PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins

Carlos H. da Silveira,[1*] Douglas E. V. Pires,[2] Raquel C. Minardi,[1] Cristina Ribeiro,[1] Caio J. M. Veloso,[3] Julio C. D. Lopes,[4] Wagner Meira Jr.,[2] Goran Neshich,[5] Carlos H. I. Ramos,[6] Raul Habesch, and Marcelo M. Santoro[1*]

[1] Department of Biochemistry and Immunology, Institute of Biological Sciences, Federal University of Minas Gerais, UFMG, Brazil
[2] Department of Computer Science, Federal University of Minas Gerais, UFMG, Brazil
[3] Pontifical Catholic University of Minas Gerais at Betim, PUCMG, Brazil
[4] Department of Chemistry, Federal University of Minas Gerais, UFMG, Brazil
[5] Structural Bioinformatics Group, National Agricultural Information Technology Research Center (CNPTIA), National Agricultural Research Corporation, EMBRAPA, Brazil
[6] Institute of Chemistry, University of Campinas, UNICAMP, Brazil

ABSTRACT

In this study, we carried out a comparative analysis between two classical methodologies to prospect residue contacts in proteins: the traditional cutoff dependent (CD) approach and cutoff free Delaunay tessellation (DT). In addition, two alternative coarse-grained forms to represent residues were tested: using alpha carbon (CA) and side chain geometric center (GC). A database was built, comprising three top classes: all alpha, all beta, and alpha/beta. We found that the cutoff value at about 7.0 Å emerges as an important distance parameter. Up to 7.0 Å, CD and DT properties are unified, which implies that at this distance all contacts are complete and legitimate (not occluded). We also have shown that DT has an intrinsic missing edges problem when mapping the first layer of neighbors. In proteins, it may produce systematic errors affecting mainly the contact network in beta chains with CA. The almost-Delaunay (AD) approach has been proposed to solve this DT problem. We found that even AD may not be an advantageous solution. As a consequence, in the strict range up to 7.0 Å, the CD approach revealed to be a simpler, more complete, and reliable technique than DT or AD. Finally, we have shown that coarse-grained residue representations may introduce bias in the analysis of neighbors in cutoffs up to 6.8 Å, with CA favoring alpha proteins and GC favoring beta proteins. This provides an additional argument pointing to the value of 7.0 Å as an important lower bound cutoff to be used in contact analysis of proteins.

Proteins 2009; 74:727–743.
© 2008 Wiley-Liss, Inc.

Key words: cutoff evaluation; protein packing; residue contacts; residue interactions; Delaunay tessellation; Voronoi diagrams; contact methods.

INTRODUCTION

Atom and residue contacts have been used in a wide range of studies involving proteins and other biomolecules. Its correct and precise assignment comprise the touchstone of the most important structural analysis algorithms, which should be able to perform: packing calculations,[1–4] functional similarities,[5] evolutionary relationships,[6] topological classifications,[7,8] structural alignments,[9] structural assessment,[10] protein structure prediction,[11] threading experiments,[12,13] network contact analysis,[14–16] empirical potentials,[17–19] thermodynamic stability previews,[20] folding inferences,[21,22] protein–protein and protein–ligand interactions,[23] and so forth. Here we will focus our attention on some methods that underlay contact characterizations in most of these applications.

Perhaps as diverse as these broad applications are the forms that a contact may be defined. The classical and simplest method is through the establishment of thresh-

Additional Supporting Information may be found in the online version of this article.

Abbreviations: AD, almost-Delaunay technique for defining contacts; ALPHA, SCOP all alpha subset; BETA, SCOP all beta subset; ALPHA/BETA, SCOP alpha/beta subset; BIC, Bayesian Information Criterion; CA, alpha carbon; CD, traditional cutoff dependent technique for defining contacts; DT, Delaunay tessellation; GC, amino acid side chain geometric center; VD, Voronoi diagrams.

Grant sponsor: UFMG (PhD Bioinformatic program).
*Correspondence to: Carlos H. da Silveira or Marcelo M. Santoro, 6627, PO Box 486, Pampulha, Minas Gerais 31270-901, Brazil. E-mail: carlos.silveira@gmail.com or santoro@ icb.ufmg.br
Received 24 January 2008; Revised 27 May 2008; Accepted 8 June 2008
Published online 14 August 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22187

old distances. For two given different points or sites $\{i,j\}$ in the atom or residue set, i will be in contact with j if the latter is inside of a sphere centered in the former with radius r, called cutoff. The real challenge is how to optimize the process of selecting the right cutoff. The literature offers a wide range of options: 3.8,[24] 4.5,[25] 5.0,[26] 5.5,[27] 6.0,[22] 6.5,[28] 7.0,[15] 8.0,[29] and 9.0[6] Å. Despite important attempts to rationalize these choices to certain contexts,[28–32] in most cases it seems that the values were quite arbitrary. Probably they were established mainly in order to satisfy optimization of the data processing in each particular case.

There are many other methods that use this traditional cutoff approach as a basis for their implementations. One of them adjusts the maximum distance to the length of the sites (generally, the van der Waals radius), by taking into account the radii of i and j plus a fix range r in-between.[33] Making r small enough is a way to try to consider only the first-order contact (the closest layer of neighbors). Again, the problem is the choice of a reasonable value for r. In general, a value in the range from 0.6 to 3.0 is chosen.[6,33–36] The contacts may also be weighted by some function, representing a contact area,[37] an energy potential,[14] an Euclidean distance,[8] or some type of normalization[25] as implemented by the radial distributions functions.[18,38] Additionally, these contacts may be seen not only as a pairwise collection of sites. It is possible to extend it to n-tuples, being common to form three[39] and four-body contacts.[40] There are also other more complex forms to assign contacts, like the occluded surface packing (OSP) metric,[2] the small-probe contact dot method,[10] and the relative contact order (CO).[22] But, apart some divergent details, all of these techniques make explicitly or implicitly (in general through a probe radius) use of a cutoff distances.

Furthermore, contact functions may handle the transition of cutoff in a discrete or continuous form. For the first, the contact is counted in an all-or-nothing fashion, generally using an unit step function.[28] In this definition, the contact list is susceptible to slight changes in the coordinates of the points at the limit of cutoff. An alternative approach is to smooth this transition state by a sigmoidal function,[41] resulting in a fractional (or continuous) number of contacts. In the empirical potentials context, Maiorov and Grippen[42] have shown that in spite of this smoothening be useful for the comparison of homologous structures, when extended to any protein set the discrete contact functions correlates linearly well with the continuous form (correlation coefficient of 0.997).

Yet another mode of assigning contacts in protein is through Voronoi[43] and Delaunay tessellations (DTs).[44] Its utilization in proteins was pioneered by the historical works of Richards[1] and Finney[45] in volume and packing calculations, and it has been growing in recent years, through numerous other applications.[46] We can define tessellation as a form of tiling a D-dimensional space. For an Euclidean space R^d, it implies in the possibility of using a collection of convex polytopes (the generalization term for any dimension of our more familiar "polygons" in 2D) to fulfill a region with no overlaps and no gaps. This strict fulfillment allows tessellation to capture special relationships among the sets of points. Voronoi and DTs are two correlated types of tiling that, following exact geometric criteria, yield the closest connectivity information about the neighborhood of the points. In proteins, the DT will result in a polyhedrization of the points so that all traced contact will be represented by edges and all sites (atoms or residues) by vertices.

It is common to classify contacts in cutoff dependent (CD) and cutoff free. In the former the cutoff parameter is an essential requirement for the definition of contact, as in the traditional contact approaches described earlier. Cutoff is also used to infer the energy of contacts in molecular simulations methods, especially for the truncation of long-ranges forces.[47] Conversely, the latter does not require the cutoff for the characterization of contacts. The DT, while a mathematic abstraction, is an example of cutoff free method, given that it is defined in a pure geometric ground where a threshold limit is unnecessary. Another instance of contacts that can be thought as cutoff free is when the distance threshold is set as infinity. A distance dependent energy function may weight the contacts so that devaluates the furthest points. The Veloso's occlusion method is an example of such strategy.[14] Two sites are in contact if there is no intervening site in-between, that is, if they are not occluded by others. An energy function is used to depreciate long, nonoccluded contacts. In molecular dynamics, the Ewald techniques[48] used mainly to measure the Coulomb energy of electrostatic contacts may also be assorted as cutoff free.

The contacts in proteins may yet to be categorized by the granularity of the points.[49] In the fine-grained models, the sites are conceived at atomic level, producing a more detailed (but more complex) representation of the protein. This fine granularity may also be utilized to map contacts at residue level. The most common approach is to assume that two residues are in contact if any of its heavy atoms are close enough.[26] In some models, the choice of heavy atoms may be more strict.[50] Others make use of statistics over the collection of closest atoms to attribute a weight to its residues in contact.[25] Coarse-grained models, on the other hand, may be used to lower the complexity of the system.[49] It is possible to simplify the residue presentation by designating one representative point, called centroid, from where the contact calculations are performed. Usual choices for representative points are: alpha carbons (CA),[29,31] beta carbons (CB),[51] geometric center (GC),[28] or barycenter (BC)[52] of the side chains (that may include or not some atom of the backbone).

In spite of all diversity of above mentioned contact definitions, we can infer here a common objective for

most of them: to map the presence or the location of the sites in a given space aiming to extract or to explore the underlying preferences (if any) in its spatial distribution. Important to this general definition is the realm of the contact concept, in special its terminological differentiation from interaction. As the etymology of the word suggest, interaction concerns to some action among agents in response to some type of mutual force. This force may be real as in the Coulomb interaction or apparent as in the case of hydrophobic "interaction," which may be regarded as a side effect of the entropic behavior of the system. Contact, as aforementioned, relates only to the presences and spatial distribution of the sites. Hence, if in a given system there are interactions among their constituent elements then it is expected that interactions will affect the components in someway, imposing some type of observable order or preference in-between. If (ideally) there are no interactions, it would be possible to calculate the contacts yet, but in this case it is expected that they would have a more random profile, without any apparent order or preference.

The correct identification of the first layer (or order) of neighbors is of crucial importance in many types of contact research. For instance, in the evaluation of empirical potentials, which intend to extract or to infer the relative energy of interactions from the statistical profile of contacts, in an approach generally recognized as an inversion of the Boltzmann law.[17–19] For these knowledge-based potentials, the first layer of neighbors may have a determinant role, because it will be more deeply affected by interactions, inasmuch as its influence rapidly decays in direction to high order contacts (in highest layers). Hence, it is generally accepted that the profile of the closest contacts contains more useful information about possible interactions. Other example of the relevance of the correct isolation of the first layer of neighbors concerns to the role of packing in the protein folding. There is an intense debate in the literature[1–4,50,87,88] about if there is or not a "packing code" ruling the hydrophobic residue aggregation when protein chain collapse towards its compact native state. Certainly, a reliable determination of the first layer of neighbors may be a mandatory procedure in approaching this question, given that packing involves exclusively residues in direct contact.

In this work, we are proposing to scrutinize some important questions, crucially dependent on contact definitions, with the explicit aim to best characterize the first layer of neighbors: How these methodologies may influence the contacts statistics? How the choice of the centroid type may change contact definitions? Is there a way to define the preferential cutoff that might be considered more appropriate than others? In addition to the previous challenge, we also wanted to compare the CD methods with the cutoff free approaches, such as the Voronoi/DT.

