Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

The interpretation of protein structures based on graph theory and contact map

This article was published in the following Dove Press journal:
Open Access Bioinformatics
18 September 2010
Number of times this article has been viewed

Mahnaz Habibi[1]
Changiz Eslahchi[1]
Mehdi Sadeghi[2]
Hamid Pezashk[3,4,5]

[1]Faculty of Mathematics, Shahid-Beheshti University, GC, Tehran, Iran; [2]National Institute of Genetic Engineering and Biotechnology, Tehran, Iran; [3]School of Mathematics, Statistics and Computer Sciences, College of Science, University of Tehran, Tehran, Iran; [4]Center of Excellence in Biomathematics, College of Science, University of Tehran, Tehran, Iran; [5]Bioinformatics Group, School of Computer Science, IPM, Tehran, Iran

Purpose: The analysis of a protein's structure allowing detailed exploration of the protein's biological function is one of the most challenging problems in bioinformatics. There are efficient algorithms to calculate main properties of a protein structure, such as packing density, buried or surface residues, and accessible surface area. But these algorithms need the three-dimensional (3D) coordinates of the proteins.

Methods: We used the contact map of a protein to construct a graph. By considering several features of the corresponding graph, we proposed some algorithms to discuss the above-mentioned properties of a protein. We also introduced a new measure for the hydrophobicity of an amino acid by defining an average degree for the amino acid as a vertex on the graph.

Results: We compared our results with those obtained by some other existing algorithms. We found strong correlations between the popular methods, which use 3D coordinates, and our methods, which only use a predicted contact map.

Conclusion: Many features of a protein can be predicted without having 3D coordinates, based on the contact map of the protein. The programs are freely available from http://www.bioinf.cs.ipm.ir/softwares/asa/asa.rar.

Keywords: accessible surface area, buried residue, surface residue, packing density, hydrophobic

Introduction

Proteins are linear chains that fold into characteristic shapes and features. It is widely believed that the three dimensional (3D) structure of a protein is the key to how this molecule carries out its biological function. To determine a protein's function, we need to analyze its structure.

Two main parameters used to determine a protein's structure are the measurements of occupied volume and volume of voids inside the protein molecule that provide information on the protein's stability.[1–3] There are different geometrical methods used to calculate the protein occupied volume. Richards used the Voronoi diagram to compute the occupied volume of a protein.[4] A Voronoi diagram is a geometric construction that is used to measure the packing density of a protein.[4–6]

The packing density of a protein is measured by the ratio of the protein's van der Waals volume to the occupied volume.[4,7] Packing density is a main feature of a protein's structure. Several studies have been conducted to analyze this feature.[8–13] In addition, functional roles of a protein molecule are determined by the interactions of the protein with other molecules. Physical interactions occur on the surface of the protein. Therefore, the identification of the surface of the protein is fundamental when studying the protein

Correspondence: Changiz Eslahchi
Faculty of Mathematics,
Shahid-Beheshti University,
GC, Tehran, Iran
Tel +98 21 29903015
Fax +98 21 22431652
Email ch-eslahchi@sbu.ac.ir

submit your manuscript | www.dovepress.com
Dovepress
DOI: 10.2147/OAB.S13790

structure. Lee and Richards recognized the importance of the protein surface.[14] They developed the widely used solvent-accessible surface model. The solvent-accessible surface is obtained by rolling a ball of radius *r* around the van der Walls surface of each residue. The computation of the surface area of a protein is always a difficult and time-consuming task. Recently, several analytical and numerical algorithms have been developed.[15,16] Braun et al presented a complete review on several methods and algorithms.[17]

The elucidation of a residue location in a protein is one of the most central problems that used the accessible surface area. The buried residues within a protein are almost hydrophobic residues which assemble in the protein interior. The accessible surface area in these residues is lower than others.

One-dimensional (1D) sequences are easily understood, but recognizing a 1D string does not tell us enough about the overall features and functions of a protein. We need a way to extract the information of a protein, without having the 3D coordinates of the protein. Recently, different approaches have been introduced to predict the contact map of a protein.[18–21] In this paper, we introduce two algorithms by using the contact map predicted by the PoCM algorithm.[21] The first algorithm generates a polymer with the maximum packing density for a given protein. Using the packed polymer, we determine the packing density of a molecule. The second algorithm (HYDCORE: hydrophobic core algorithm) is based on graph theory. We determine the core of a protein as a subgraph which has a large average score. In addition, we obtain the accessible surface area of each atom, only by using the contact map of a protein and some statistical information in the dataset. Because of strong correlations between the popular methods of calculating the accessible surface by using the 3D coordinates of each atom, we compared our results with these obtained from GETAREA (http://www.scsb.utmb.edu/). In general, despite the high difference in rationale behind our method and GETAREA, our results are almost the same as the solvent accessible surface area reported in GETAREA.

Furthermore, by using the fact that the hydrophobic groups tend to be assembled in the center of a protein, we scale amino acids corresponding to their hydrophobic properties.

Material and methods

In this section, we introduce some algorithms to determine the main features of a protein from its contact map using properties of graph theory.

