, named VoTAP (Voronoï Tessellation Assignment Procedure), which assigns secondary structures of a polypeptide chain using the list of α-carbon coordinates. This program uses three-dimensional Voronoï tessellation. This geometrical tool associates with each amino acid a Voronoï polyhedron, the faces of which unambiguously define contacts between residues. Thanks to the face area, for the contacts close together along the primary structure (low-order contacts) a distinction is made between strong and normal ones. This new definition yields new contact matrices, which are analyzed and used to assign secondary structures. This assignment is performed in two stages. The first one uses contacts between residues close together along the primary structure and is based on data collected on a bank of 282 well-refined nonredundant structures. In this bank, associations were made between the prints defined by these low-order contacts and the assignments performed by different automatic methods. The second step focuses on the strand assignment and uses contacts between distant residues. Comparison with several other automatic assignment methods are presented, and the influence of resolution on the assignment is investigated. Proteins 2004;55:519–528.** © 2004 Wiley-Liss, Inc.

Key words: voronoï tessellation; α-helices; β-strands; contact matrices

INTRODUCTION

Since the prediction of α-helices,[1] π-helices,[1] and β-sheets,[2] different methods of assignment have been developed. Initially, the only way for crystallographers to assign secondary structures was to perform a visual inspection. Since then, a number of authors have designed different algorithms to automatically assign a part or all the different kinds of structures present in proteins.[3–10] Nowadays, the most widely used assignment programs are DSSP,[11] DEFINE,[12] P-Curve,[13] and STRIDE.[14] These methods, which are based on different approaches, logically produce results that can be slightly different. Colloc'h et al.[15] showed that the percentage match score between DSSP, P-Curve, and DEFINE was only 63%, mainly because of discrepancies in the length of assigned structural elements. As already noticed by several authors,[11,12,14] this problem reveals that no method can be considered as the best one because each one is correct but only in the context of its own definition. Moreover, each algorithm produces artifacts and Colloc'h et al.[15] showed that they can be attenuated by the use of a consensus assignment in which each residue is assigned to the state determined by at least two of the three tested algorithms. One part of our work is based on the same approach because we used the assignments of four programs (DSSP, DEFINE, STRIDE, and P-SEA[16]) as reference data to elaborate a new assignment method.

This approach was combined with the use of a very sensitive geometrical tool based on the three-dimensional (3D) Voronoï tessellation.[17] A tessellation is a way of subdividing space into regions associated with each element of a set of discrete points to characterize their topological relations. For each of these elements, the Voronoï process associates a polyhedron called Voronoï cell defined by the intersection of contact planes built midway between points. Therefore, a cell defines the neighborhood closer to its associated point than to all the others, and its faces define the contacts with its nearest neighbors. For a given set of points, the Voronoï decomposition is unique and absolute because there is no empty space between cells. The set of Voronoï polyhedron is called a Voronoï diagram, and the cells' characteristics are very informative on the packing of the associated point set. In fact, for protein structures, two different scales can be investigated. The atomic level, which associates one cell with each atom or each group of atoms present in the structure, is the most commonly used.[18–23]. The amino acid (AA) level associates one cell with each residue, and the starting point set may be constituted of existing atoms (α- or β-carbons for instance[24–28]) or virtual points as the geometric centers of each residue.[29,30] Not only is this second approach less represented but also it is not a Voronoï tessellation in every case, but rather, a Delaunay tessellation, which is a dual of a Voronoï diagram.[25–28]

Abbreviations: AA: amino acid; PDB: Protein Data Bank; VoTAP: voronoï tessellation assignment procedure; 3D: three-dimensional; Cα: α carbon.

*Correspondence to: Jean-Paul Mornon, Laboratoire de Minéralogie Cristallographie, Universités Paris 6 et 7, case 115, 4 place Jussieu, 75252, Paris, France. E-mail: mornon@lmcp.jussieu.fr

Received 20 March 2003; Accepted 9 June 2003

Published online 1 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.10566

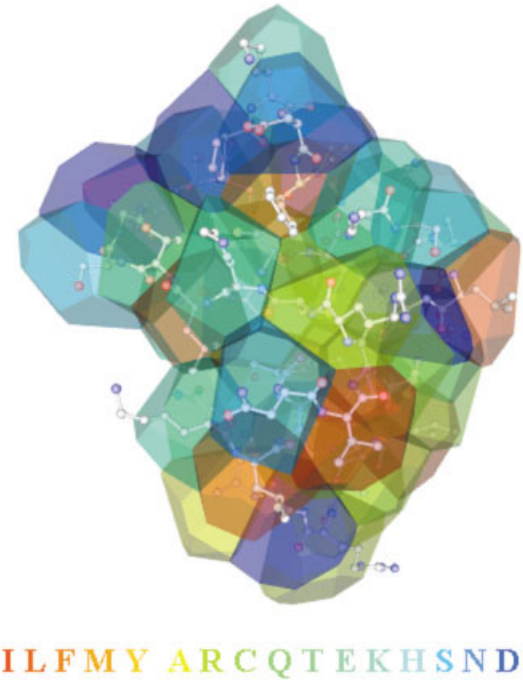

Fig. 1. Voronoï tessellation of the plant toxin β-purothionin from *Triticum aestivum* (PDB code 1bhp; 45 AA; resolution 1.70 Å). The colors of cells correspond to hydrophilic/hydrophobic properties of AA (blue to red). The protein atomic coordinates are represented in ball and stick: the sphere color depends on the atom nature (white for carbon, blue for nitrogen, red for oxygen, and yellow for sulfur). In this view, N- and C-termini are hidden and are not indicated. All the 45 cells corresponding to the 45 AA are represented, but the considered 587 spheres mimicking their environment are not shown. The protein does not contain Tryptophan.