As these methodologies compose the base for important applications used today in structural bioinformatics, it is of fundamental importance to know about their idiosyncrasies, their limits, their divergences, and in which conditions they may bias the results. Here, we will be focusing on one representative method of each principal class of the contact definitions described above. We have examined the relationships between the traditional CD technique and the Voronoi/DT, at the level of residues, having as centroid the alpha carbons (CA) or the side chain GC. We have seen how these contact definition approaches behave when applied to three groups of 91 non-sequence related proteins: all alpha, all beta, and alpha/beta as defined by the SCOP classification.[76] We verified in preliminary tests that alpha+beta class generate outputs that behaved as a mix of alpha and beta pattern, so we will not present this result here. Indeed, the alpha+beta class is formed by segregated regions of helices and sheets, and is expected that it have a contact pattern that is approximately a sum of alpha and beta isolated classes. The remaining SCOP classes were also not considered. As a case study, we have analyzed how these methodologies recognize the structured neighborhood of the first-order contacts. Some intriguing results emerged from these comparisons, concerning not only to the topological fundaments of residue packing, but also to the applicability of DT and related techniques in proteins.

METHODS

Proteins sets

Thereafter, we will call "alpha" and "beta" any helix and sheet protein secondary structure elements, respectively, independent of subtype classifications. Hence, included in alpha are, for example, these known helices: α-helix, 3_{10}-helix, π-helix, and the rarer left hand helix. Through PDB[54] advanced search engine and STING_DB,[55] we composed three sets (ALPHA and BETA and ALPHA/BETA) with equal sizes, sampled from SCOP.[53] As STING_DB uses much more stringency annotation (coincidence of DSSP,[7] STRIDE,[56] and PDB annotation both in length and secondary structure element annotation) we used only DSSP as the filtering agent in order to get more structures into our data mart. All these proteins were filtered after applying the following general selection criteria: X-ray resolution less than 2.0 Å, *R*-Work less than 0.2, sequence identity less than 30%, and chain length between 50 and 600 amino acids. An initial search with these parameters found, at November of 2007, 248 proteins for ALPHA and 314 for BETA. To enhance the secondary structure signal in both sets, we checked its relative assignment according to DSSP.[7] For ALPHA, we accepted those that have more than 35% of alpha content and less than 12% of beta content. For BETA, in addition to the alpha content had less than

Table I
Protein Structure Databases

Database	Proteins[a]
ALPHA (91 chains)	1LMB3, 1B0N1, 1M451, 1VRK1, 1A7W1, 1ALV1, 1AMZ1, 1BGF1, 1DK81, 1DNU2, 1EYV1, 1FC31, 1FT51, 1G331, 1G4I1, 1GPQ2, 1GV21, 1HBK1, 1HBN2, 1HE11, 1I2T1, 1I8O1, 1J7Y2, 1JFB1, 1K0M1, 1KG21, 1KQF3, 1L9L1, 1LJ81, 1LKP1, 1M1N2, 1M4R1, 1M8Z1, 1M9X2, 1MTY2, 1MTY3, 1MXR1, 1MZ41, 1N1J1, 1N1J2, 1N2A1, 1NOG1, 1O081, 1O831, 1OOH1, 1OR01, 1OW41, 1OWL1, 1PBW1, 1PPR1, 1Q081, 1QGI1, 1QMG1, 1QOY1, 1R8S2, 1RRM1, 1SQ21, 1T6U1, 1T7R1, 1TX41, 1TZV1, 1TZY2, 1TZY4, 1VDK1, 1VLG1, 1W531, 1WDC2, 1WKU1, 1WOL1, 1WPB1, 1WVE2, 1K961, 1YOY1, 1YYD1, 1Z101, 2ABK1, 2BAA1, 2CCH2, 2CIW1, 2CZ21, 2EUT1, 2GC44, 2GKM1, 2I5N1, 2I5N3, 2I5N4, 2INC1, 2INC2, 451C1, 5CSM1,1BZR1
BETA (91 chains)	1JIW2, 1F582, 1SBW1, 1TGS1, 1A121, 1BHE1, 1C9O1, 1CRU1, 1EAJ1, 1EUR1, 1EUW1, 1F8E1, 1FLT2, 1FNS1, 1FNS2, 1GQ81, 1GSK1, 1GUI1, 1HOE1, 1I0C1, 1IBY1, 1J831, 1K121, 1KV71, 1LK33, 1LR51, 1M9Z1, 1NSZ1, 1O5U1, 1O6S2, 1OFL1, 1OFZ1, 1OH41, 1PBY2, 1PMH1, 1PNF1, 1PQ71, 1PXV2, 1QHV1, 1RG81, 1RMG1, 1ROC1, 1RWI1, 1SFD1, 1SQ91, 1SR43, 1SVB1, 1SVP1, 1T2W1, 1T611, 1T612, 1TCZ1, 1TUD1, 1UAC2, 1UMH1, 1USR1, 1UV41, 1UWW1, 1UXZ1, 1V051, 1V6P1, 1VPS1, 1WD31, 1XQH1, 1Y0M1, 1Y7B1, 1ZE31, 1ZE32, 1ZGO1, 2A2Q3, 2ADF2, 2ADF3, 2AG41, 2AGY1, 2BCM1, 2DJF1, 2FCB1, 2FGQ1, 2FK91, 2GC42, 2GC43, 2H3L1, 2HS11, 2IAV1, 2IVZ1, 2J1N1, 2O8L1, 2POR1, 2SIL1, 3EZM1, 1K5C1
ALPHA/BETA (91 chains)	1F2T2,1ABA1,1VSR1,1JW92,1I9C1,1A4Y2,1AY71,1AY72,1C1Y2,1CCW1,1CSE2,1DTD2,1F602,1GMX1,1H4X1, 1H751,1J3A1,1JF81,1GV81,1MJH1,1OGD1,1QTN2,1QZM1,1R0R2,1RLK1,1SCJ2,1SRV1,1U0S1,1U0S2, 1UC71,1UGH2,1VC11,2SIC2,1QNT1,1G2I1,1IM51,1J2R1,1C1Y1,1DF71,1FBT1,1IO01,1JUV1,1K1E1,1NQU1,1NXJ1, 1QTN1,1R2Q1,1SC31,1SHU1,1SVI1,1WDJ1,1JFX1,1JAY1,1IU81,1LK51,1M5W1,1UOK1,1EDG1,1EQC1,7A3H1, 1DXE1,1GPE1,3GRS1,4UAG1,1NZY2,1DCI1,1QF51,1TCA1,1CVL1,1B4Z1,1HX01,1JAK1,1I9C2,1KQP1,1KDG1,1LWD1, 1J181,1OCK1,1OI71,1P1J1,1GQN1,1T4B1,1TA32,1T9H1,1UJM1,1UMK1,1CCW2,1CSE1,1NMM2,1SCJ1,1U7P1

[a]Brookhaven data bank codes concatenated with the chain ID number, conform the order of its occurrence in the PDB files.

12%, we imposed that the relative number of residues in beta had at least two times more than that in alpha, and a nonstructured content less than 65%. These values were chosen after several tests which yielded less than the desired number of structures for further analysis. The cardinality of our set was reduced to 158 for ALPHA and 148 for BETA, at this stage. For ALPHA/BETA we started with a set of 554 proteins. No restriction was imposed to its relative secondary structure composition.

Improving the PDB content

To clean and filter the PDB content of the data mart that we extracted for analysis in this work, we used the PDBEST package,[57] a tool in development by our group that first accesses and then assesses the annotation quality of PDB files. The PDBEST is composed by a group of PERL scripts that apply a set of rules, defined by the user, on the original PDB files and releases them cleaned and filtered. To compose our PDBEST data the following rules (in high level language) were applied to each PDB file in the ALPHA, BETA, and ALPHA/BETA sets: detach chains; re-enumerate chains, residues and atoms; exclude chains with the same sequence in SEQRES tag; delete chains with residues missing atoms; exclude chains whose residue names are not standard amino acids (exception to selenomethionine); if there are atoms with more than one occupancy, choose that with the larger probability; if there are models, get the first; if annotation error in chain, residue or atom fields is detected, correct or warn about it. In the end of this process, the ALPHA and BETA sets were equalized in 182 proteins, 91 for each (Table I). The ALPHA/BETA set has also 91 chains and it was manipulated with intention to generate a chain length distribution close to ALPHA and BETA set.

A statistical summary of the quality of our data is shown on Figure 1. A complementary way to assess the homogeneity of our databases is to verify the surface/volume distribution in ALPHA, BETA, and ALPHA/BETA sets. This is also a way to certify that all sets have statistically the same globular character. Chothia and Janin[58] demonstrated, in an approximation to solid bodies of similar shapes, that the relation between the solvent accessible surface area (A_s) and molecular weight (M) may be given by:

$$A_s = k_a M^d \tag{1}$$

where k_a and d are constants. They found $k_a \approx 11.1$ and $d \approx 0.70$. The latter was assumed to be close enough to 2/3 as expected by a perfect sphere. We modified Eq. (1) to give a surface (A_s) to volume (V) ratio as a function of the number of residues (chain length) n:

$$\frac{A_s}{V} = k_b n^{-\frac{1}{3}} \tag{2}$$

The linear regression applied to a log–log transformation of Eq. (2) for ALPHA gave intercept 1.00 ± 0.24 and slope -0.36 ± 0.04, for BETA intercept 0.95 ± 0.22 and slope -0.36 ± 0.04, and for ALPHA/BETA intercept 0.82 ± 0.16 and slope -0.34 ± 0.04, all at 0.95 level of confidence. We saw that the regression parameters for all sets were homogeneous.