Preliminaries

Contact maps

A contact between two given atoms (or residues) exists when a certain distance is below a given threshold. The distance between two residues may be defined by the distance between two carbon alpha (C_α) atoms,[22] or between two carbon beta (C_β) atoms,[23] or it may be the minimum distance between any pair of atoms belonging to the side chain or to the backbone of two residues.[18,19]

Let P be a protein with n atoms (or residues), which are labeled $1, 2, ..., n$. The Euclidean distance between two atoms i *and* j is denoted by $d(i, j)$. We define the contact map of the protein as a matrix $T = (t_{ij})_{1\le i,j\le n}$, where $t_{ij} = 0$ and $i \neq j$.

$$t_{ij} = \begin{cases} 1 & \text{if } D_{Cutoff}, \\ 0 & \text{otherwise}. \end{cases}$$

Some aspects of graph theory

Let $G = <V, E>$ be a graph in which V denotes the set of vertices, E denotes the set of edges, and $|V(G)| = n$. Two vertices u and v of G are called adjacent, or neighbors, if uv is an edge of G. The degree $d(v)$ of a vertex v is defined as the number of neighbors of v. The average degree of G is defined by

$$d(G) = \frac{1}{n} \sum_{v \in V(G)} d(v).$$

Let $\eta: V(G) \times V(G) \to [0,1]$ be a score function. We define $score(v)$ and $score^1(v)$ of a vertex v in G by

$$Score_G(v) = \sum_{vu \in E(G)} \eta(v,u),$$

$$score^1_G(v) = \sum_{vu \in E(G)} (\eta(v,u) + \eta(u,v)).$$

The average score of G is defined by

$$score(G) = \frac{1}{n} \sum_{v \in V(G)} score_G(v).$$

Suppose $\varepsilon(G) = score(G)/2$. It is clear that if a graph has a large minimum score, then it also has a large average score. It should be noted that the average score may be large even when its minimum score is small. However, the vertices of large score may not be distributed uniformly among all vertices. It is a classic algorithm of graph theory (its proof is similar to the proof of theorem 1.2.2) that every graph $\overline{G}$ with at least one edge has a subgraph H with

$$min_{v\in V(H)} score^1_H(v) > \varepsilon(H) \geq \varepsilon(G) \qquad (1)^{24}$$

This theorem implies that every graph G could have a subgraph H with the minimum score greater than $\varepsilon(G)$.

A protein can be considered as a graph $G = \langle V, E\rangle$, for which each vertex $v_k \in V$ represents a residue of the protein and each $v_iv_j \in E$ represents a contact between two residues v_i and v_j. On the other hand, there is an edge $v_iv_j \in E$, *if* $t_{ij} = 1$. By degree of a residue (atom) in a protein, we mean the degree of this amino acid (atom) in its corresponding graph.

The measurement of packing density

It is well known that each amino acid is approximated by a ball of radius1.7*Å*. Since two consecutive residues could be overlapped, a residue is considered a ball of radius 1.5*Å*.[25]

Step 1: Generating an approximately maximum compact polymer

The maximum packing density polymer is defined as a protein with the approximate maximum number of contacts. We generate a polymer of length n, say C_P, inside the sphere with the minimum radius, for which each residue has the maximum number of neighbors.

Assume that O, the 3D location of the center of the first monomer, is the origin of the coordinates. We find the centers of all external tangent balls to the first sphere. Let $\vartheta = \arcsin(1.5/3)$, where 2ϑ is the minimum angle between two vectors, starting from O and ending up in the centers of neighboring balls in each longitude. Therefore, the maximum number of balls for each longitude is approximately $N = [180/2\vartheta]$. For each latitude with radius R^*, the minimum angle between two vectors starting from O and ending up in the centers of neighboring balls is 2β where $\beta = \arcsin(1.5/R^*)$. Then, the maximum number of balls on each latitude is $[360/2\beta]$.

Now, suppose k balls, $k < n$, of radii 1.5 *Å* are located within the sphere of radius r. Similar to the above case, it could be shown that for each longitude, the minimum angle between two line segments drawn from O to the centers of the two neighboring balls located on the surface of the sphere of radius r is 2ϑ, where $\vartheta = \arcsin(1.5/1.5 + r)$.

For each latitude with radius R^*, the minimum angle between two lines starting from O and ending up in the centers of neighboring balls is 2β where $\beta = arcsin(1.5/R^*)$. So, the maximum number of balls on each latitude is $360/2\beta$. The above process terminates if n balls are located in a sphere that has a high packing density.

Step 2: Calculation of packing density

Let P be a protein with n residues, and T be its contact map of P, and G be a graph corresponding to P. We generate an approximate maximum packing polymer C_p with n residues. Let G_p be the graph corresponding to C_p. We define the packing density of P, P_d, as the number of edges of the graph G divided by the number of the edges of the graph G_p. Formally, the packing density of the protein is denoted by

$$P_d = \frac{|E(G)|}{|E(G_p)|_{max}},$$

where $|E|$ is the size of the set of the edges of a graph G. Since the number of edges of graph G is at most $n(n-1)/2$, we can obtain the packing density of a protein by an algorithm of order $O(n^2)$.

Definition of surface and buried residues in a protein

In this section, the new algorithm, HYDCORE, is presented to predict surface and buried residues from an input contact map. HYDCORE contains two main steps. In the first step, we define a score function η, and in the second step we obtain the subgraph H that satisfies in (4) to predict buried and surface residues.