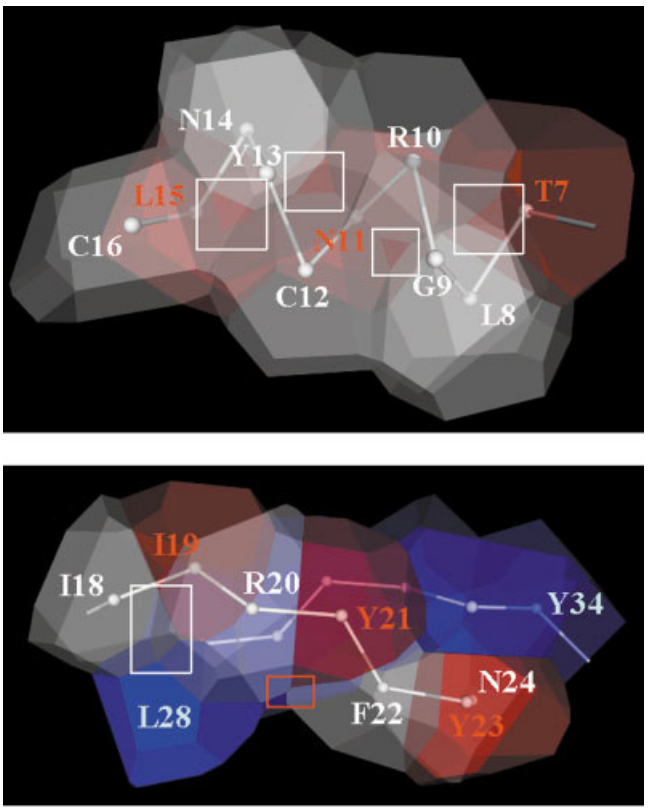

Fig. 3. **a:** Voronoï tessellation of the helix T7-C16 from 1bhp. Only the Cα trace is represented; the cells of T7, N11, and L15 from right to left are in red for visual commodity. The white squares highlight some typical helix triangle faces between AA i and i ± 2. **b:** Voronoï tessellation of two antiparallel strands in contact 1bpi I18-N24 is in red and gray, L28-Y34 is in blue, and L28 cell is not transparent for visual commodity. The two strands share a lot of faces. Red cells have no faces in common, whereas gray ones have a very small one (0.11 $Å^2$ between F22 and R20, red rectangle) and a bigger one (4.45 $Å^2$ between I18 and R20, white rectangle). These two features, added to the fact that strands do not contain a lot of AA, explain why intrastrand contacts are less informative than interstrand ones.

In this study, we considered α-carbons (Cα) as a starting point set, and we used the Voronoï tessellation to establish contact maps. Contact maps are derived from distance maps, also called distance plots or distance matrices, the utility of which has been recognized for a long time.[31,32] The most commonly used distance matrix is that containing all pairwise distances between Cα. It is independent of the coordinate frame and except for the overall chirality, it contains all the information present in an ordinate coordinate file concerning the backbone. This very practical tool is in fact a bidimensional representation of a 3D structure. Distance constraints between residue pairs lead to contact maps or contact matrices. Many authors have used these matrices, but their definitions of contacts between residues are not always the same because they always depend on a cutoff distance.[33–38] We propose here a new definition of contact between residues based on the Voronoï tessellation performed on Cα and independent of any cutoff. Contact matrices are then established, analyzed, and used to assign secondary structures.

MATERIALS AND METHODS

Voronoï Tessellation, Environment, and Relaxation

In this work, we used a code derived from a previous one[30] in which the geometrical centers of the lateral chain were considered. Here, we performed the tessellations on the Cα atoms because these are always present even at low resolution and also because they are more adapted to the detection of the secondary element regularity. Cα coordinates are extracted from the original file deposited in the Protein Date Bank.[39] To construct the Voronoï cells, the first step of the algorithm concerns the determination of the Delaunay tetrahedral decomposition. This unique decomposition is made of tetrahedra, the vertices of which are the Cα of the structure. The most important property of this decomposition is that the circumscribed sphere of each tetrahedron does not contain any other Cα. Thus, the set of tetrahedra sharing a particular Cα as a vertex unambiguously defines its closest neighborhood. Each face of the resulting Voronoï polyhedron is then associated with one point of this neighborhood. For the Cα close to the protein surface, the lack of neighbors often produces Voronoï cells that cannot be closed.

To overcome this difficulty, we used a model environment in which the structure is embedded. The method is the same as the one used as the precedent.[29] This environment is a relaxed random packing of spheres spread around the structure. It was designed to play the role of a

solvent with packing properties close to those of proteins with a thickness corresponding to at least three layers of spheres. The volume of these spheres is taken close to the mean volume occupied by amino acids, leading to a diameter of 6.5 Å. To eliminate the discrepancies appearing at the protein surface, we performed an iterative relaxation of the environment as follows: a Voronoï tessellation is performed on the Cα and the centers of the environment spheres. The geometric centers of the environment cells are calculated and are then considered as the new environment points. At this stage, a new Voronoï tessellation can be performed, and this operation is repeated nine times to reach convergence, which means that the geometric centers and the environment points are at the same positions.

Figure 1 shows an example of a Voronoï tessellation of the plant toxin β-purothionin from *Triticum aestivum* (PDB code 1bhp ; 45 AA ; 1 chain).[40] The atomic coordinates from crystallographic X-ray structure obtained at 1.70 Å resolution were used. The color of cells corresponds to hydrophilic (in blue) to hydrophobic (in red) properties of AA.[41] All the 45 cells of the protein are represented and can be seen by transparency.

Contact Map

Because a Voronoï cell defines the closest neighborhood of its corresponding Cα, we can consider that two amino acids are in contact if their corresponding cells share a face. This topological definition needs no cutoff; moreover, each contact can now be characterized by different properties such as the face area, the number of edges, the perimeter, and the distance between the two Cα of the amino acids in contact.

With this definition, it becomes possible to establish the contact map of a protein structure. A protein of N AA can be described by a $N \times N$ matrix, each element noted a_{ij} is either 1 (or 2 for further considerations) if AA i and j are in contact or 0 if not. The matrix is symmetric because, by contact definition, $a_{ij} = a_{ji}$ and $a_{ii} = 0$.

To obtain a more detailed information for contacts between AA close together along the primary structure (typically with a sequence gap $\leq$ 6 AA), we chose to separate the "strong" contacts from the "normal" ones. To that aim, 282 protein chains with resolution < 2.5 Å were selected from the PDB release of July 2002 and constitute what we call the StatBank (see Table I). To avoid structural redundancy, each chain of this bank contains domains from different superfamilies as defined in SCOP.[42] We determined the mean area of contact faces for all the 210 possible pairs of AA for each sequence gap ranging from 1 to 6 AA; for sequence gaps $>$ 6 AA, the number of contacts between some pairs of AA became too small. These statistics were performed on the 282 protein structures.

When the contact matrix of a particular structure is established, for contacts between AA with small sequence gap, we compared the face area to the corresponding mean. If the area was larger than the mean increase of 2 Å^2, we considered that the contact was a strong one and reported 2 instead of 1 on the matrix; if it was smaller, we considered that it was a normal one and reported 1. The Voronoï tessellations are used here to describe in a simple way the topological relations between AA; that is the reason why the calculated area must not be considered as the real contact surface between two residues. The value of 2 Å^2 was chosen to minimize the number of discordances between our assignment algorithm and DSSP, PSEA, DEFINE, and STRIDE.