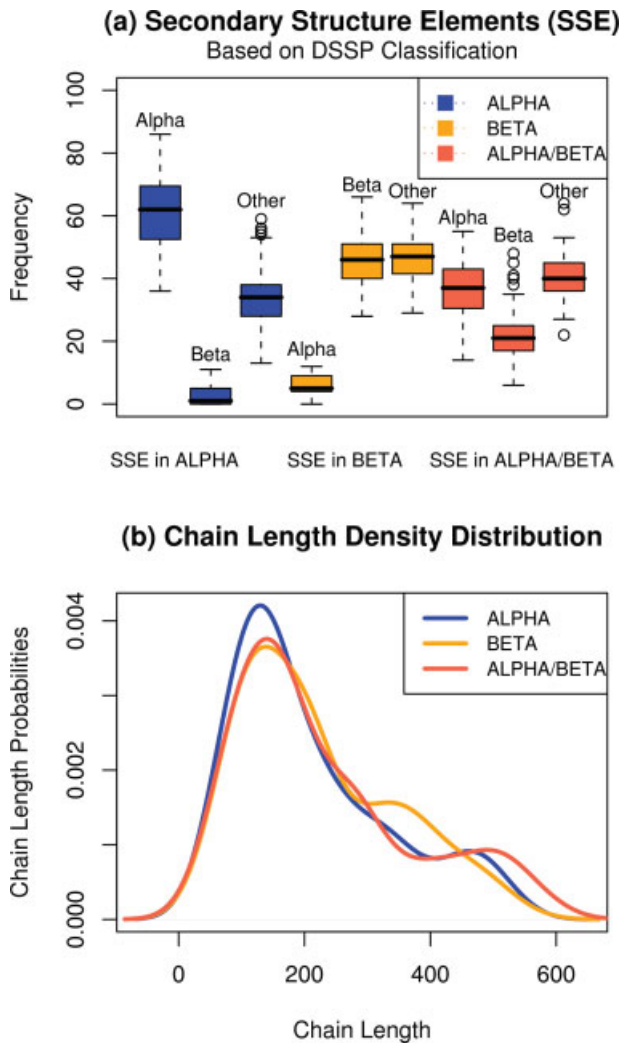

Figure 1

The assessment of ALPHA, BETA, and ALPHA/BETA data sets. (**a**) The DSSP assignment for secondary structure in ALPHA is in blue (dark gray), BETA is in orange (light gray) sets, ALPHA/BETA is in red (median gray). The alpha content in ALPHA had a mean and standard deviation of 61.6% ± 11.6%, with min/max of 36.0%/86.0%. The beta content in BETA had a mean and standard deviation of 46.6% ± 7.9%, with min/max of 28.0%/66.0%. The alpha content in ALPHA/BETA had a mean and standard deviation of 36.0% ± 9.6%, with min/max of 14.0%/55.0%; the beta content in ALPHA/BETA had a mean and standard deviation of 22.1% ± 7.8%, with min/max of 6.0%/48.0%. (**b**) Distribution density of chain length in ALPHA, BETA, and ALPHA/BETA sets, in blue (dark gray), orange (light gray), and red (median gray), respectively. The mean and standard deviation in ALPHA was 210 ± 125 residues, with min/max of 61/522 residues, totalizing 19163 residues. The mean and standard deviation in BETA was 221 ± 122 residues, with min/max of 59/534 residues, totalizing 20,127 residues. The mean and standard deviation in ALPHA/BETA was 230 ± 141 residues, with min/max of 51/587 residues, totalizing 20,894 residues. All sets had very similar chain length distributions. The homogeneity of the center tendencies for the three sets was calculated by nonparametric Kruskal-Wallis[86] test, giving *P*-value of 0.63. Kolmogorov-Smirnov[80] test assured the goodness-of-fit between ALPHA and BETA distributions with *P*-value of 0.31, ALPHA and ALPHA/BETA distribution with *P*-value of 0.41, BETA and ALPHA/BETA with *P*-value of 0.87. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

The volume and surface was calculated by Gerstein programs[59,60] "calc-surface" and "calc-volume," with the traditional Richard's method B[1] and also adopting Richard's radii.

Contacts

The definition of contact that we have adopted here was essentially geometric and not energetic. As aforementioned, we are working only with contacts, not with interactions. It involved the determination of a set of neighboring points belonging to a space around one given referential point or centroid, with the Euclidean distances used as a weight. All contact data were defined at the residue level, but a low-resolution approach was adopted: each residue was characterized by one centroid. We compared two types of referential points: by alpha carbons (CA) and by side chain geometrical centers (GC). The latter was reduced to the coordinates of alpha carbon in the case of Glycine.

Cutoff dependent contacts—traditional approach

We implemented and then analyzed the CD traditional approach: a contact is defined between any given pair of residues $\{i,j\}$ if the Euclidean distance between their centroids was less than or equal to one arbitrary cutoff distance. For a more mathematical description, see the Methodologies section at the Supplementary Material.

Cutoff free contacts—Voronoi diagrams and Delaunay tessellation

The Voronoi diagram (VD) is a geometric construct named in honor to the Russian mathematician Georgy Voronoi (1868–1908) who employed in 1908 the n-dimensional case.[43] The basic ideas in low dimensions can be traced back to works of Dirichlet,[62] Gauss,[63] and Descartes.[64] See also references 65 and 66.

In 3D, VD will partition the volume associating a polyhedron to each site, which is called a Voronoi cell. Each face of these polyhedrons will comprise planes that bisect the line linking a site to each of its near sites, mapping a neighborhood with the closest contacts. For a more mathematical description, see Aurenhamer[68] in the Methodologies section of the Supplementary Material.

One construct related to VD is the DT[44]. Voronoi in his celebrated paper[43] of 1908 had already realized that the dual graph of his diagrams on lattices seem to have important characteristics. A dual graph is obtained by assigning a vertex to each region of the target graph and making vertices links if and only if those regions share an edge[61] [Fig. S1(g) at Supplementary Material]. Another Russian, Boris Delaunay (1890–1980), extended the original work of Voronoi from lattices to irregularly placed sites through an ingenious method: in 3D, four

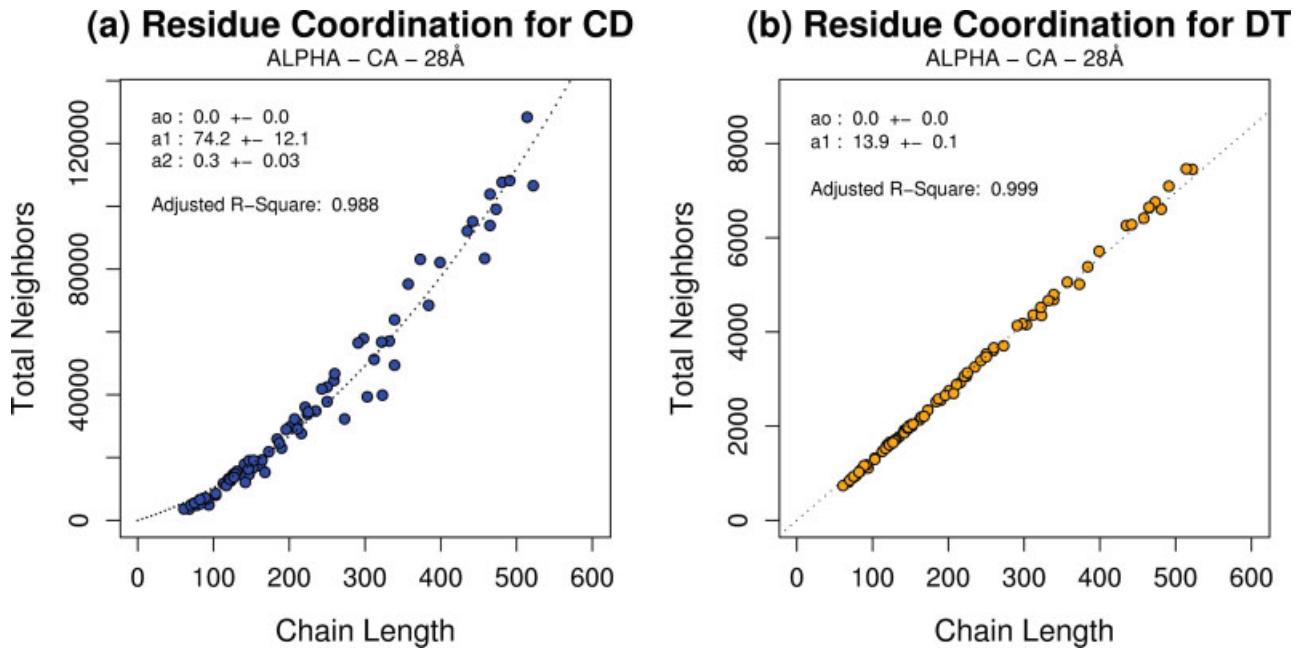

Figure 2

Cumulative coordination as a function of the number of residues (chain length) in 91 proteins of ALPHA CA set. The ordinate represents the total neighbors in a sphere of 28.0 Å, centered at CA. The numbers at the upper left corner indicate the fitting curve coefficients with corresponding standard errors. For both regressions, the fitting curve was forced to pass through zero. (**a**) Coordination for CD contacts is well fitted by a parabola. (**b**) Coordination for DT contacts is linear. The slope indicates the average coordination number in 28.0 Å as 13.9 ± 0.1 neighbors per residue in ALPHA CA, at 0.95 level of confidence. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

sites compose a Delaunay tetrahedrization if and only if the circumsphere of each polyhedron is empty of other sites. Applying this algorithm over all sites will tessellate a volume by tetrahedrization, with the remarkable property that only the nearest neighbor to each point will be connected by an edge.

Here we used the program ADCGAL[69] to compute DT and almost-Delaunay (AD) contacts. We have also used the same equations (vide the Methodologies section at the Supplementary Material) than those utilized in the CD methodology to compose the DT/AD contacts statistics, but considering only the edges returned by the latter.

RESULTS AND DISCUSSION

First, we would like to discuss comparatively some of the main contact properties and intrinsic problems of the two methodologies: CD and cutoff free DT.

Cutoff dependent properties

One of the greatest advantages of the traditional CD method is that it makes a complete scanning of all combinations for pairwise edges inside the search spheres built by cutoff (r) and centered in each site. So, if there are n vertices or sites delimited by these spheres, there are $C(n,2)$ edges combinations counted by CD, what gives $O(n^2)$ contacts [Fig. 2(a)]. In a volume with near uniformly distributed points the number of sites raises with $O(r^3)$. Hence, the number of contacts as a function of cutoff will be of $O(r^6)$. This is not an exponential, but it is a polynomial of high order, and it is expected that the number of contacts may grow vigorously with r (although we have evidences that the coefficients for the highest exponents may be very small, data not shown). This is an estimate for infinity condition without border limits. In well-packed proteins, insofar as r increases, the search spheres extrapolate all coordination shells until reaching the last atoms in the boundary region. Sites more buried feel this border effect in larger cutoff values than sites near the surface, but the general effect is that the increase in the contact number will be contained. The result is a sigmoidal cumulative distribution asymptotic to $C(n,2)$. Naturally, given the symmetry of the sigmoidal curves, its density distribution (first derivative) tend to be Gaussian like but only in the shape, because it continues to be a polynomial now with $O(r^5)$.

Emphasizing, the best quality of the CD method is that it is exhaustive or total. It enumerates all contacts that can exist for a given cutoff. This seems ideal for a global vision of the packing, but it is not so efficient for an unambiguous analysis of local arrangements. If the goal is to count

the contacts only in the first coordinate shell (the first layer of neighbors), for example, CD does not offer an easy way of prospecting it without the risk of confronting with both false-positive (occluded and counted) and false-negative (not-occluded and not-counted) contacts. The usual approach is to try to find a statistically optimum cutoff that minimizes the possibility of this risk, observing the contact density behavior along the cutoff.

Manavalan and Ponnuswamy[29] were likely to be the first to endeavor in this direction. They found that hydrophobic residues represented by CA atoms in a set of 14 proteins were maximally clustered between 6 and 8 Å of cutoff, suggesting 8 Å as an ideal value. Afterward, Miyazawa and Jernigan[28] looking at the radial density profiles of interior residues, whose location was up to 7 Å from the protein GC, evinced a peak at the shell between 5.0 and 5.5 Å, with a subsequent valley about 6.5 Å. The latter was elected as the ideal cutoff for the empirical potential analysis that they were carrying out using a special set of 42 proteins. We cannot leave unnoticed that the residues in their model were reduced to GC of side chain atoms (CA for Glycine). Following a similar methodology, Zhang *et al.*[30] estimated their best cutoff as 6.0 Å, but computing the residue contacts by the distances between their heavy atoms, from a selected collection of 89 proteins. Furuichi and Koehl[31] tried to find the ideal cutoff for a set of 125 nonhomologous proteins contrasting the effects of the chain lengths in the contact profiles. Using a CA residue representation, they built a reduced database with 68 proteins, dividing it into two subsets S (Short) and L (Long), where S had proteins with less than 130 residues and L the remaining ones. The first interesting fact that they identified was that the two distributions were reasonably similar up to 10.0 Å, indicating that short range interactions were independent of protein size. Other fact is that they have observed that the predictive power of their empirical potential in the two sets S and L diverged about 7.0–8.0 Å. In face of this, they have suggested the ideal cutoff as 8.0 Å. More recently, Kamagata and Kuwajima[32] introduced an experimental ground in the optimum cutoff definition. They verified a surprisingly high correlation between the number of contact clusters (*Nc*) and the log of intermediate rate constants in the kinetics analysis of folding from 12 non-two-state proteins. In the cluster definition, they used a fine-grained approach establishing that two residues are in contact if any of their heavy atoms were close enough. They tested the influence of the variation of cutoff on the changing of linear correlation coefficients, and found that no meaningful alteration was found for cutoff beyond 5.5 Å.