In the first step, for every two amino acids, $u, v \in \{1,2,\ldots,20\}$, we define a matrix $P = (p(u, v))_{1\leq u, v\leq 20}$ by

$$p(u,v) = \frac{1}{2}\left(1 - \frac{\sqrt{R_u^2 + 2R_uR_v}}{R_u + R_v}\right),$$

where R_u is the radius of the sphere that has its surface area equal to the surface area of amino acid u extracted from DSSP. The value $p(u, v)$ is the ratio of the surface area of an amino acid u which is covered by the amino acid v when the sphere of these amino acids are tangent to each other (see Figure 1).

Let G be the corresponding graph of a contact map. The score function η for G is defined by

$$\eta(i, j) = p(u, v),$$

where u and v are two amino acids corresponding to residues i and j of the sequence of the protein.

We seek to find a subgraph H of G satisfying the following process. Let $G_0 = G$ and v_1 be a vertex of G with minimum $score^1$. If $score^1_{G_0}(v_1) \leq \varepsilon(G_0)$, then consider $G_1 = G_0 - v_1$, otherwise $H = G_0$. Construct a sequence

$$G = G_0 \supseteq G_1 \supseteq \ldots. \supseteq G_i$$

of subgraphs of G.

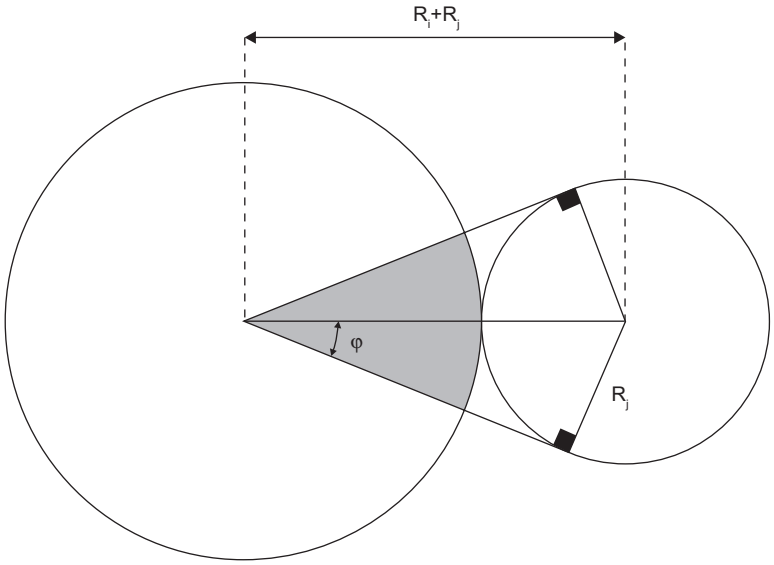


Figure 1 The colored region shows the part of amino acid *i* covered by amino acid *j*, and its surface area is defined by $2\pi R_i^2(1-cos(\varphi))$, where $0 \le \varphi \le \pi/2$ and $cos(\varphi) = \sqrt{R_i^2 + 2R_iR_j}/(R_i + R_j)$.

If G_i has a vertex v_i, such that

$$score^1_{G_i}(v_i) \le \varepsilon(G_{i-1}),$$

then $G_{i+1} = G_i - v_i$, otherwise consider $H = G_i$. This process stops when there are no vertices v_i in G_i such that

$$score^1_{G_i}(v_i) \le \varepsilon(G_i).$$

Now we use the subgraph H to identify the buried and surface residues. The vertices of H which have the score of at least $score(H)$ will be considered as the buried residues, and the vertices of $G-H$ are considered as surface residues.

Calculation of accessible surface area for each atom

In this section, we first define two radii for two contact atoms i and j in a way that these atoms with the defined radii are expected not to overlap.

Let $d_1, d_2, ..., d_m$ be the distances between two contact atoms, i and j, in all proteins in the dataset. We define

$$r_{ij} = \frac{1}{m}\sum_{k=1}^{m}\frac{d_k R_i}{R_i + R_j},$$

where R_i is the radius of the atom i.[26]

Let P be a protein with N atoms and G be its corresponding graph with respect to atom contact map of P. Let $S = \{i_1, i_2, \ldots, i_t\}$ be the set of the atoms in contact with i. Suppose $r_i = min\{r_{ii_1}, r_{ii_2}, ..., r_{ii_t}\}$. We define the occupied surface area of i as

$$O_i = \sum_{k=1}^{t} 2\pi r_i^2\left(1 - \frac{\sqrt{r_i^2 + 2r_i r_{i_k i}}}{r_i + r_{i_k i}}\right).$$

We define the accessible surface area of the atom i by

$$ASA(i) = \begin{cases} 0 & if\ \ 4\pi r_i^2 < O_i; \\ 4\pi R_i^2 - O_i(4\pi R_i^2/4\pi r_i^2) & \text{otherwise.} \end{cases}$$

Generating a random contact map of a protein

When we analyze the contact map of a protein, we can calculate the number of contacts between two amino acids u and v of a protein P by $f_{uv}(P)$. Therefore the probability of having a contact between u and v is defined by

$$Prob_{u,v} = \frac{\sum_{P \in Dataset} f_{uv}(P)}{\sum_{P \in Dataset} n_u(P) n_v(P)},$$

where $n_u(P)$ is the number of amino acid u in a protein P.