The contact map derived from the protein 3-isopropylmalate dehydrogenase from *thiobacillus ferrooxidans* (PDB code 1a05; 357 AA; chain A)[43] is represented in Figure 2. The N-terminal end is in the left bottom corner. Below the matrix and on its sides is represented the assignment of the structure according to PROMOTIF[44]: white boxes

TABLE I. List of the 282 and 194 Proteins (PDB code) Constituting, StatBank and CheckBank, Respectively

StatBank

1a05	1ail	1b5e	1c3m	1d1p	1e19	1f5m	1kpf	1qb2	1ycq
1a28	1air	1b8o	1c52	1d3v	1e20	1f60	1kuh	1qc7	256b
1a2z	1aj8	1b93	1c5e	1d4t	1e2a	1f8y	1lau	1qf9	2a0b
1a34	1ako	1bb9	1c76	1d6r	1e6t	1f94	1lbe	1qhw	2acy
1a3a	1amk	1bdo	1c9h	1d8d	1eay	1fd3	1ldt	1qjb	2afg
1a44	1amw	1bhd	1c9o	1d9t	1ed1	1flj	1lki	1qje	2ahj
1a4m	1aoe	1bhp	1c9s	1dan	1edy	1fqt	1luc	1qk2	2bbk
1a4y	1aoh	1bj1	1cbj	1dce	1ei7	1fs1	1mat	1qsd	2ctc
1a58	1apx	1bjp	1cbk	1dcp	1ej1	1fup	1mka	1qtn	2e2c
1a6o	1aqb	1bk5	1cc8	1dd3	1ejf	1fxo	1mml	1rav	2end
1a7w	1ars	1blu	1cfy	1dd6	1ekg	1fzd	1mro	1ris	2kau
1a80	1aun	1bm8	1cfz	1dif	1ekj	1g24	1msk	1rop	2mhr
1a8b	1ava	1bm9	1ciq	1dk0	1el6	1g71	1nhk	1rpx	2nsy
1a99	1avb	1bo4	1cjd	1dk7	1em9	1gak	1noa	1sfp	2pth
1aa7	1aw8	1bov	1ck4	1dlm	1emv	1gci	1opc	1tbg	2sic
1aac	1awc	1bpl	1cku	1dlw	1enh	1gen	1oun	1tfe	2wrp
1aap	1awd	1bs4	1cl8	1dm9	1eq6	1gnk	1oyc	1tig	3cla
1aaz	1ay7	1bs9	1cmb	1doz	1esc	1got	1p35	1toa	3daa
1abe	1ayx	1btn	1coz	1dqi	1eur	1gym	1pbw	1tup	3eip
1abr	1azp	1bu7	1cq3	1dqo	1euv	1hfe	1pdo	1tyf	3pyp
1ad1	1azz	1bue	1cqd	1ds7	1eyv	1hoe	1php	1ugi	3tdt
1ad6	1b00	1bx4	1cqy	1dsz	1ez3	1hyp	1pml	1utg	4aah
1ae1	1b0n	1bx7	1cvw	1dtd	1f05	1icf	1poc	1vhh	4icb
1aew	1b0w	1byf	1cxy	1dun	1f0c	1imb	1poh	1vhr	4pah
1ag9	1b0x	1bzy	1d0b	1dwv	1f3u	1jdw	1pud	1vid	4ubp
1agi	1b33	1clk	1d0i	1dxe	1f3v	1jpc	1pyt	1wgj	6ins
1agj	1b3a	1c26	1d0q	1dxj	1f3z	1knb	1qau	1whi	6prc
1ah7	1b3t	1c2t	1d1j	1dzr	1f41	1kp6	1qb0	1who	7cei
1aho	1b4b								

CheckBank

1a17	1b7d	1bxy	1dio	1egw	1fxd	1lts	1qex	1tml	2cro
1a1x	1bbp	1byq	1dj8	1ej8	1g31	1mjc	1qgh	1ukz	2dor
1a6j	1bbz	1byr	1djr	1eo9	1g43	1mnm	1qhv	1unk	2erl
1ah4	1bd8	1c08	1dk8	1ep0	1g6g	1moq	1qip	1uro	2izh
1ahs	1beo	1cex	1dly	1eqo	1gdo	1mwp	1qkj	1ute	2lis
1aj2	1bf4	1chd	1dp4	1ert	1gpr	1mzm	1qq8	1vcc	2mta
1al3	1bix	1cjb	1dpj	1ew4	1gux	1nba	1qtw	1vmo	2pii
1amx	1bkp	1cjw	1dqe	1eyq	1hcb	1ndd	1reg	1wap	2plc
1ann	1bkr	1cnu	1dv8	1f2k	1hcq	1nec	1rfs	1wba	2poo
1apa	1blx	1co6	1dyn	1f21	1hle	1nnc	1rkd	1xxa	2prd
1arb	1bn8	1ctf	1dz3	1f7d	1htp	1npk	1rvv	1ycs	2prg
1arv	1bou	1cv8	1e6i	1fas	1hxn	1otf	1sei	2abk	2rn2
1aug	1bpi	1cyn	1e79	1fkj	1ibr	1pbv	1sml	2arc	2tnf
1ax0	1bqk	1cyo	1eai	1fle	1ida	1pfk	1spp	2asr	2trc
1ayf	1br9	1dlg	1ecm	1flm	1ido	1phr	1srv	2ay1	3fap
1b16	1bu5	1d2z	1ecp	1flt	1inp	1ppf	1stm	2bnh	3fib
1bli	1buo	1d4o	1ecy	1fof	1iro	1ptf	1sup	2bop	4fgf
1b4f	1bv1	1d7d	1edm	1fq0	1jac	1puc	1tcd	2cev	4mon
1b66	1bvy	1dea	1efv	1fua	1kpt	1qd9	1tif	2cpg	4nos
1b67	1bxs	1dfu	1egp						

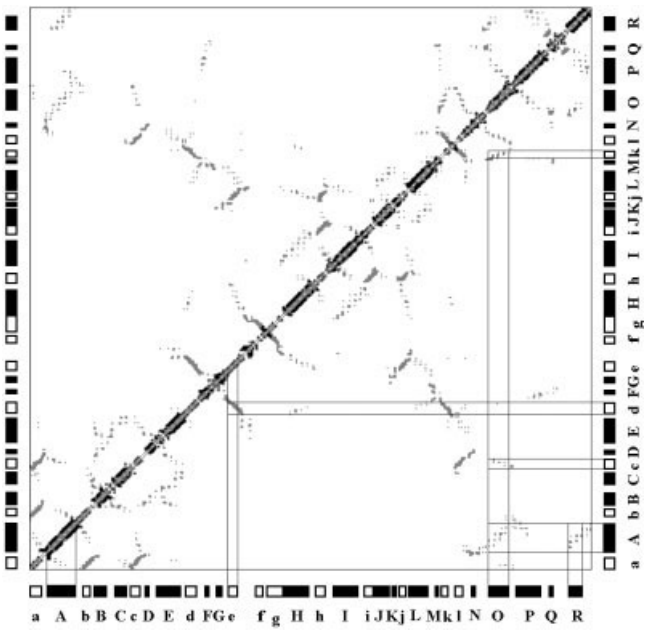

Fig. 2. Contact map derived from the protein 3-isopropylmalate dehydrogenase from *thiobacillus ferrooxidans* (PDB code 1a05; 357 AA; resolution 2 Å; chain A). The N-terminal end is in the left bottom corner. Contacts are represented in gray or in black for strong contacts with sequence gap < 6 AA. The secondary structures are represented on the sides and below the matrix. Black box represent helices and are labeled with capital letters (from A to R). White box represent strands and are labeled with small letters (from a to l). Vertical and horizontal lines highlight contacts between some of the secondary structures.

represent strands, and black ones represent helices. Normal contacts are represented by gray dots; strong contacts are in black.