Delaunay tessellation properties

Cutoff free techniques like DT can reduce the neighborhood ambiguity problem found in CD as a consequence of their exact geometric definition. For most cases, there is a mathematical guarantee that contacts in DT are occluding-free, because the bisection or separators in VD are constructed between the closest pair of sites, which practically vanishes the DT edges crossing intermediate sites. Exceptions could occur for sites near the surface, with a region without neighbors. In certain configurations, edges can be traced between those sites that were in a quasi-linear condition and stay partially occluded, with the respective center of the common sphere near infinity. But these cases tend to be rare and may be circumvented by some usual procedures like solvation.[1] It is precisely that geometric appeal and apparent non-ambiguity some of the reasons that VD/DT have been largely used as a efficient method to identify nearest contacts in proteins.[46] And the method seems to be better suited specially when the target is the first layer of neighbors, the major difficulty of CD techniques.

There is an interesting property of DT when the number of edges is analyzed by distances for sites in nearly uniform distribution: regardless of the dimension, this number tends to grow in $O(n)$ for the average case.[70] In proteins, we could verify that this was really true. In Figure 2(b) we show a plot with 91 ALPHA CA protein set, clearly indicating a high linear correlation, in the range of up to 28 Å, between the number of neighbors and the number of residues (chain length). This linearity indicated that DT captured a limited number of neighbors in the vicinity of each residue, and made the number of contacts to be scaled up to the size of the protein.

As a consequence, its slope contained worthy information about the protein packing characteristics, indicating the average number of neighbors per residue at a given distance. For our ALPHA CA data, in 28.0 Å, we see that this average was 13.9 $\pm$ 0.1 neighbors [Fig. 2(b)]. The calculated mean for the other sets was (using the same cutoff): BETA CA: 14.0 $\pm$ 0.1, ALPHA GC: 13.4 $\pm$ 0.1, BETA GC: 13.6 $\pm$ 0.1, ALPHA/BETA CA: 14.1 $\pm$ 0.1, ALPHA/BETA GC: 13.6 $\pm$ 0.1 neighbors, all at 0.95 level of confidence. Our results were very close to the value 13.97 found by Soyer *et al.*[52] for the average number of faces in Voronoi polyhedrons in a collection of 40 proteins, where residues were represented by the side chain barycenter.

The lines in both Figure 2(a,b) were forced to intercept zero, given that it is reasonable to assume that with zero residues it would expect to have zero neighbors. But, it was possible to note that small proteins with few residues tended to stay slightly below the fitting curve, in both CD and DT plots. Probably, this happened because for high cutoff value the contribution to the neighbor counting for shorter proteins was smaller than for larger proteins. For CD, the search spheres reach the size limit of the small proteins faster as the cutoff value grows, exhausting its capacity to contribute with more edges. For DT, at high cutoff, the contribution may be given by

sites in the surface of proteins, which also were fewer in small proteins. Hence, the average number of neighbors in extended cutoff has a bias dictated by the amount of small and large proteins present into the database. As noted by Furuichi and Koehl,[31] this naturally imposes an upper limit in the cutoff to be used, that is, the select cutoff value must not be larger than the average radii of the smallest proteins in the database.

To see how this fact may affect our data, we estimated the Voronoi volume of our proteins, using the cited Gerstein *et al.* programs.[59,60] We approximated these volumes as spheres, to make a rude estimation of the average radii for our data mart proteins. This returned the following basic statistics: modal = 16 Å, mean and median = 18 Å, min = 10 Å, and max = 27 Å. This alerted us that the cutoff of 28.0 Å used before was too large to be reliable. The volume statistic above suggested that our cutoff should not go beyond the minimal estimative of 10 Å, constituting it an upper bound limit.

However, DT is not all problem-free. It is known that DT is not robust to the noise in the location of the points. There are situations where small movement of the centroids may lead to substantially different arrangement on the simplices (Fig. S2, Supplementary Material and Refs. 82–85). DT in 3D requires, for a complete tetrahedrization, that all points should stay in general position, for example, that no five points are cospherical, because five points in a sphere admit five tetrahedra that satisfy the empty-sphere criterion. Sites whose coordinates are near these situations are sensible for DT to small change in their position. The edges may flip with minimal movements of the centroids, changing the pattern of contacts. Note that CD also shares this problem, but only for sites in the cutoff frontier zone; for the remaining sites, CD is insensible to this difficulty.

A real example illustrates how serious this DT intrinsic missing edges problem may be in protein. In Figure 3 are shown in yellow (or light gray) all contacts found by DT around the residue ILE 167 (in light blue or median gray) from the all beta Endopolygalacturonase (1K5C),[71] solved at 0.96 Å resolution. All residues were represented by their alpha carbons (CA). The DT correctly identified 10 neighbors of ILE 167, but ignored two other legitimate presences: the ASN 188 and CYS 190 in violet (or dark gray). This happened because the four residues CYS-166, ASN-188, GLN-189, and ILE-167 are near the degenerated case, that is, they are almost in a cospherical condition, and as the contact {CYS-166, GLN-189} was closer than the pair {ILE-167, ASP-188}, DT traced the former contact. There was a clear symmetry in the first order contacts of ILE 167 that DT seemed not be able to encompass fully. This error may be systematic, with DT tending to omit for many residues in one strand, one or two edges with residues in the companion strands. For example, the {THR-136, VAL-168} contact was also ignored because of a difference of 0.14 Å with the contact

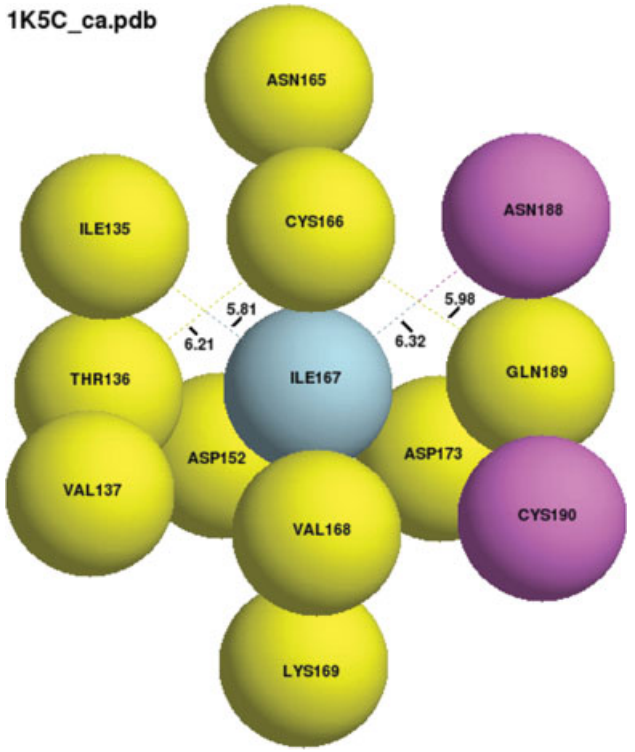

Figure 3

The first order contacts in the vicinity of residue ILE 167 in light blue (or median gray) from the all beta Endopolygalacturonase 1K5C.[71] Each residue is represented here by its alpha carbon (CA) in CPK model. In yellow (or light gray), the 10 residues that DT made an edge with ILE 167 are shown. In violet (or dark gray), are presented the residues that DT was unable to recognize as neighbors. The dashed line and respective numbers indicate the distances in Å.

{ILE-167, ILE-135}. Therefore, DT was not so absolutely free of ambiguity as it is believed and often reported in literature.[52,46]

Bandyopadhyay and Snoeyink[67] have tried to address this problem by what they called AD simplices. Given a set of sites S in R^2, $Q \subset S$ points will comprise a set of AD simplices AD(ϵ) if and only if by perturbing each site of S up to a finite threshold ϵ the altered Q has an empty circumscribing circle; see Figure S3 (Supplementary Material) for more details. The AD actually seemed to be able to identify missing edges not detected by traditional DT near the degenerate position. It, for example, detected the {ILE-167,ASP-188} and {ILE-167,CYS- 190} contacts in the case cited earlier. We will return to this question but before, we have to see how DT relates to CD.

Confronting cutoff dependent and Delaunay tessellation

After describing above some of the idiosyncrasies of both methodologies, we will now compare the way how CD and DT prospect contacts as a function of the cutoff.

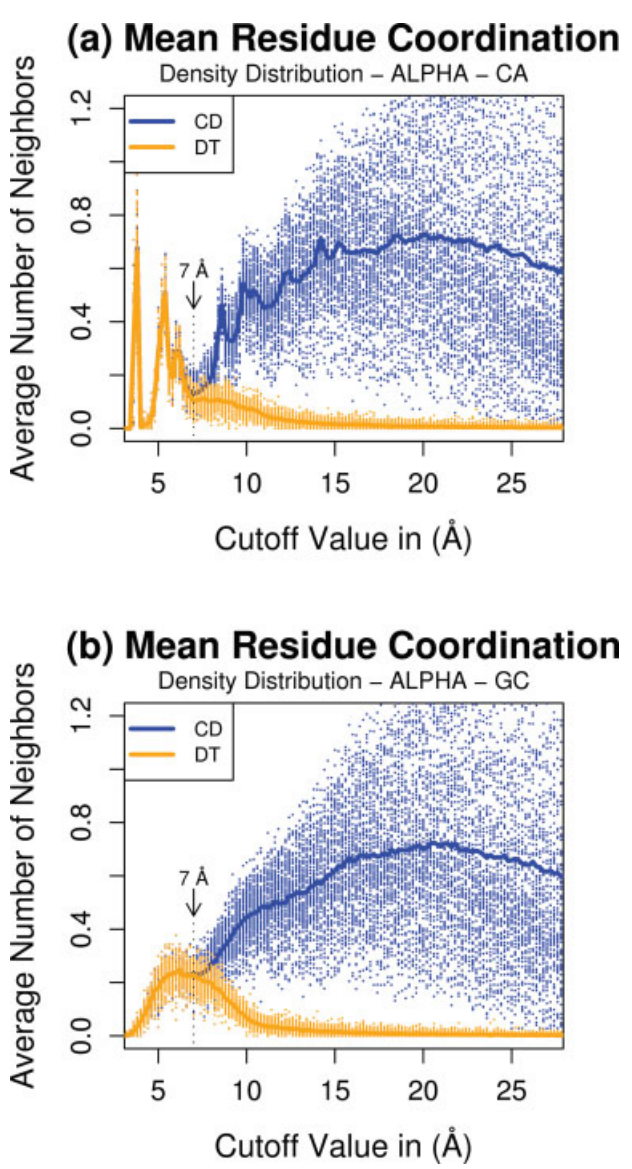

Figure 4

Comparative density distributions between the mean residue coordination of CD and DT methodologies as a function of cutoff for ALPHA set. The thick continuous line in blue (or dark gray) and orange (or light gray) denotes the mean number of neighbors for CD and DT models, respectively. The scatter data for all the sampling points are shown offering a complete overview of the behavior and variance of the data. (**a**) Curve patterns for ALPHA set with alpha carbon (CA) residue representation. 7.0 Å is a bifurcation point between CD and DT. (**b**) The distribution profiles with side chain GC. We see the same bifurcation point at 7.0 Å. For both CA and GC residue representation, the CD data present a large variability beyond 7.0 Å. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Figure 4 shows the density distribution of the number of neighbors per residue in the range of cutoff 0.0–28.0 Å for both CD and DT, in ALPHA set. At cutoff value about 7.0 Å, the CD profile bifurcated from DT profile. Below 7.0 Å, the distributions were in essence the same, independent of the residue representation. Above 7.0 Å, the CD exploded with great variability, mainly as a consequence of the diversity in the protein sizes. Note that this impressive divergence is occurring with the data normalized by the number of residues. This may indicate the existence of combinatory process in the edges enumerations that seems to be extremely sensible to the number of sites. It is important to remember that the number of edges by cutoff may grow with a polynomial of high degree.