Let V be a fixed set of n residues of a protein, say $V = \{1, 2, ..., n\}$. We would like to generate a graph G, such that each vertex corresponds to a residue and the adjacency matrix corresponds to a contact map of the protein. Intuitively, we generate G randomly as follows:

First, we define the probability that the adjacency matrix of the graph is a contact map of a protein, by

$$Prob_{G_0} = \frac{good_{vertex}}{|V(G)|},$$

where $good_{vertex}$ is the number of vertices that have degrees between the maximum and minimum degree (see Table 1).

For each edge $e = ij \in E(G)$, we decide whether or not e could be an edge of G; if $x > Prob_{\{i=u,\, j=v\}}$ where x is the random number chosen from the interval [0, 1], we accept e as an edge of G.

Let ℓ be a fixed number between 0 and 1, and $G_0, G_1, ..., G_t$ be the sequence of random graphs, such that $Prob_{G_i} < \ell$ ($i = 1, ..., t$). Randomly choose vertex i of graph G_t and remove all adjacent edges to i. The new graph G^* is constructed by adding some new edges to i obtained by the random experiments. If $Prob_{G_t} < Prob_{G^*}$, then $G_{t+1} = G^*$, otherwise repeat the last step. The algorithm stops if $Prob_{G_t} \geq \ell$.

Results

Dataset

A representative set of X-ray protein structures with resolution $<1.7\text{Å}>$ is gathered from the Protein Data Bank (PDB) by using the advanced search in RCSB (http://www.rcsb.org). The structures with more than 40% similarity in sequences are excluded. Taking these criteria into account, 1988 proteins were selected.

Performance evaluation measures

There are many studies that identify buried or surface residues. These algorithms require a 3D structure to find the location of residues of a protein, while our results about the location of residues are based on the contact map of the protein. However, to evaluate the performance of our method, we compared the predicted surface and buried residues with surface and buried residues reported by DSSP. We use two known measures of

$$Sn = \frac{TP}{TP + FN} \qquad SP = \frac{TN}{TN + FP}.$$

Table 1 The summary of some statistics of each amino acid

Amino acids	A	R	N	D	C	Q	E	G	H	I
Mean	5.7	4.8	5	4.9	6.9	4.7	4.7	5.2	5	5.7
Minimum	3	4	2	3	4	4	2	2	3	3
Maximum	10	8	9	8	7	8	8	7	7	6
Amino acids	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**
Mean	5.6	4.5	6	6.2	4.4	5.1	5.2	5.8	5.6	5.8
Minimum	3	2	2	2	2	4	3	4	4	4
Maximum	8	7	8	8	6	8	7	8	7	9

These measures are based on the relation between the number of residues correctly assigned positive (*TP*), the number of residues correctly assigned negative (*TN*), the number of residues incorrectly assigned positive (*FP*), and the number of residues incorrectly assigned negative (*FN*).

The agreement, A, between two methods is another measure used in this paper. Agreement is defined as the number of residues for which both methods agree ($TP + TN$), divided by the total number of residues.

$$A = \frac{TP + TN}{TP + TN + FP + FN}.$$

Observations

D_{cutoff} threshold

The minimum Euclidean distance between two consecutive residues will be assumed to be 3Å.[25] Also, the maximum Euclidean distance between two contact residues is assumed to be 9Å. 9Å is chosen, because any amino acid could not be able to locate between two contact amino acids. We first change D_{cutoff} threshold between $3 - 9\text{Å}$ for two C_α atoms to obtain different contact maps. Figure 2 shows the standard deviations of packing densities of all proteins in the dataset for different D_{cutoff} thresholds. We find that $D_{cutoff} = 8\text{Å}$ yields the highest standard deviation of packing densities. Therefore $D_{cutoff} = 8\text{Å}$ could be used to distinguish proteins.

Statistical analysis

It is well known that the number of contacts relates linearly to the chain length.[23,27] The slope of the linear relation depends only on how a contact map is defined. Using C_α atoms and 8Å as cutoff distance, the number of contacts in a compact globular protein is approximately three times the length of the protein, with a relatively small standard deviation of ±0.4. In Table 1, some of the descriptive statistics of each amino acid are presented.

Packing density

Figure 3 shows the correlation of the packing density, P_d, with the number of residues in real proteins. We find that P_d decreases when the number of residues increases. The short-chain protein has a higher packing density, which is from 0.5 to 0.65, and the proteins with length of at least 175 residues have packing densities near to 0.53.