As expected, most of the contacts can be found along the diagonal and represent lower order contacts. Therefore, they often represent intrasecondary structure contacts, and the characteristics of the diagonal geometry/shape can be relevant to the nature of the corresponding secondary structure. For instance, helices are characterized by a thick diagonal with sides continuously black. The geometry of the α-helix favors the contacts between AA i and i $\pm$ 4 (and also i $\pm$ 3), and the contacts observed are strong ones, whereas trivial contacts between AA i and AA i $\pm$ 1 or AA i $\pm$ 2 are most of the time normal ones. Figure 3(a) represents a tessellated helix and shows one of the most typical topological features of helices: contact faces between AA i and AA i $\pm$ 2 are most of the time small triangles. This observation means that these contacts are not favored; nevertheless, the helix topology induces them. For strands (Fig. 2), the diagonal is thinner because contacts between AA i and i $\pm$ 3 and above are rarely observed, and contrary to the helices for which strong contacts between AA i and i $\pm$ 2 are very rare, they are much more frequent for strands even though they are not systematic, as illustrated in Figure 3(b). It shows two strands in contact (1bpi[45] I18-N24 in gray and red and L28-Y34 in blue). There are no intrastrand contacts with a sequence gap ≥ 3 AA. It can be seen that gray cells make contacts with a sequence gap of two AA (with normal or very small faces), whereas red ones do not. These observations added to the fact that strands are shorter than helices (in number of AA) explain why they are less recognizable. The loops between regular secondary structures do not seem to have specific detectable characteristics.

Most of the contacts not found along the diagonal (Fig. 2) are not evenly distributed on the map because they form lines that can be classified according to their direction and their texture. The lines parallel/perpendicular to the diagonal represent contacts between parallel/antiparallel regular secondary structures of the same nature, respectively. This is simply explained by the fact that when AA i and j from two different secondary structures are in contact, AA I + δ is also in contact with j + δ in the parallel case and with j $-$ δ in the antiparallel case. The line texture helps to distinguish contacts between two helices or two strands. A thin continuous line indicates that two strands are in contact (parallel and antiparallel); in this case, $\delta = 1$ and AA i touches AA j and AA j + 1 and j $-$ 1 but rarely j + 3 or j $-$ 3. Faces between two strands are numerous and quite large, explaining why these contacts seem more informative than the lower order contacts (intrastrand contacts); see Figure 3(b). The contact line between antiparallel strands e and d in Figure 2 is an example. Contacts between two helices are represented by thick broken lines; in this case, $\delta = 4$ and AA i touches AA j and AA j + 4 or/and j $-$ 4 but not necessarily AA j $\pm$ 1 or AA j $\pm$ 2. In Figure 2, an example is given by the contact lines between parallel helices A and O and A and R, which are the first and the last helix of the protein, respectively. The other lines represent contacts between helices and strands. In this case, if AA i and j are in contact, AA I + δ_1 is in contact with AA j $\pm$ δ_2 (δ_1 and δ_2 vary in each different case): as a consequence, these lines are neither parallel nor perpendicular to the diagonal, but they have directions that stand between these two extremities. Because a helix is always involved in this kind of contacts, the line is always a broken one, like the one between helix O and strand k in Figure 2.

Statistics

From these observations, we aimed at evaluating if it was possible from the Cα coordinates deposited in a PDB file and from an appropriate Voronoï tessellation to detect the secondary elements present in the protein structure, and therefore, to assign each AA to one of the three states: helix (α, 3_{10}, and π), strand (parallel and anti-parallel), or coil (turn, loop). Our method is based on statistical data extracted from the StatBank 3D structures.

These structures were processed by four programs of secondary structure assignment, DSSP, DEFINE, STRIDE, and P-SEA. The outputs of these programs were treated in such a way that the final results consist only of four three-state assignments: A for helix (α, 3_{10}, and π), B for strands (parallel, antiparallel, and β-bridges), and C for everything else.

Each protein of the StatBank was put in its environment, which was relaxed nine times and then finally

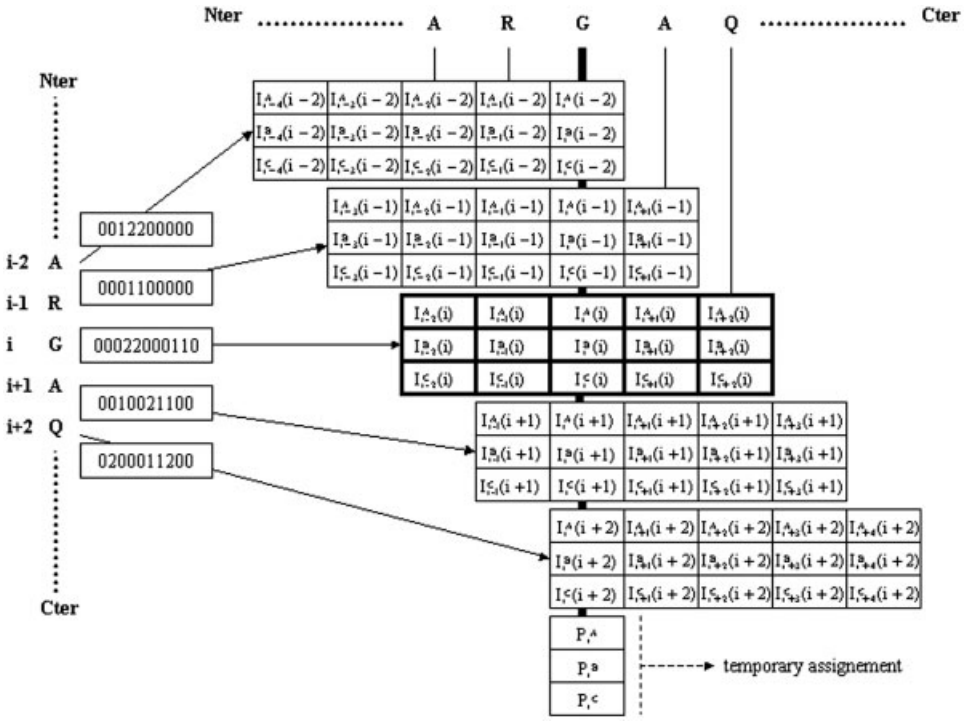

Fig. 4. This scheme presents a part of the process described in part (i) for AA number i. From left to right: for each AA from AA i − 2 to i + 2, the corresponding prints are extracted from the contact map. Each print is associated with the corresponding array derived from Tab_Cen (or Tab_Ext for the first or last six residues). One column of each array (from the last one for AA i − 2 to the first for AA i + 2) gives a probability for each conformation (A, B, or C) for AA i. For each conformation the mean probability P_i^X (X stands for A, B, or C) is calculated.