The same data for BETA set are shown in Figure 5. Although the bifurcation point continued to be about 7.0 Å, the differences seem to have initiated at a lower cutoff value, around 6.2 Å. Note that in the interval 6.2 up to 7.2 Å, despite the differences, both curves were still correlated. Certainly, the anticipation (from 7.0 to 6.2 Å) occurred as a consequence of the DT problem already described earlier and illustrated in Figure 3.

Figure 6 shows the graphics for ALPHA/BETA set. We also see the presence of DT missing edges problem, as expected by its relatively higher beta content. The bifurcation point also seems to occur between the intervals 6.2 up to 7.2 Å.

Computing the difference between the mean curve areas of CD and DT up to 7.0 Å, it was possible to estimate the percentage of edges not considered by DT. The average error was about 5.1% $\pm$ 0.3% for BETA CA, 0.6% $\pm$ 0.2% for BETA GC, 1.9% $\pm$ 0.3% for ALPHA CA, 0.3% $\pm$ 0.2% for ALPHA GC, 3.3% $\pm$ 0.3% for ALPHA/BETA CA, and 0.4% $\pm$ 0.2% for ALPHA/BETA CG, all at 0.95 level of confidence. In spite of being low, it is important to remember that this error may not be random. Note also that it will affect mainly proteins rich in beta structures due the almost flat topology of its strands, which may put four sites in a cocircular condition.

In our point of view, the fact that these distributions were in essence the same up to 7.0 Å had two possible outcomes. The first was that it unified the CD and DT properties: all edges up to 7.0 Å will be complete, enumerating combinatorially all possible contacts that can exist inside the searching sphere (a CD property); and all edges will also be legitimate contacts, with a geometrical guarantee to be completely free of occlusions (a DT property). This, together with the fact that the results were independent of the protein classes analyzed and also independent of CA or GC residue representations, made 7.0 Å a candidate to a referential lower bound cutoff limit to be used in contacts definitions applications that want to adopt the coarse-grained models used here. Although this lower bound is independent of the quality of the database, the upper bound is dependent of the protein size distribution. So, for our databases an ideal cutoff might be found in the range of 7.0–10.0 Å. However, for CD, at 7.0 Å there is a certain guarantee that contacts will be occlusion-free, but above this the chances of getting false-contacts will increase.

The second outcome was that, as a consequence of the single perspective, up to the bifurcation limit CD should

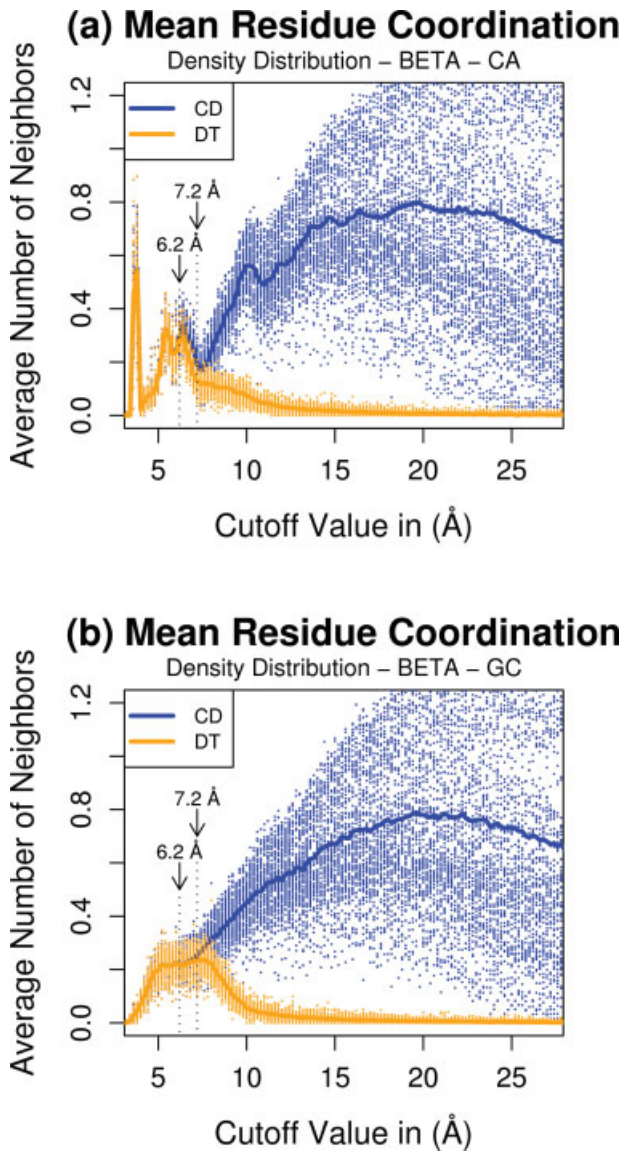

Figure 5

Comparative density distributions between the mean residue coordination of CD and DT methodologies as a function of cutoff for BETA set. The thick continuous line in blue (or dark gray) and orange (or light gray) denote the mean number of neighbors for CD and DT sets, respectively. The scatter data for all the sampling points are shown offering a complete overview of the behavior and variance of the data. (**a**) Curve patterns for BETA set with alpha carbon (CA) residue representation. Because of the DT missing edges problem the sites near the degenerate state are not recognized and the separation point is anticipated to about 6.2 Å. But a bifurcation between CD and DT seems to occur between 6.2 and 7.2 Å. (**b**) Distribution profiles with side chain GC residue representation. We see the same interval where bifurcation point occurs: between 6.2 and 7.2 Å. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

also inherit the linearity behavior of DT. As CD is quadratic by nature, it is expected that it makes a transition from parabolic to linear model, staying essentially linear around 7.0 Å. To check this possibility, we conducted a model selection test using the Bayesian Information Criterion (BIC)[72] to evaluate if linear or quadratic behavior was statistically more adequate to fit our CD data in each range of cutoff. BIC returns a number that measures the fitting quality of the model to the data. The lower

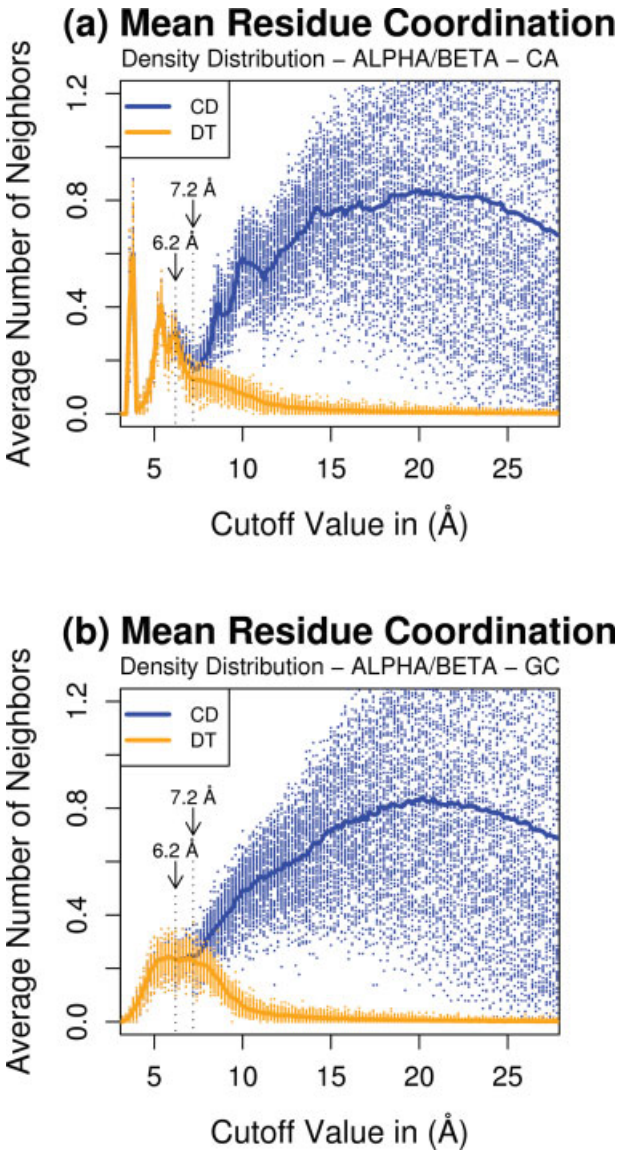

Figure 6

Comparative density distributions between the mean residue coordination of CD and DT methodologies as a function of cutoff for ALPHA/BETA set. The thick continuous line in blue (or dark gray) and orange (or light gray) denote the mean number of neighbors for CD and DT sets, respectively. The scatter data for all the sampling points are shown offering a complete overview of the behavior and variance of the data. (**a**) Curve patterns for ALPHA/BETA set with alpha carbon (CA) residue representation. Like in BETA set, the DT missing edges problem is also present but in minor degree. A bifurcation between CD and DT seems to occur also between 6.2 and 7.2 Å. (**b**) Distribution profiles with side chain GC residue representation. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