Now, for each protein with n residues, we generate an $n \times n$ random contact map. Figure 4 shows the correlation of P_d with the number of residues in both the actual and random proteins. In Table 2, some of the descriptive statistics

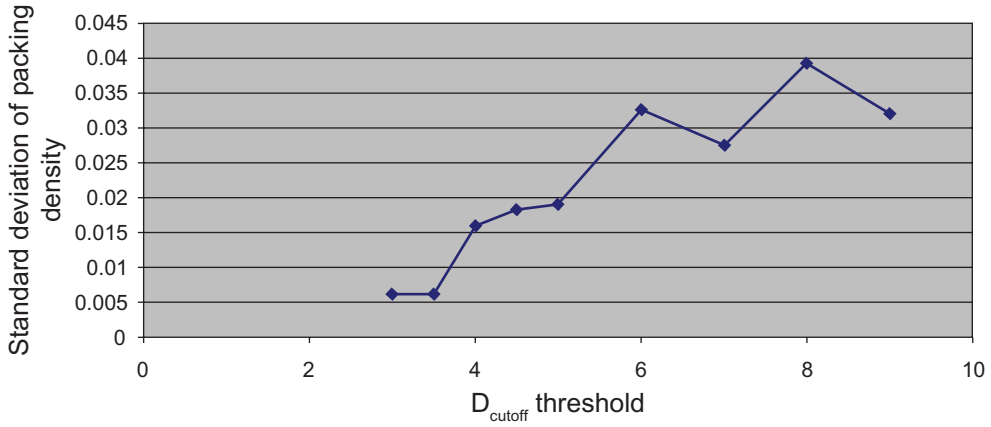


Figure 2 The D_{cutoff} threshold verses standard deviation of packing densities.

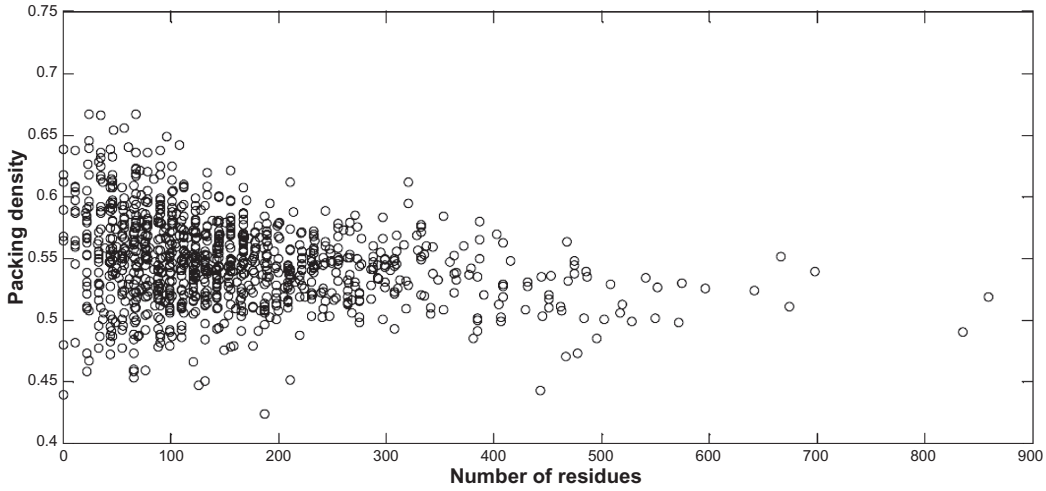


Figure 3 The relation between packing density and the number of residues of proteins.

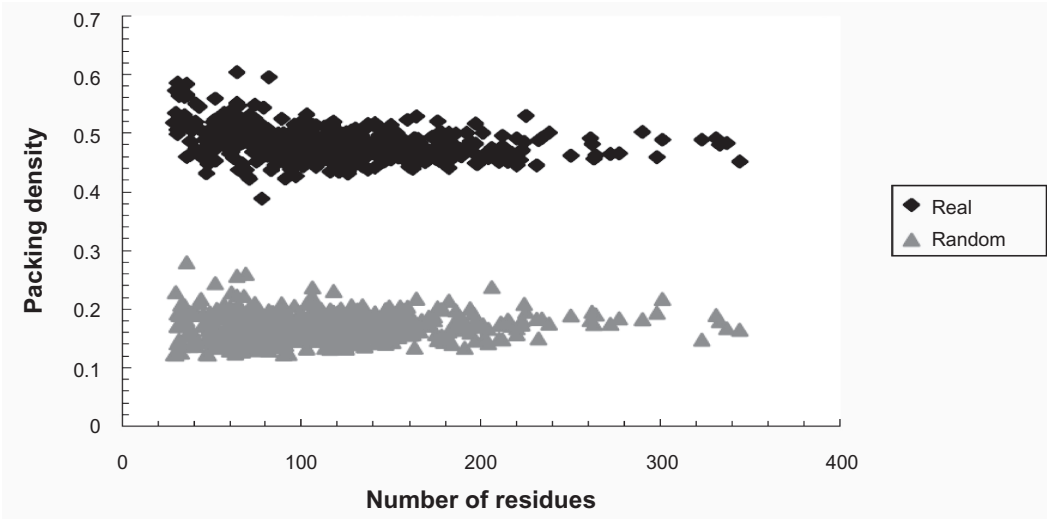


Figure 4 Packing density versus the number of residues of actual (real) and randomly constructed proteins.