tessellated. From the corresponding contact matrix, and for AA number i, we extracted the values from a_{ii-6} to a_{ii-2} and a_{ii+2} to a_{ii+6} [the central triplet (a_{ii-1}, a_{ii}, a_{ii+1}) was of no interest because it is almost always equal to (1,0,1)] to obtain a linear matrix of 10 elements representing the sequential neighborhood of the considered residue, which we will call in this article the print of AA number i. This print was associated to the four quintuplets (one for each algorithm) formed by the assignments (A, B, or C) of residues from i − 2 to i + 2. This operation was performed for all the residues of the StatBank except for the first six and the last six residues of each chain. In these cases, this kind of print was not suitable because some residues were necessarily missing to establish it. In fact, for these amino acids, the distinction was no longer made between strong or normal contacts but between possible or impossible ones. For these residues, a 0 in the print still means that there was no contact, a 1 means that there was a contact (strong or normal), but a 2 means that there could not be a contact. For instance, for the first residue at the N-terminus of a chain, the print starts by (22222…). To obtain quintuplets, the existing assignments were completed with coils.

All of the 243 (3^5) possible quintuplets were enumerated for each encountered print, and the occurrence of each kind of quintuplet was noted $N(X_1\ X_2\ X_3\ X_4\ X_5)$. For instance, the print (0012112100) was encountered 4519 times in the StatBank and 4425 times (97.92%) with the quintuplet (AAAAA) (for this print N(AAAAA) = 4425) and a few times with (CAAAA) or (AAAAC). It was then possible to associate each print with the frequency of observation of each kind of assignment for each position in the quintuplet. This can be summed up in a (3 × 5) array as the central one in Figure 4, in which each frequency is noted $I_{i+p}^X(i)$ with p ranging from −2 to +2 as the position in the quintuplet centered on I; X one of the three possible assignments A, B, or C; and i the position in the sequence of the considered AA. For instance, for the print associated to AA i, $I_{i+2}^A(i)$ is the frequency of apparition of AA i + 2 assigned A and can be calculated as follows:

$$I_{i+2}^A(i) = \sum_{Xn=A,B,C;n\epsilon[1,4]} N(X_1X_2X_3X_4A) \Big/ \sum_{Xn=A,B,C;n\epsilon[1,5]} N(X_1X_2X_3X_4X_5).$$

From this definition, it can be easily seen that $I_{i+2}^A(i) + I_{i+2}^B(i) + I_{i+2}^C(i) = 1$, explaining why $I_{i+2}^A(i)$ was considered as the probability for the last residue of the quintuplet to be in the helix conformation for the print associated to AA i. The 3718 different prints and corresponding arrays were stored in a table named, Tab_Cen. For the residues at the N- and C-termini, the 388 prints were stored in a table named Tab_Ext.

Algorithm

To assign any residue from any PDB file, the procedure starts by selecting the Cα coordinates. The environment is

added and relaxed nine times, and the final tessellation is then performed to obtain the corresponding contact matrix. Then the assignment process can be divided in two different parts.

The first part uses the information contained near the matrix diagonal. The print of each residue is determined and associated with the corresponding array from Tab_Cen or Tab_Ext. For all the amino acids, each in its turn, the same process is applied and is represented in Figure 4. For AA i, the arrays from AA i − 2 to AA i + 2 are considered and three mean probabilities are calculated:

$$P_i^X = \frac{1}{5} \sum_{j=-2}^{2} I_i^X(i + j)$$

where X can be A, B, or C. The greatest mean will give the temporary assignment to AA i. At the end of this part of the process, a succession of temporary secondary structures is obtained. The probability P_i^X associated with each residue gives an estimation of the reliability of the temporary assignment. For instance, >0.6, this assignment always agrees with most of the four automatic methods used (typically intrahelices residues). Between 0.5 and 0.6, the agreement with the most is not systematic but there is almost always one or two automatic programs that agree (typically helix ends or some strand ends). Below 0.5, discordances with all the automatic assignments are much more frequent and occur between helix or strand and coil (strands ends). If a print (or more) is not found in one of the two tables, the mean is calculated on the other probabilities. This property also provides a good robustness under variations such as loop deletions.

The criteria applied in the second part were refined to minimize the number of discordances between our algorithm and DSSP, PSEA, DEFINE, and STRIDE. This part of the assignment process uses the interstrand information. Contacts between temporary strands are searched and analyzed to lengthen or shorten them. Short regular secondary elements are suppressed. The different steps of this process are performed in the order as follows.

When a residue i assigned to strand with $P_i^B > 0.5$ is detected, the algorithm looks for contacts with every residue j assigned to strand or to coil with $P_j^C < 0.6$ and $|i - j| > 6$ to avoid intrastructure contacts. When found, contacts between AA i + d or i − d and j + d or j − d with d starting at 1 are searched with the same conditions. If this contact exists, the search is repeated for d + 1 and so on, the last value of d is noted D. If D > 3, then all the residues between j ± (D-1) are assigned to strand. This new assignment is not taken into consideration in the next step, which follows the same principle but with different conditions. When a residue i assigned to strand or to coil with $P_i^C < 0.6$ is detected, the program looks for contacts with residue j assigned to strand or to coil with $P_j^C < 0.6$ and $|i - j| > 6$. When found, contacts between AA i + d or i − d and j + d or j − d with d ranging from 1 to D are searched with the same conditions for i ± d but for j ± d the only condition is that its temporary assignment must be strand. If all the contacts exist and if D > 3, then all the residues between i ± (D-1) are assigned to strand.

For the next steps, these new assignments are used, but helices and strands composed of 2 residues or less are assigned to coil. The first stages of this part of the assignment process were created to try to lengthen the strands; the following step is created to check if the strands are correctly assigned.

To that aim, contacts between residues assigned to strand are detected and their neighborhoods along the primary structure are checked. All the contacts between two residues i and j assigned strands are listed and for each one, the following conditions must be fulfilled at least once. For AA i, if the next and/or previous residue in the sequence is also assigned to strand, then the next or/and previous residue of AA j must also be assigned to strand. If these conditions are never fulfilled and if $P_i^B < 0.6$, then the residue is finally assigned to coil; the last step of the process is to suppress once again all the strands with <3 AA.

TABLE II. Comparison of Secondary Structure Assignment Algorithms for Each Kind of Elements

	Agreement level of VoTAP with previous algorithms				
	DSSP (%)	PSEA (%)	STRIDE (%)	DEFINE (%)	CONSENSUS (%)
Helix	93.0	92.7	96.7	96.7	95.6
Strand	77.3	79.7	79.1	73.1	82.1
Coil	79.3	83.1	78.3	64.2	82.7
Total	83.2	85.3	84.4	76.9	86.7

On a residue per residue basis, the percent of agreement of the secondary structure elements are detailed for each of the three states (helix, strand, and coil). Our results are compared with those of DSSP, PSEA, DEFINE, STRIDE, and a consensus approach based on these algorithms.