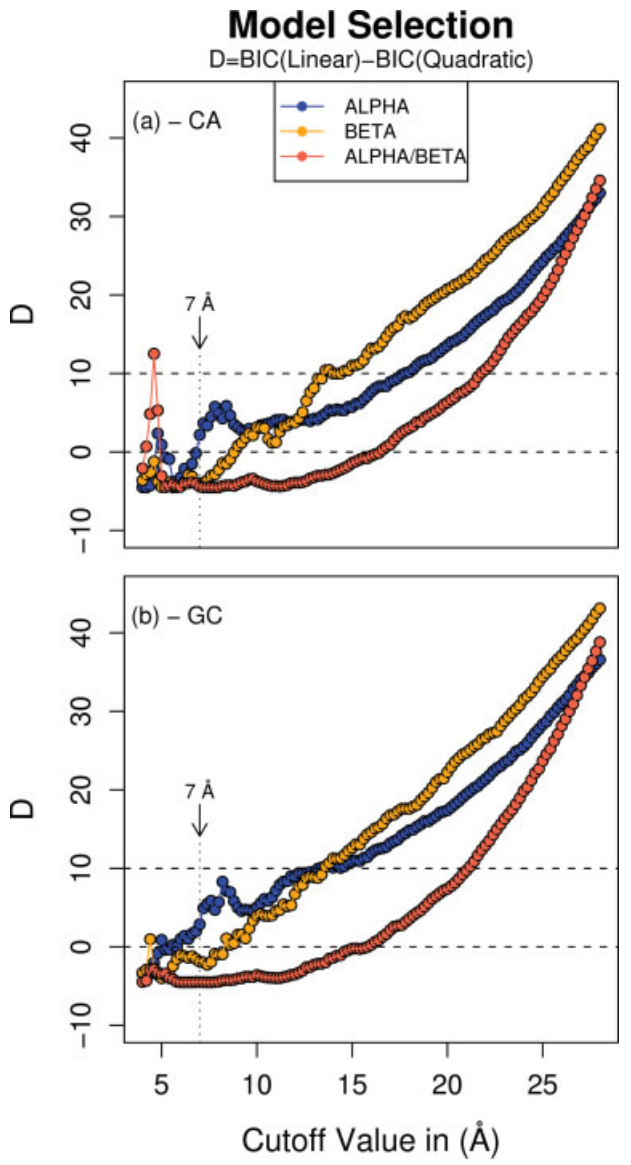

Figure 7

Linear against quadratic model selection test using Bayesian Information Criterion[72] (BIC) for each cutoff range. The ordinate contains the difference between BIC numbers, D = BIC(Linear) − BIC(Quadratic). The smaller the BIC number the more adequate is the model. Hence, positive D values favor the quadratic model. (**a**) Model selection test for alpha carbon (CA) residue representation. For high cutoff, a quadratic version is favored. Insofar as the cutoff goes down, the model was switching to a linear version. At 7.0 Å, D is near zero for ALPHA, BETA, and ALPHA/BETA sets. ALPHA/BETA has an intriguing behavior. Unlike ALPHA and BETA pattern, ALPHA/BETA decays to a linear mode (D = 0) more quickly, at cutoff around 15 Å. (**b**) The same as in (a) applied to the side chain GC residue representation. Again, the behavior is similar as for the CA representation. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

this number, the better is the model. But, as BIC absolute values are meaningless, we computed the relative difference D = BIC(linear) − BIC(quadratic). Positive values for D indicate that the quadratic model is superior, and vice-versa. Burnham and Anderson,[73] as a rule of thumb, suggested $D \leq 10$ as upper limit to assessing the relative merits of the models, recommending $D \leq 2$ as a suited value. Summarizing, the closer D is from zero, more indistinct are the models, and we should select those models with fewer parameters (if we accept the Occam's razor criterion). Figure 7 presents the results of the BIC test. We can see that, for high cutoff, the better model was in fact quadratic. From that same figure one can clearly see that as the cutoff values decreases, a quadratic dependence property was lost and a linear form prevailed. At 7.0 Å, all curves were either near or below zero. Curiously, for high cutoff values both ALPHA and BETA seemed to have the same decay rate, with the former being closer to linear than the latter. In a point near to 20.0 Å, ALPHA decreased its decay rate while beta continued practically unchanged. As a consequence, BETA data reached $D = 0$ earlier than ALPHA. Another surprising result was the profile of ALPHA/BETA set. Mysteriously, it reaches $D = 0$ in a high cutoff value about 15 Å. It is not trivial to try to explain this strange behavior, which seems to touch subtle topological differences between the spatial distributions of centroids in the sets analyzed. Further investigation is necessary to clarify this point.

As noted by Bandyopadhyay and Snoeyink through their lemma 4.1,[69] this linear behavior may be a result of the nearly uniform packing of proteins. In our view, it constitutes also a topological signature of the first coordination shell. As a consequence, the cutoff value at 7.0 Å may be understood as the ideal distance from where the first-order contact is optimally separated from other higher-order contacts in proteins. It is the best point that put in evidence the immediate layer of neighbors. At this moment, as we cannot formally prove this linear statement of the first-order contact, we will only conjecture that it is true based on all evidences related herein.

Is almost Delaunay a solution?

We are now in condition to assess the Bandyopadhyay and Snoeyink solution[67] to solve the DT missing edges problem. In Figure 8, we see the plot of AD (using threshold perturbation value of 2.0 Å and prune of edges of 28.0 Å) against the CD and DT, for BETA with CA representation. Interestingly, we perceived that AD was the difference between CD and DT. This meant that, as threshold perturbation approaches to infinity, the DT+AD approaches to CD in the range of cutoff considered. If this is true, there is no apparent reason to prefer the solution DT+AD in favor of CD when an ideal cutoff is used. However, one of the advantages of AD was that with it we could determine the more precise moment where DT initiates the divergence from CD. We considered this point as that where half of proteins had at least one AD edge. The resulting values were: ALPHA CA: 6.2 Å, ALPHA GC: 7.0 Å, BETA CA: 6.2 Å, BETA GC:

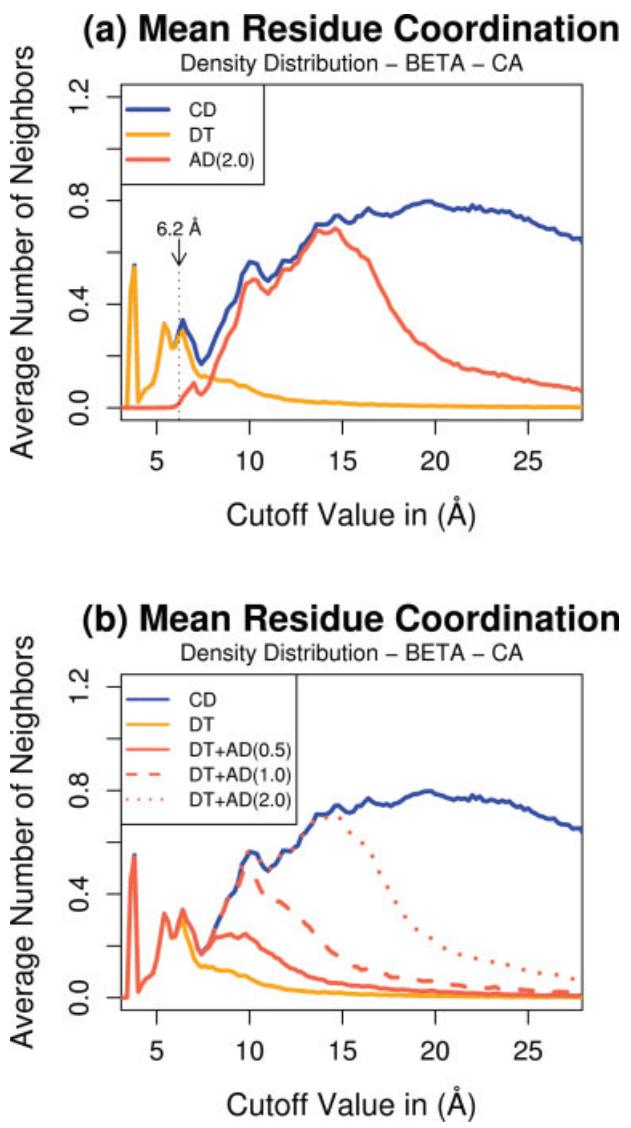

Figure 8

Comparison of the average number of neighbors per cutoff values between AD solution with CD and DT. The thick continuous line in blue (or dark gray), orange (or light gray), or light red (or median gray) denote these averages for CD, DT, and AD models, respectively. The scatter data for all the sampling points are not shown here, only the average number of neighbors. In all AD the prune parameter was of 28.0 Å. These data were collected from BETA set with alpha carbon (CA) residue representation. The number in parenthesis in AD indicates the threshold perturbation parameter. (**a**) We plotted the AD with perturbation threshold of 2 Å against CD and DT. We see that AD appears to be a complement of DT in cutoff less than 15.0 Å. We can see also that AD is different of zero in a cutoff value about 6.2 Å. (**b**) Plot with CD, DT, and the sum of DT + AD with perturbation thresholds of 0.5 Å, 1.0 Å, and 2.0 Å. We see that the sum of DT + AD tend to yield results very similar to CD as the perturbation threshold grows. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

6.8 Å, ALPHA/BETA CA: 6.2 Å, ALPHA/BETA CG: 7.0 Å. This confirmed that the DT missing edges problem occurs mainly when CA residue representation is used.

A case study: a neighborhood analysis in proteins

Finally, we implemented a case study to analyze the influence of the residue coarse-grained centroids CA and GC on the neighborhood profiles. To this point we already demonstrated that CD and DT are unified up to 7.0 Å, and that we may use the simpler CD technique. We built a cumulative scatter plot highlighted by the average number of neighbors, in the cutoff range from 4.0 to 9.0 Å (see Fig. 9). For CA representation, we could delimit at least three regions, separated by cutoffs 5.2 Å and 6.8 Å. The cumulative CA curves seemed to have different behaviors before and after this marked distances, in ALPHA, BETA, and ALPHA/BETA sets. ALPHA/BETA appears to inherit the pattern of ALPHA before 6.8 Å and of BETA after 6.8 Å. For GC representation, the profile was more homogeneous. At about 6.8 Å, like in CA representation, the two mean curves of ALPHA and BETA sets stayed very close, indicating a convergence point. ALPHA/BETA, on the other hand, seems to follows closely the BETA pattern, but always assuming high values, in the average.

To have a more reliable idea of the real differences between our ALPHA and BETA sets, we decided to explore the behavior of the central tendencies along the distance ranges. For this, we applied both the parametric Student *t*-test with Welsh[74] correction for differences in variances, and non-parametric Wilcoxon *t*-test[75] in our cumulative distribution of neighbors (Fig. S4 at Supplementary Material). We also performed a more robust nonparametric test as a form of ensuring the *t*-test results, which is known to require normality as a precondition. Indeed, it was notable that the two tests produced well correlated curves in most of the regions. Low *P*-values indicated that the difference between the average number of neighbors in the sets was statistically significant. For CA residue representation, in the ALPHA × BETA, ALPHA × ALPHA/BETA, and BETA × ALPHA/BETA comparisons, we found two sharp peaks with high *P*-values in favor of indistinctness of the means/medians, at about 5.2 Å and 6.8 Å, the same interception points found in Figure 8(a). With GC, in the ALPHA × BETA comparison, all *P*-value tended to remain below the arbitrary threshold of significance, except for a visible broader peak around also 6.8 Å. In the ALPHA × ALPHA/BETA comparison, as expected, any significant similarity between the means/medians is found, except in very lower cutoff values. In the BETA × ALPHA/BETA comparison, we see that the central tendencies had more homogeneity between 5.0 Å and 6.0 Å, and after 7.2 Å.