Table 2 The summary of descriptive statistics

	Mean	SD	Minimum	Quartile1	Median	Quartile3	Maximum
Actual	0.4851	0.0008	0.3883	0.4663	0.4831	0.501	0.6037
Random	0.1704	0.0006	0.1221	0.1531	0.1696	0.1852	0.2795

Abbreviation: SD, standard deviation.

are presented. It is concluded that the packing density of a protein is an important parameter for differentiating actual contact maps from randomly constructed contact maps. Figure 5 shows the packing density plotted against the total surface area of each protein in the dataset. This figure supports the reverse relationship between total accessible surface area and packing density.

Calculation of hydrophobicity in protein

Calculation of hydrophobicity of a protein is considered as one of the most important properties in determining biological function. There are several legitimate ways to determine the hydrophobicity of a protein. In this section, a new scale of hydrophobicity is introduced, which we call SBG (scale based on graph theory).

It is known that the residues in protein cores have higher degrees than surface amino acids, and also that the hydrophobic residues tend to sit in the core of a protein to avoid water.[28] Therefore, we expect that the hydrophobic residues have higher degrees than hydrophilic ones. We are interested in scaling each amino acid according to its average degree. We defined the total degree of each amino acid i as the sum of the degree of amino acid i in all proteins in the dataset. The total average degree of i is defined by the ratio of the total degree i to its frequency in the dataset. Figure 6 shows the total average degrees for each amino acid in dataset.

We consider the total average degree of each amino acid as the scale of hydrophobicity properties. The higher average degree of an amino acid implies the higher tendency for being in the core of a protein. We sort the amino acids from the highest total average degree to the lowest total average degree.[29–35] In Table 3, we present our scales and other scales obtained by different methods. In general, despite the different rationales behind our method and those popular methods, there is a strong correlation between them. In fact, using the Spearman correlation coefficient, we conclude our method generally has more correlation with other existing methods (with minimum correlation of 0.782).

Testing for accuracy

In this section, we only select the structures with less than 25% similarity of sequences. In this work, we introduce a

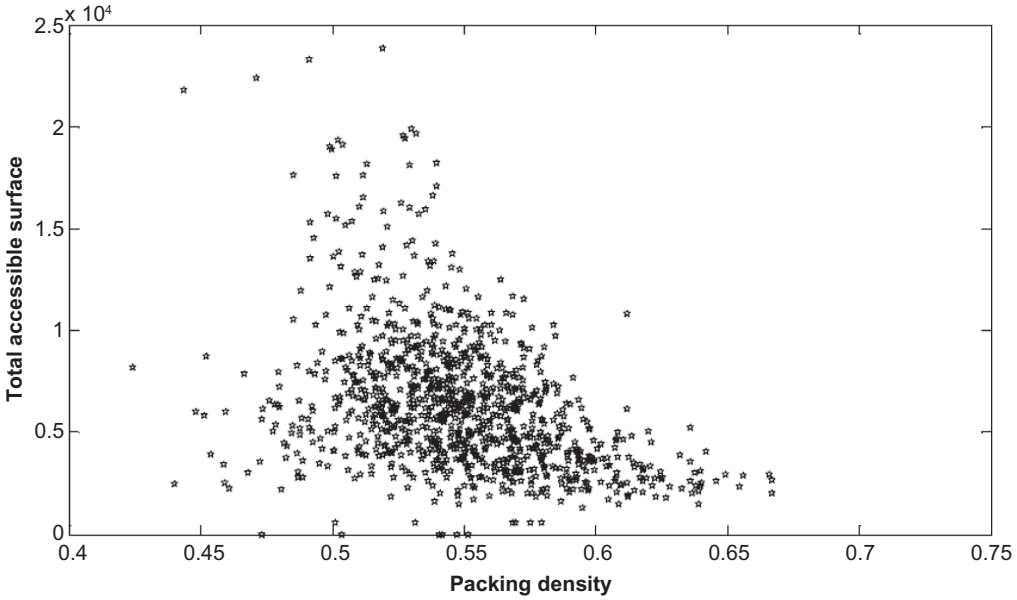


Figure 5 Packing density versus the total accessible surface area.

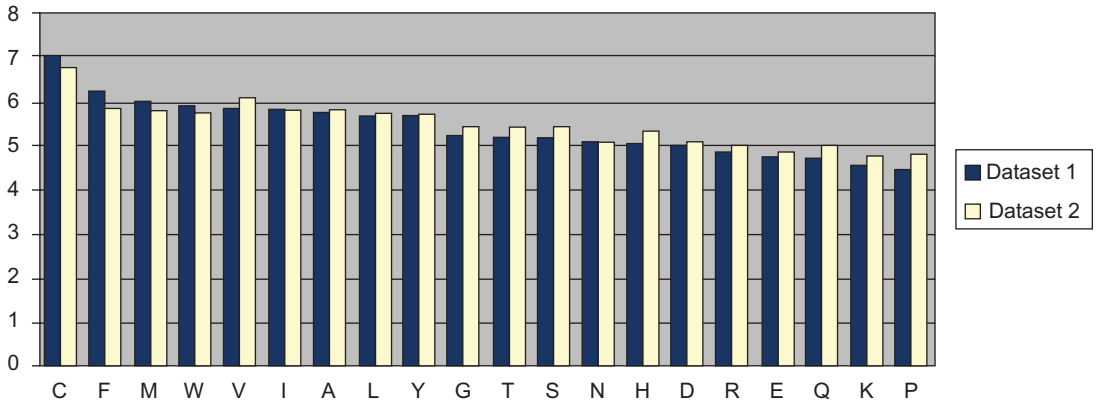


Figure 6 The total average degrees of all amino acids for two datasets.