RESULTS AND DISCUSSION

To check the validity of our algorithm, we used a second bank of protein structures called CheckBank. The construction method is the same as the one used for StatBank but with different structures because the StatBank is used to elaborate Tab_Cen and Tab_Ext. We compared our results with the assignments performed on the 194 structures (see Table I) of CheckBank by DSSP, PSEA, DEFINE, and STRIDE. Because our algorithm performs a three-state assignment, we converted the different states of each program into three classes following the convention previously detailed. The comparison of the five algorithms is given (see Table II) for each type of secondary structure element. In this table, the last column corresponds to a consensus assignment inspired by the work of Colloc'h et al.; in this consensus, each residue is assigned to the state determined by at least two of the three following methods (DSSP, PSEA, and DEFINE). STRIDE was not considered in this consensus to avoid a bias due to the great agreement between DSSP and STRIDE (>90%; see Table III).

According to the agreement (>90%; see Table II) of the five methods tested here, the helices are the better as-

TABLE III. Comparison of Secondary Structure Assignment Algorithms

Versus	DSSP (%)	PSEA (%)	STRIDE (%)	DEFINE (%)	CONSENSUS (%)
DSSP	—	80.2	95.8	73.0	87.8
PSEA	80.2	—	81.4	77.2	92.0
STRIDE	95.8	81.4	—	74.4	87.7
DEFINE	73.0	77.2	74.4	—	84.8
CONSENSUS	87.8	92.0	87.7	84.8	—
VoTAP	83.2	85.3	84.4	76.9	86.7

On a residue by residue basis, the percent of agreement of the secondary structure elements between the different methods are presented (DSSP, PSEA, DEFINE, STRIDE, consensus, and VoTAP). In the consensus method, each AA is assigned to the state determined by at least two of the three methods (DSSP, PSEA, and DEFINE). Because of the good agreement between DSSP and STRIDE (>95%), the latter was not considered to avoid a bias in the consensus.

signed secondary structure elements. This is mainly explained by two factors that are combined: the regularity of the helical geometry and the number of residues involved in a helix conformation. The consequence is a great number of residues corresponding to a few types of prints; for instance, the 10 most represented prints out of a total of 3718 (0.3%) correspond to 6325 amino acids (out of 48911 AA of the StatBank: 12.9%), which are assigned to the helix conformation by one of the four studied algorithms. For strand assignment, the agreement is lower, which could be explained by the likeness of the prints corresponding to strands and to certain coils. Moreover, the extended strands are generally short (in AA), and the regularity of the strands is consequently more difficult to detect than for helices, and as it was already noticed,[15] strands are generally more difficult to assign than helices. That is why our strand assignment is divided into two parts, combining intrastrands and interstrands contacts and finally a verification and, if necessary, a correction of these assignments.

DEFINE is an algorithm based on the distance matrix concept and compares inter Cα distances to ideal structures. This algorithm is sensitive to ϵ, the cumulative discrepancy between the ideal and observed distance matrix. We used an ϵ of 0.75 Å for helices and of 0.5 Å for strands instead of the default value of 1 Å which produced an excess of secondary structures.[46,47] Nevertheless, the agreement with DSSP, PSEA, and STRIDE (see Table III) showed the deficiencies of this parameterization. One of the main consequences is a too important composition in regular secondary elements (63.5%). Despite the fact that DEFINE contributes to the computation of the consensus, its agreement with the consensus is lower than that of VoTAP. DSSP is in better agreement than VoTAP with the consensus but only by 1%, a small value if we consider that DSSP also contributes to the consensus computation. Because STRIDE is in a very good agreement with DSSP (>95%), it is not surprising that the agreement with the consensus is quite the same. PSEA is the only algorithm that actually surpasses VoTAP. It is interesting to notice that PSEA, like VoTAP, uses only Cα and is based on a consensus because it uses simultaneously distance and angle criteria.

The proportion and length of regular secondary elements assigned by VoTAP stand between the extremities of all the other programs (data not shown); the length of helices and strands are almost the same as the ones assigned by PSEA, whereas the proportions are a little different. From this point of view (proportion and length), our method seems to be a good compromise between these programs.

For 8 of 33,193 residues (0.02%) in the CheckBank, the assignment was not possible because the five consecutive prints were not present in the Tab_Cen or in the Tab_Ext. By default, these residues were assigned to the coil conformation; in the eight cases, this assignment was in agreement with the other methods.

Our method is based on statistics derived from a new form of consensus assignment, which has several advantages. With the consensus proposed by Colloc'h et al.,[15] the assignment of the considered residue depends only on the assignments of this particular AA determined by at least two of three different methods. This implies that the information of the nonretained assignment is lost. But a regular secondary structure is mainly defined by the topological or geometrical relations that its constitutive elements make with their neighborhood. Our method presents the advantage to take into account not only the four assignments of the considered residue but also the four assignments of the two residues preceding and following the residue. The artifacts inherent to each algorithm are dissolved into the pertinent information, which is the result of the concordance between the different methods. It would have been very interesting to use the program P-curve,[13] which is based on very different properties such as the helicoidal parameters of each peptide unit. This very different point of view would have enriched our statistics but this algorithm was not available at the time.

VoTAP is independent of any geometrical cutoff because it has the advantage of not considering any particular criterion. The only feature of interest is the local topology, which is given by the Voronoï tessellation. For a particular structure, this information is absolute and unique, and contrary to pure numerical criteria, the derived neighborhood can accommodate some distortions. To investigate to which extent this assertion was true, we studied the influence of the structure resolution on the behavior of the different algorithms. To do so, we used 26 structures, which were superseded in the PDB by better refined structures. These structures are archived in the PDBObs, which is accessible from the PDB. The 52 structures are listed in Table IV; the first column lists the 26 superseded entries with the lower resolution (mean resolution: 3.0 Å), which constitutes the LR bank, and the third column lists the 26 corresponding structures with a better resolution (mean resolution: 1.9 Å), which constitutes the HR bank. Assignments were performed on these structures and the results of the different methods were compared. For some structures in the LR bank, only Cα coordinates were available; that is why DSSP and STRIDE were not able to assign all the residues of this bank. We also noticed that for some proteins, DEFINE was not able to assign all the

TABLE IV. PDB Codes List of the 26 Proteins of the Lower Resolution Bank (LR) and of the Corresponding Proteins of the Higher Resolution Bank (HR)