Looking conjointly these data we noted that, between the cutoff 5.2 Å and 6.8 Å, CA and GC residue representations do not agree whether ALPHA or BETA had more average neighbors per residue. The former seemed to bias the result in favor of ALPHA and the latter in favor of BETA. If two different researchers had chosen one of

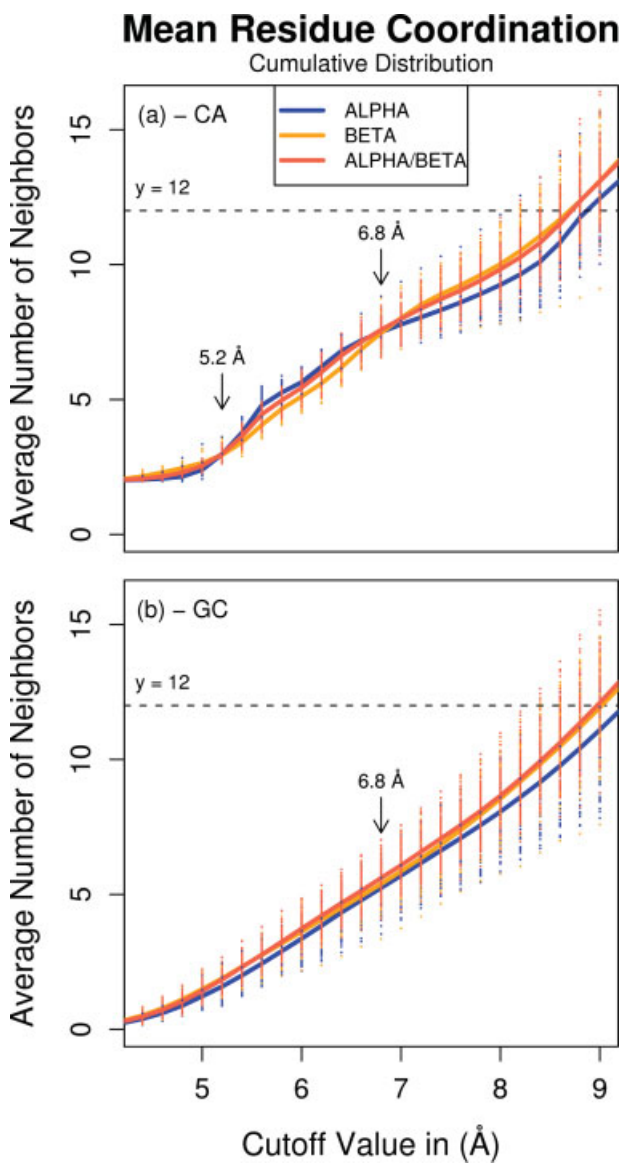

Figure 9

Cumulative distribution of the average neighbors per residue as a function of the cutoff using CD data. The thick continuous line in blue (or dark gray) orange (or light gray) and red (or median gray) denotes the mean coordination for ALPHA, BETA, and ALPHA/BETA sets, respectively. The scatter data for all the sampling points are shown offering a complete overview of the behavior and variance of the data. The dashed line traces the 12 neighbor limit, the characteristic number of the close packing of identical spheres. (**a**) The average coordination with alpha carbon (CA) residue representation. The arrows indicated some regions of convergence between the mean curves, comprehending the cutoff values at 5.2 Å and 6.8 Å. ALPHA/BETA tend to follow ALPHA before 6.8 Å and to follow BETA after 6.8 Å. (**b**) The average coordination with side chain GC residue representation shown the same convergence region of 6.8 Å as in CA for ALPHA and BETA. ALPHA/BETA tends to stay always over BETA pattern. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

these residue representations, and if they had used a cutoff below 6.8 Å, they might reach different conclusions about the neighborhood number profile in alpha and beta proteins. ALPHA/BETA would contribute to distort the data in favor of alpha if had used a cutoff below 6.8 Å, and in favor of beta if had used a cutoff above 6.8 Å. As a consequence, this difference in the number of neighbors might affect, for example, packing inferences, inducing contradictory results about if alpha or beta proteins are more compacts. Above 6.8 Å, this ambiguity is minimized, with both CA and GC residues representation at least agreeing about the statistically larger average number of neighbors in beta than in alpha. Curiously, our CD/DT unifier cutoff number 7.0 Å stays in a region less vulnerable to these problems of contradiction. Therefore, this constitutes an additional argument in benefice of the 7.0 Å as a lower bound cutoff for contact analysis of proteins.

At last, we need to point out two of what we consider as most important remarks. First, that we had established here a candidate value to lower bound cutoff for prospect contacts in proteins. As the upper bound cutoff is dependent of the protein sizes in database, the ideal cutoff value will be as well. Second, 7.0 Å is a lower bound cutoff if the research goal is as complete enumeration of the first-order contacts as possible, that is, the reliable characterization of the first layer of neighbors. If the interest, on the other hand, is to map only the nearest interactions, like for example those closer than 4.0 Å, a lower bound cutoff value will obviously not be applicable.

Limitations and perspectives

The work presented here carries intrinsically some limitations and it does have equally many perspectives that could be considered as complements, because the former may be a motivation for the latter. So, we will comment both conjointly.

Generating a database that contains appropriate structures, selected according to well designed filtering procedure is a general problem in practically any bioinformatics work. In this article, we choose to respect the skewed protein size distribution returned from the PDB filtering for ALPHA and BETA set, but we forced the ALPHA/BETA distribution to follow the formers. We are confident that the results raised by our investigation for cutoff values up to 7.0 Å is protein size independent, as long as this value stays in a region below the 10.0 Å limit found by Furuichi and Koehl.[31] Moreover, the smallest proteins in the data set we created had also 10.0 Å of approximated radius.

Other coarse-grained forms to represent residues is also a future research target, like the beta carbon (CB) representation; or even more fine-grained ways, like the residue contacts prospected by the closer heavy atoms. A deep fine-grained CD/DT model in atom level will be computationally challenging but certainly welcome complementation to our results.

The definition of surface in the context of this work is particularly tough. It is known that sites on surface can yield open Voronoi cells, with infinite volumes. Certainly, this is a problem for some algorithms like those performing calculations on both the volume and the area of a protein. Fortunately, this does not preclude the Delaunay decomposition, which requires only that the circumsphere of each tetrahedron is empty of other sites. But, sites on the protein surface will be free to compose edges with any other site in the same condition, creating an unreal number of neighbors by residues. These edges may not be informative with respect to structure and may introduce dangerous bias on contact statistics, as demonstrated by our analysis at large cutoff of 28.0 Å. Two forms to deal with this question are commonly used. One way requires removing any DT edges and tetrahedrons that are not completely inside of the overlapped volumes of the sites. This subset of DT is frequently called the "α-shape" of the molecule.[4,89] The question here is how to choice pertinent virtual volumes for sites in coarse-grained models? The second well-known way demands the introduction of fictitious surface solvent molecules.[1] But, this solution comes with certain arbitrariness and must be made carefully. We consciously opt in the present study do not treat protein surfaces because preliminary tests using the well-elaborated surface solution of VORO3D program[77] have given us strong evidence that the solvation should have a prominent effect in DT only in edges beyond 7.0 Å (vide Fig. S5 at Supplementary Material). In fact, this comprised yet another support to the hypothesis that the edges distribution up to 7.0 Å was generated mainly by buried sites of the first coordination shell.

Another question that remains open is whether the number of neighbors in a given cutoff can be used as a way to measure the local packing density in proteins. The number of neighbors will depend in a complex way not only of the packing, but also of the size and form of the sites. Fleming and Richards,[2] describing some packing properties of 152 nonhomologous proteins, and using the OSP metric have shown that helices seem to be more efficiently packed than strands. They also found indications that aromatic residues tend to be better packed than aliphatic ones. Liang and Dill[4] found that larger proteins tend to be more loosely packed than smaller proteins. Angelov *et al.*,[78] analyzing the Voronoi properties in a collection of 39 proteins, have shown that there is a tendency of positive linear correlation between the number of faces per cells (neighbors) and residue Voronoi volumes, although glycine, alanine, lysine, and arginine stay outlines. As expected, glycine and tryptophan were the extremes, with the first having in average 13.36 neighbors and the second 14.86 neighbors, a difference of 1.50 neighbors. As the authors did not give the standard errors, we cannot judge the statistical relevance of these differences, which seems to be very low. While Kuntz and Grippen[79] found inhomogeneities in local density primarily related to differences in the clusters of nonpolar side chains and backbone secondary structures, Tsai *et al.*[3] demonstrated that if surface waters are included in calculations, the overall packing became high and fairly uniform. But, in spite of this complexity, perhaps we may give an alternative interpretation to data of Figure 9. If we a priori assume (in a Bayesian fashion) that the conclusions of Fleming and Richards[2] are true and that the difference in packing among the residues is not significant in our data, and if we also look at the CA and GC residue representations as signatures of backbone and side chain contributions, respectively, perhaps it may be possible to conjecture that the tendency of alpha proteins to be more compact than beta proteins comes more from the backbone.

We are also finalizing a work where very stringent filtering conditions are applied in order to obtain as large as possible data mart containing protein structures with ONLY alpha helices or beta strands, and then conduct a statistical analysis (including multivariate) of 43 parameters from STING_DB, all in order to distinguish what in fact is determining the packing characteristics of those two major secondary structure elements, contributors to the protein fold.

CONCLUSIONS

We have scrutinized several and different questions related to the methodologies commonly used for prospecting protein contacts, producing some intriguing results. Primarily, we came up with the discovery of the cutoff number at about 7.0 Å as an important distance parameter in analysis of contact in proteins. We saw that, at this distance the CD and DT results converge, what allowed us to unify the main properties of both techniques: that all pairwise contacts are complete and true-positive (counted and not occluded). We believe that these characteristics comprehend a topological signature of the first coordination shell and 7.0 Å is the lower bound distance number that delimits it from other highest order layers of neighbors. This was also corroborated by some preliminary tests applying radius distribution function to our data, where we found that 7.0 Å is always near a valley, independent of the protein class and residue representation (data not shown). It was striking too that at this distance we could not distinguish alpha carbons (CA) from side chain GC residue representations. This unification of CD/DT also affirmed the linear condition of the first order contacts and we noted that there is a passage from linear to quadratic model when, in CD, the cutoff probably extrapolates beyond the first order of neighbors.

Note that the cutoff value at 7.0 Å will have the meaning of optimum parameter candidate to better isolate the

first layer of neighbors if it had been used the coarse-grained residues representation of CA or GC. Other granularities may produce different reference cutoff values and even different neighbor profiles, but this does not invalidate the technique presented here. We are proposing a method to estimate the better cutoff that can be applied at any granularity, specifically to enclose as safe as possible the first shell of neighbors.

Other unexpected conclusion concerned the applicability of DT to prospect contacts in proteins. We viewed that DT had a bothersome characteristic that made it to ignore some sites close to the degenerate condition. Although our estimated DT error was low (in the order of 5% in BETA CA set), we have shown that it may be systematic and not random. One alternate solution used to solve this missing edges problem is the AD methodology. However, we have empirically demonstrated that AD tended to be a complement of DT, and therefore, their sum converged to CD as the perturbation parameter grows. If with CD, which is a much simpler technique than DT or AD, within the range up to 7.0 Å we have the same guarantee of full true-positive contacts, why use DT or DT+AD? It is worth emphasizing that we are not condemning DT and correlating techniques, as useless in protein contact analysis. Angelov *et al.*[78] have found intriguing properties in protein contacts, exploring topological parameters of the Voronoi cells, like the average number of edges by face, that may be seen as a weight of the contact related to the symmetry and type of interactions among neighbors. Our work can only state, in the strict range up to 7.0 Å in graphs of contacts with Euclidean distance weight, that DT or DT+AD seem to be not necessary since CD will yield, in a simpler way, complete and reliable results.

Finally, our case study comparing CA and GC residue representations shown that the use of cutoff below 6.8 Å may conduce to contradictory results regarding the question if either alpha or beta has a larger average number of neighbors. We saw that at a lower value than this cutoff, CA representation may induce a bias in favor of alpha and GC representation in favor of beta. Further investigations will be needed to verify if these biases are relevant or not, mainly for applications where the precision of contacts may be crucial, like empirical potentials. On the other hand, this final conclusion enforce the reliability of the 7.0 Å value as an ideal lower bond cutoff to be used in prospecting contacts in proteins, that is independent of the coarse-grained CA or GC residue representations.