new algorithm, HYDCORE, to predict the location – surface or core – of each residue of a protein. The PoCM algorithm is first used to predict the contact map of the protein before HYDCORE is run. To check the validity of our method, we compared our results predicted by HYDCORE with the results reported by DSSP. Tables 4 and 5 show the results of these comparisons. According to DSSP, the residues with the percentages of accessible surface area lower than *k*% are identified as the buried residues, and the residues with the percentages of accessible surface area greater than *k*% are identified as residues located on the surface, named surface amino acids. We find that there is a strong correlation between HYDCORE and DSSP for buried and surface residues in $k = 15\%$ and $k = 80\%$, respectively. We also use 3D coordinates to calculate the contact map of the protein before our method is run. Tables 4 and 5 reveal that the results of our method has considerable agreement to the results using the predicted contact map. So, the more accurate the contact map prediction, the more accurate the HYDCORE we obtain.

Table 3 The scale of hydrophobicity properties. Comparison of the SBG scale of the dataset with other scales

Amino acid	SBG scale	Kyte, Doolittle[34]	Engelman et al[31]	Eisenberge et al[30]	Hopp, Woods[32]	Cornette et al[29]	Rose et al[35]	Janin[33]
C	6.74	2.5	−1	4.1	0.29	0.91	0.9	2
F	5.83	2.8	−2.5	4.4	1.19	0.88	0.5	3.7
M	5.8	1.9	−1.3	4.2	0.64	0.85	0.4	3.4
W	5.76	-0.9	−3.4	1	0.81	0.85	0.3	1.9
V	6.08	4.2	−1.5	4.7	1.08	0.86	0.6	2.6
I	5.81	4.5	−1.8	4.8	1.38	0.88	0.7	3.1
A	5.81	1.8	−0.5	0.2	0.62	0.74	0.3	1.6
L	5.76	3.8	−1.8	5.7	1.06	0.85	0.5	2.8
Y	5.73	−1.3	−2.3	3.2	0.26	0.76	−0.4	−0.7
G	5.46	−0.4	0	0	0.48	0.72	3	1
T	5.45	−0.7	−0.4	−1.9	−0.05	0.7	−0.2	1.2
S	5.45	−0.8	0.3	−0.5	−0.18	0.66	−0.1	0.6
N	5.08	−3.5	0.2	−0.5	−0.78	0.63	−0.5	−4.8
H	5.34	−3.2	−0.5	0.5	−0.4	0.78	−0.1	−3
D	5.07	−3.5	3	−3.1	−0.9	0.62	−0.6	−9.2
R	5.02	−4.5	3	1.4	−2.53	0.64	−1.4	−12.3
E	4.84	−3.5	3	−1.8	−0.74	0.62	−0.7	−8.2
Q	5.01	−3.5	0.2	−2.8	−0.85	0.62	−0.7	−4.1
K	4.77	−3.9	3	−3.1	−1.5	0.52	−1.8	−8.8
P	4.8	−1.6	0	−2.2	0.12	0.64	−0.3	−0.2

Abbreviation: SBG, scale based on graph theory.

Table 4 Comparison of buried residues obtained by HYDCORE and DSSP

	Setting[a]	TP	TN	Fp	FN	Sn	Sp	A
10%	1	12883	25664	10699	3070	0.807	0.705	0.736
	2	14263	26993	9370	1690	0.894	0.742	0.788
15%	1	15132	24825	8450	3909	0.794	0.746	0.863
	2	15955	25597	7678	3086	0.83	0.769	0.794
20%	1	16227	22833	7355	5901	0.733	0.756	0.746
	2	18222	24777	5411	3906	0.823	0.82	0.821
25%	1	18336	22369	5246	6365	0.742	0.81	0.778
	2	19971	23953	3662	4730	0.808	0.867	0.839

Note: [a]1 refers to HYDCORE's results obtained by using the PoCM algorithm to predict the contact map, and 2 refers HYDCORE's results obtained by using 3D coordinates to predict the contact map.
Abbreviations: HYDCORE, hydrophobic core algorithm; DSSP, definition of secondary structure of proteins.

Table 5 Comparison of surface residues obtained by HYDCORE and DSSP

	Setting[a]	TP	TN	Fp	FN	Sn	Sp	A
95%	1	1213	46873	3508	722	0.626	0.93	0.919
	2	1471	46571	3810	464	0.76	0.924	0.918
90%	1	1556	46845	3065	750	0.688	0.938	0.927
	2	1925	46554	3356	481	0.8	0.932	0.926
85%	1	2581	46567	2140	1028	0.715	0.956	0.939
	2	2766	46192	2515	843	0.766	0.948	0.935
80%	1	3472	45888	1249	1707	0.67	0.973	0.943
	2	4132	45988	1149	1047	0.797	0.975	0.958

Note: [a]1 refers to HYDCORE's results obtained by using the PoCM algorithm to predict the contact map, and 2 refers HYDCORE's results obtained by using 3D coordinates to predict the contact map.
Abbreviations: HYDCORE, hydrophobic core algorithm; DSSP, definition of secondary structure of proteins.