LR bank		HR bank	
PDB code	R (Å)	PDB code	R (Å)
14ps	2.6	1qjb	2.0
151c	2.4	351c	1.6
1abk	2.0	2abk	1.9
1abp	2.4	1abe	1.7
1abx†	3.5	2abx	2.5
1act	2.8	2act	1.7
1afn	2.6	2afn	2.0
1alp	2.8	2alp	1.7
1baa	2.8	2baa	1.8
1bjl	3.0	3bjl	2.3
3cha‡	2.8	5cha‡	1.7
4cna	2.9	5cna	2.0
1cpp†	2.6	2cpp	1.6
2dpv‡	3.3	4dpv‡	2.9
1erl	1.6	2erl‡	1.0
2fnr	3.0	1fnd	1.7
1fxb	2.3	1iqz‡	0.9
1grs†‡	3.0	3grs	1.5
1mhr†	5.5	2mhr	1.7
5pfk†	7.0	6pfk	2.6
2psi	2.9	1qlp	2.0
2rhn	3.5	1ayn	2.9
1scp†	3.0	2scp	2.0
1sdh	2.4	3sdh	1.4
1sga	2.8	2sga	1.5
1trc	3.6	1fw4	1.7

The 26 protein structures composing the low resolution bank (LR) are extracted from PDBObs, which is a bank of superseded entries of the PDB, the resolution of which is indicated in column 2. The corresponding new entries of the PDB and their resolutions are listed column 3 and 4, respectively. †Only Cα coordinates were available. ‡DEFINE was not able to perform the assignment.

residues. The results are shown in column 2 of Table V. The difference between LR and HR assignment is indicated for each algorithm in column 3. This table highlights the deficiencies of DEFINE, which has difficulties to stay coherent between LR and HR with an agreement of 86.8% with only 77.5% of the bank. The best agreements are performed by DSSP and STRIDE, but they only assign 74.4% of the bank and are not able to detect secondary elements of the structures with the lowest resolutions (e.g., 5pfk and 1mhr). PSEA and our algorithm are the only ones that were able to assign all the residues of the bank; the agreement between HR and LR is a little better for PSEA than for our method and are both comparable to the one of DSSP. When the data were available in the PDB file, we used the crystallographers assignments as "standard of truth" to compare the assignments of the different methods and the evolution between the LR bank and the HR one. The results are reported in columns 4 and 5 of Table V. Despite the fact that the agreement between LR and HR is 86.8%, the agreement between the DEFINE and crystallographers assignments does not evolve between LR and HR and is the lowest of the five methods. The other gaps between the two banks vary from 1.6% for STRIDE to 3.1% for DSSP. For the only methods that can perform an assignment for all the residues, the gaps are 3.0% for PSEA and 2.7% for our method. VoTAP has a little advantage on PSEA because for the two banks, the agreement with the crystallographers assignments is better, and 78.6% of assignments are in agreement with the crystallographers' ones. For the protein 5pfk (resolution R = 7 Å), our algorithm detected 86.2% of the crystallographers assignment of 6pfk (resolution R = 2.6 Å), PSEA detected 81.5%, and DEFINE 82.4%.

CONCLUSION

This study shows that Voronoï tessellations provide an efficient tool to assign the secondary elements of any protein structure from the Cα positions. Because the only criterion used is the neighborhood of each residue, numerical values such as distances or angles and their associated cutoff are not necessary. Artifacts introduced by these cutoffs are then avoided. Moreover, the suppleness of the Voronoï tessellation gives good results even at low resolution. The comparison with other assignment methods or with crystallographer assignments show that VoTAP results are on average of the same quality, but a closer look at two actual cases reveals some interesting features. Figure 5(a) shows the assignments of the different methods used in this study (crystallographer assignment, DSSP, PSEA, STRIDE, DEFINE, and VoTAP) performed on the human thioredoxin (PDB code 1ert; 105 AA; R = 1.70 Å).[48] This protein provides typical examples of problems encountered when evaluating assignment procedures. The different methods agree on the number of secondary elements (9 structures with 5 sheets and 4 helices) and the number of assigned AA except for DEFINE which tend to assign longer secondary structures (crystallographers 73, DSSP 74, PSEA 75, STRIDE 74, VoTAP 72, and DEFINE 85). The dissensions between automatic methods lie at the secondary structure extremities. This problem is logical because its reasons are intrinsic to each algorithm and is at the same time a problem of detection (cutoff) and of definition. These methods consider different criteria to define regular secondary structures, and if they agree on the overall definition of what is a helix or a strand, they do not define their ends in the same way. It may be then "natural" that on average a maximum agreement exists. For instance, with 1ert one disagreement at each end of each secondary element would give a percentage agreement of about 83% (1–2 × 9/105). This hypothesis can be generalized to the CheckBank (33,193 AA and 2,183 secondary structures) for which one discrepancy on average by structure extremity would then lead to a maximum agreement of 86.8% (1–2 × 2,183/33,193). It is interesting to note that the accordance between VoTAP and the consensus is 86.7% (see Table II); this may imply that the performance quality of the consensus methods (e.g., VoTAP, which reexamines and enlarges this notion) resides in the fact that they tend to regularize the structure ends without defining them.

Figure 5(b) proposes another example of the different assignments performed on the bovine cytochrome b(5)

TABLE V. Comparison of the Five Assignment Methods

Methods	No. of assigned AA in LR (%)	Agreement between LR and HR (%)	Agreement between LR and crystallographers (%)	Agreement between HR and crystallographers (%)
VoTAP	100.0	88.2	78.6	81.3
DSSP	74.4	90.2	81.5	84.7
PSEA	100.0	90.1	76.9	79.9
STRIDE	74.4	91.4	81.0	82.6
DEFINE	77.5	86.8	73.0	73.8

Fig. 5. **a:** A typically well-assigned protein: different assignments of the human thioredoxin (PDB code 1ert; 105 AA; resolution 1.70 Å). **b:** A typical example of difficult assignment: the bovine cytochrome b(5) (PDB code 1cyo; 88 AA; resolution 1.50 Å). The first line is the crystallographers assignment present in the PDB file. Secondary structures are labeled according to this assignment: A or a for helices, B or b for strands, and are highlighted underlined. Capital letters represent AA, which have the same six assignments; in bold are the AA that have the same assignment as the one performed by the crystallographers. Italic letters represent discrepancies in secondary structures length between the different methods and the crystallographer assignment.

(PDB code 1cyo; 88 AA; R = 1.50 Å).[49] This case highlights another difficulty encountered with automatic assignment methods because they here disagree on the number of secondary elements. For instance, B3 is not assigned by PSEA, B5 like A6 is neither assigned by PSEA nor DEFINE and B4 is only detected by DSSP and STRIDE but as a single β-bridge. This last case is of further interest because VoTAP initially assigned the first two residues of B4 to the strand conformation, but VoTAP, like PSEA, assigns finally a strand if it has at least 3 AA, which is not the case of the other methods. For instance, DEFINE assigns a strand at the C-terminus of the chain and DSSP and STRIDE assign a strand between A1 and B2. These elements are neither detected by the other automatic methods nor by the crystallographers. The case of helix A2 shows that a consensus method can detect a secondary element when classical methods fail to do so because VoTAP is the only algorithm to detect this helix. The qualities of consensus methods, which were highlighted by Colloc'h et al.,[15] are thus confirmed here. It is noteworthy that these qualities are not specific to the assignment problem because it was shown[50] that it was better to combine secondary structures prediction methods than to use a single classical one.