ACKNOWLEDGMENTS

The authors thank Michael Waisberg, not only by the fecund discussions on subject, but also by his incredible ability of finding rare and old publications. We also thank Leonardo Vieira, Ricardo Leão, and Laila Barros for helpful reviews of this paper. We are also indebted to Bráulio Couto and Deive Oliveira for invaluable aid and advice on the statistical analysis.

REFERENCES

1. Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 1974;82:1–14.
2. Fleming PJ, Richards FM. Protein packing: dependence on protein size, secondary structure and amino acid composition. J Mol Biol 2000;299:487–498.
3. Tsai J, Taylor R, Chothia C, Gerstein M. The packing density in proteins: standard radii and volumes. J Mol Biol 1999;290:253–266.
4. Liang J, Dill KA. Are proteins well-packed? Biophys J 2001;81:751–766.
5. Lisewski AM, Lichtarge O. Rapid detection of similarity in protein structure and function through contact metric distances. Nucleic Acids Res 2006;34:1–10.
6. Lesk AM, Chothia C. How different amino acids sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 1980;136:225–270.
7. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
8. de Melo RC, Lopes CE, Fernandes FA Jr, da Silveira CH, Santoro MM, Carceroni RL, Meira W, Jr, Araújo Ade A. A Contact map matching approach to protein structure similarity analysis. Genet Mol Res 2006;5:284–308.
9. Holm L, Sander C. Protein structure comparison by alignment of distance matrix. J Mol Biol 1993;233:123–138.
10. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 1999;285:1711–1733.
11. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
12. Ponder JW, Richards FM. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1987;193:775–791.
13. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a know three-dimensional structure. Science 1991;253:164–170.
14. Veloso CJ, Silveira CH, Melo RC, Ribeiro C, Lopes JC, Santoro MM, Meira W, Jr. On the characterization of energy networks of proteins. Genet Mol Res 2007;6:799–820.
15. Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. Biophys J 2004;86:85–91.
16. Kannan N, Vishveshwara S. Identification of side-chain cluster in protein structures by a graph spectral method. J Mol Biol 1999;292:441–464.
17. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
18. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 1997;266:195–214.
19. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
20. Richards FM. Protein stability: still an unsolved problem. Cell Mol Life Sci 1997;53:790–802.
21. Ptitsyn OB, Ting KH. Non-functional conserved residues in globins and their possible role as a folding nucleus. J Mol Biol 1999;291:671–682.

22. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
23. Melo RC, Ribeiro C, Murray CS, Veloso CJ, da Silveira CH, Neshich G, Meira W, Jr, Carceroni RL, Santoro MM. Finding protein–protein interactions patterns by contact-map matching. Genet Mol Res 2007;6:946–963.
24. Mancini AL, Higa RH, Oliveira A, Dominiquini F, Kuser PR, Yamagishi ME, Togawa RC, Neshich G. STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. Bioinformatics 2004;20:2145–2147.
25. Heringa J, Argos P. Side-chain clusters in protein structures and their role in protein folding. J Mol Biol 1991;220:151–171.
26. Godzik A, Kolinski A, Skolnick J. Topology fingerprint approach to the inverse protein folding problem. J Mol Biol 1992;227:227–238.
27. Gregoret LM, Cohen FE. Protein folding: effect of packing density on chain conformation. J Mol Biol 1991;219:109–122.
28. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
29. Manavalan P, Ponnuswamy PK. A study of the preferred environment of amino acid residues in globular proteins. Arch Biochem Biophys 1977;184:476–487.
30. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707–726.
31. Furuichi E, Koehl P. Influence of protein structure databases on the predictive power of statistical pair potentials. Proteins 1998;31:139–149.
32. Kamagata K, Kuwajima K. Surprisingly high correlation between early and late stages in non-two-stage protein folding. J Mol Biol 2006;357:1647–1654.
33. Tanaka S, Scheraga H. A Model of protein folding: inclusion of short-, medium-, and long-range interactions. Proc Nat Acad Sci USA 1975;72:3802–3806.
34. Rodionov MA, Johnson MS. Residue-residue contact substitution probabilities derived from aligned three-dimensional structures and the identification of common folds. Protein Sci 1994;3:2366–2377.
35. Hernández G, LeMaster DM. Hybrid native partitioning of interactions among nonconserved residues in chimeric proteins. Proteins 2005;60:723–731.
36. Blades MJ, Ison JC, Ranasinghe R, Findlay JB. Automatic generation and evaluation of sparse protein signatures for families of protein structural domains. Protein Sci 2004;14:13–23.
37. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms: lyzozyme and insulim. J Mol Biol 1973;79:351–371.
38. Kirkwood JG. Molecular distribution in liquids. J Chem Phys 1939;7:919–925.
39. Godzik A, Skolnick J. Sequence structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc Natl Acad Sci USA 1992;89:12098–12102.
40. Tropsha A, Carter CW, Jr, Cammer S, Vaisman II. Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins. Methods Enzymol 2003;374:509–544.
41. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy minimization and dynamics calculations. J Comput Chem 1983;4:187–217.
42. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
43. Voronoi GM. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième Mémoire: Recherches sur les parallélloèdres primitifs. J Reine Angew Math 1908;134:198–287.
44. Delaunay B. Sur la sphère vide. A la memoire de Georges Voronoi. Izv Akad Nauk SSSR 1934;7:793–800.
45. Finney J. Volume occupation, environment and accessibility in proteins. The problem of the protein surface. J Mol Biol 1975;96:721–732.
46. Poupon A. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. Curr Opin Struct Biol 2004;14:233–241.
47. Loncharich RJ, Brooks BR. The effects of truncating long-range forces on protein dynamics. Proteins 1989;6:32–45.
48. Darden T, York D, Pedersen L. Particle mesh Ewald: a Nlog(N) method for Ewald sums in large systems. J Chem Phys 1993;98:10089–10092.
49. Monge A, Lathrop EJ, Gunn JR, Shenkin PS, Friesner RA. Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. J Mol Biol 1995;247:995–1012.
50. Bahar I, Jernigan RL. Coordination geometry of nonbonded residues in globular proteins. Fold Des 1996;1:357–370.
51. Yuan X, Bystroff C. Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. Bioinformatics 2005;21:1010–1019.
52. Soyer A, Chomilier J, Mornon JP, Jullien R, Sadoc JF. Voronoi tessellation reveals the condensed matter character of folded proteins. Phys Rev Let 2000;85:3532–3535.
53. Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J, Orengo CA. The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Res 2003;31:452–455.
54. 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.
55. Neshich G, Mazoni I, Oliveira SR, Yamagishi ME, Kuser-Falcao PR, Borro LC, Morita DU, Souza KR, Almeida GV, Rodrigues DN, Jardine JG, Togawa RC, Mancini AL, Higa RH, Cruz SA, Vieira FD, Santos EH, Melo RC, Santoro MM. The Star STING server: A multiplatform environment for protein structure analysis. Genet Mol Res 2006;05:717–722.
56. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
57. Pires D, Silveira C, Santoro M, Meira W, Jr. PDBEST—PDB enhanced structures toolkit. In Proceedings of the 3rd International Conference of Brazil Association for Bioinformatics 2007. São Paulo: AB3C Publishing; p 39.
58. Chothia C, Janin J. The nature of the accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–14.
59. Gerstein M, Tsai J, Levitt M. The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. J Mol Biol 1995;249:955–966.
60. Harpaz Y, Gerstein M, Chothia C. Volume changes on protein folding. Structure 1994;2:641–649.
61. Townsend M. Discrete mathematics: applied combinatorics and graph theory. Menlo Park: The Benjamin/Cummings Publishing Company; 1987.
62. Dirichlet GL. Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. J Reine Angew Math 1850;40:209–227.
63. Gauss CF. Recursion der 'Untersuchungen iiber die Eigenschaften der positiven ternaren quadratischen Formen von Ludwig August Seeber. J Reine Angew Math 1840;20:312–320.
64. Decartes R. Principia philosophiae. Amsterdan: Ludovicus Elzevirius;1644.
65. Thiessen H. Precipitation averages for large areas. Monthly Weather Rev 1911;39:1082–1084.
66. Wigner E, Seits F. On the constitution of metallic sodium. Phys Rev 1933;43:804–810.
67. Bandyopadhyay D, Snoeyink J. Almost-Delaunay simplices: nearest neighbor relations for imprecise points. In Proceedings of the 15th

Annual ACM-SIAM Symposium on Discrete Algorithms, 2004. Philadelphia: SIAM Publishing. Session 5A; pp 410–419.
68. Aurenhammer F, Klein R. Voronoi diagrams. In: Sack J, Urrutia G, editors. Handbook of computational geometry. Amsterdam: Elsevier Science; 2000. pp 201–290.
69. Bandyopadhyay D, Snoeyink J. Almost-Delaunay simplices: Robust neighbor relations for imprecise 3D points using CGAL. Comp Geom 2007;38:4–15.
70. Dwyer RA. Higher-dimensional Voronoi diagrams in linear expected time. Disc Comp Geom 1991;6:343–367.
71. Shimizu T, Nakatsu T, Miyairi K, Okuno T, Kato H. Active-site architecture of endopolygalacturonase I from Stereum purpureum revealed by crystal structures in native and ligand-bound forms at atomic resolution. Biochemistry 2002;41:6651–6659.
72. Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6:461–464.
73. Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 2004;33:261–304.
74. Welch BL. The generalization of "student's" problem when several different population variances are involved. Biometrika 1947;34:28–35.
75. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947;18:50–60.
76. 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.
77. Dupuis F, Sadoc JF, Julien R, Angelov B, Mornon JP. Voro3D: 3D Voronoi tessellations applied to proteins structures. Bioinformatics 2005;21:1715–1716.
78. Angelov B, Sadoc JF, Julien 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.
79. Kuntz LD, Crippen GM. Protein densities. Int J Pept Protein Res 1978;13:223–228.
80. Fligner MA, Killeen TJ. Distribution-free two-sample tests for scale. J Am Stat Assoc 1976;71:210–213.
81. Massey FJ, Jr. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc 1951;46:68–78.
82. Icking CR, Klein P, Köllner L. Java applets for the dynamic visualization of Voronoi diagrams. In: Computer science in perspective. New York: Springer-Verlag; 2003. pp 191–205.
83. Edelsbrunner H, Mücke EP. Simulation of simplicity: a technique to cope with degenerate cases in geometric algorithms. ACM Trans Graph 1990;9:66–104.
84. Barber CB, Dobkin DP, Huhdanpaa H. The quick hull algorithm for convex hulls. ACM Trans Math Soft 1996;22:469–483.
85. Duncan CA, Goodrich MT, Ramos EA. Efficient approximation and optimization algorithms for computational metrology. In Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms 1997. Philadelphia: SIAM Publishing; pp 121–130.
86. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583–621.
87. Behe MJ, Lattman EE, Rose GD. The protein-folding problem: the native fold determines packing, but does packing determine the native fold. Proc Natl Acad Sci USA. 1991;88:4195–4199.
88. Richards FM, Lim WA. An analysis of packing in the protein folding problem. Q Rev Biophys 1993;26:423–498.
89. Edelsbrunner H. The union of balls and its dual shape. Disc Comp Geom 1995;13:415–440.