This program will be part of a software especially conceived to study and visualize 3D Voronoï tessellation on protein structures at the amino acid level. Assignment and contact maps will be then available.

REFERENCES

1. Pauling L, Corey RB, Branson HR. The structure of proteins: two hydrogen-bonded helical configurations of the polypetptide chain. Proc Natl Acad Sci USA 1951;37:205–234.
2. Pauling L, Corey RB. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc Natl Acad Sci USA 1951;37:729–740.
3. Lewis PN, Momany FA, Scheraga HA. Folding of polypeptide chains in proteins: a proposed mechanism for folding. Proc Natl Acad Sci USA 1971;68:2293–2297.
4. Kuntz ID. Protein folding. J Am Chem Soc 1972;94:4009–4012.
5. Crawford JL, Lipscomb WN, Schellman CG. The reverse turn as a polypeptide conformation in globular proteins. Proc Natl Acad Sci USA 1973;70:538–542.
6. Levitt M, Greer J. Automatic identification of secondary structure in globular proteins. J Mol Biol 1977;114:181–239.
7. Rose GD, Seltzer JP. A new algorithm for finding the peptide chain turns in a globular protein. J Mol Biol 1977;113:153–164.
8. Chou PY, Fasman GD. Beta-turns in proteins. J Mol Biol 1977;115: 135–175.
9. Kolaskar AS, Ramabrahmam V, Soman KV. Reversals of polypeptide chain in globular proteins. Int J Pept Protein Res 1980;16:1–11.
10. Ramakrishnan C, Soman KV. Identification of secondary structures in globular proteins—a new algorithm. Int J Pept Protein Res 1982;20:218–237.
11. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
12. Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 1988;3:71–84.
13. Sklenar H, Etchebest C, Lavery R. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins 1989;6:46–60.

14. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
15. Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng 1993;6:377–382.
16. Labesse G, Colloc'h N, Pothier J, Mornon JP. P-SEA: a new efficient assignment of secondary structure from C α trace of proteins. Comput Appl Biosci 1997;13:291–295.
17. Voronoï G. Recherches sur les paralleloedres primitifs. J Reine Angew Math 1908;134:198–287.
18. Finney JL. Volume occupation, environment and accessibility in proteins. The problem of the protein surface. J Mol Biol 1975;96: 721–732.
19. Tsai J, Voss N, Gerstein M. Determining the minimum number of types necessary to represent the sizes of protein atoms. Bioinformatics 2001;17:949–956.
20. Quillin ML, Matthews BW. Accurate calculation of the density of proteins. Acta Crystallogr D Biol Crystallogr 2000;56:791–794.
21. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
22. Gerstein M, Chothia C. Packing at the protein-water interface. Proc Natl Acad Sci USA 1996;93:10167–10172.
23. Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 1974;82:1–14.
24. Zimmer R, Wohler M, Thiele R. New scoring schemes for protein fold recognition based on Voronoi contacts. Bioinformatics 1998;14: 295–308.
25. Zheng W, Cho SJ, Vaisman, II, Tropsha A. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. Pac Symp Biocomput 1997;486–497.
26. 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.
27. 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.
28. Wako H, Yamato T. Novel method to detect a motif of local structures in different protein conformations. Protein Eng 1998;11: 981–990.
29. Angelov B, Sadoc JF, Jullien R, Soyer A, Mornon JP, Chomilier J. Nonatomic solvent-driven Voronoi tessellation of proteins: an open tool to analyze protein folds. Proteins 2002;49:446–456.
30. 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.
31. Phillips DC. The development of crystallographic enzymology. Biochem Soc Symp 1970;30:11–28.
32. Nishikawa K, Ooi T. Comparison of homologous tertiary structures of proteins. J Theor Biol 1974;43:351–374.
33. Singer MS, Vriend G, Bywater RP. Prediction of protein residue contacts with a PDB-derived likelihood matrix. Protein Eng 2002;15:721–725.
34. Selbig J. Contact pattern-induced pair potentials for protein fold recognition. Protein Eng 1995;8:339–351.
35. Go M. Correlation of DNA exonic regions with protein structural units in haemoglobin. Nature 1981;291:90–92.
36. Galaktionov S, Nikiforovich GV, Marshall GR. Ab initio modeling of small, medium, and large loops in proteins. Biopolymers 2001;60:153–168.
37. Kim MK, Jernigan RL, Chirikjian GS. Efficient generation of feasible pathways for protein conformational transitions. Biophys J 2002;83:1620–1630.
38. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
39. 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.
40. Stec B, Rao U, Teeter MM. Refinement of purothionins reveals solute particles important for lattice formation and toxicity. Part 2. Structure of beta-purothionin at 1.7 Å resolution. Acta Crystallogr D Biol Crystallogr 1995;51:914–924.
41. 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. Cell Mol Life Sci 1997;53:621–645.
42. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
43. Imada K, Inagaki K, Matsunami H, Kawaguchi H, Tanaka H, Tanaka N, Namba K. Structure of 3-isopropylmalate dehydrogenase in complex with 3- isopropylmalate at 2.0 A resolution: the role of Glu88 in the unique substrate-recognition mechanism. Structure 1998;6:971–982.
44. Hutchinson EG, Thornton JM. PROMOTIF—a program to identify and analyze structural motifs in proteins. Protein Sci 1996;5: 212–220.
45. Parkin S, Rupp B, Hope H. Structure of bovine pancreatic trypsin inhibitor at 125 K definition of carboxyl-terminal residues Gly57 and Ala58. Acta Crystallogr D Biol Crystallogr 1996;52:18–29.
46. Colloc'h N, Cohen FE. Beta-breakers: an aperiodic secondary structure. J Mol Biol 1991;221:603–613.
47. Presnell SR, Cohen BI, Cohen FE. A segment-based approach to protein secondary structure prediction. Biochemistry 1992;31:983–993.
48. Weichsel A, Gasdaska JR, Powis G, Montfort WR. Crystal structures of reduced, oxidized, and mutated human thioredoxins: evidence for a regulatory homodimer. Structure 1996;4:735–751.
49. Durley RCE, Mathews FS. Refinement and structural analysis of bovine cytochrome B(5) at 1.5 angstrom resolution. Acta Crystallogr D Biol Crystallogr 1996;52:65–76.
50. King RD, Ouali M, Strong AT, Aly A, Elmaghraby A, Kantardzic M, Page D. Is it better to combine predictions? Protein Eng 2000;13:15–19.