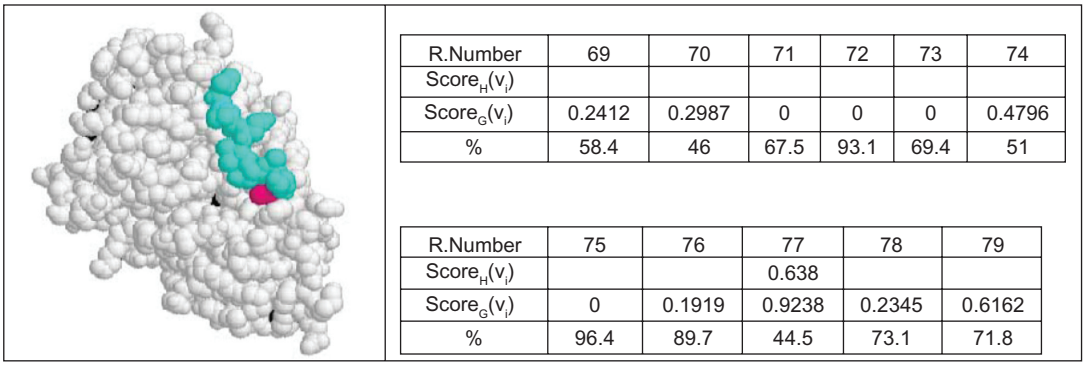
To calculate accessible surface area of each atom of a protein, we present a method based on the atom contact map of a protein. We compared our results with the results of GETAREA. Let μ_0 and μ_1 be the means of our results and GATAREA results, and h_0 be the assumption of $\mu_0 = \mu_1$ respectively, and h_1 be the null hypothesis that $\mu_0 \neq \mu_1$. We found that the prob($h_1|h_0$) is about 0.039. With a type 1 error of 5%, we reject the null hypothesis. It is shown that the results of our method and the GATAREA method are usually similar; however, we found these results by using the contact map and GATAREA using 3D coordinates. This test is done for the several proteins based on the contact map computed by the 3D structure of the protein.

The other test for the performance of this method is done by calculating the time of our algorithm. It is clear that our algorithm is a fast algorithm. This algorithm is of the order $O(n^2)$, where n is the number of vertices of the graph corresponding to a contact map.

Conclusion

In the first part of this work, we introduce a new measure to compute the packing density of a protein using the corresponding contact map. The correlation of the packing density with the number of residues is shown in Figure 3. We also plot packing density against total surface area in Figure 5. It is revealed that there exists a reversed relationship between them.

In the second part of this work, we proposed HYDCORE to assign surface or buried residues, using the predicted contact map of a protein. With HYDCORE, the score of each residue is calculated from the adjacent vertices in the corresponding graph. The score of a residue indicates the surface area of that amino acid occupied by adjacent residues. Therefore, the score value of a residue can be used as a parameter to describe the occupied surface of a residue. So, the lower score of an amino acid implies the lower probability of being in the core of a protein. We expected



R.Number	69	70	71	72	73	74
$Score_H(v_i)$						
$Score_G(v_i)$	0.2412	0.2987	0	0	0	0.4796
%	58.4	46	67.5	93.1	69.4	51

R.Number	75	76	77	78	79
$Score_H(v_i)$			0.638		
$Score_G(v_i)$	0	0.1919	0.9238	0.2345	0.6162
%	96.4	89.7	44.5	73.1	71.8

Figure 7 A part of the protein of lambda exonuclease (PDB code1AVQC, residues 69–79) and scores value of its residues in graph *G* calculated by HYDCORE. Residue 77 has a high score value and locates in the subgraph *H*, but its score value is not higher than score(*H*) = 0.7588.
Abbreviations: PDB, Protein Data Bank; HYDCORE, hydrophobic core algorithm.

Table 6 Two score values of a part of the protein of *Trehalose repressor from Escherichia coli* (PDB code 1BYK) in subgraph *H* calculated by HYDCORE

PDB code	Amino acid	Residue number	Accessible surface	%	ScoreH(v_i)	ScoreG(v_i)
1BYK	THR	279	4.25	4	1.3032	1.3032
1BYK	VAL	280	0	0	1.4541	1.4541
1BYK	ASP	281	39.86	35.3	0.5198	0.5198
1BYK	PRO	282	1.15	1.1	1.259	1.259
1BYK	GLY	283	0	13.6	1.1144	1.1144

Note: Residue 281 has score value lower than, and the other greater than, *score*(*H*) = 0.8363. Therefore, we expect all residues except residue 281 to be inside the hydrophobic core.
Abbreviations: PDB, Protein Data Bank; HYDCORE, hydrophobic core algorithm.

not to see any residue with a high score value in the surface amino acids.

HYDCORE finds two subgraphs *H* and *G-H* of *G*. The subgraph *G-H* includes the vertices with low score values; therefore, the residues in subgraph *G-H* reside on the surface of a protein. Figure 7 shows a part of the protein surface of lambda exonuclease (PDB code 1AVQ, residues 69–79). The score values of all residues except for residue 77 are lower than the average score value, thus we expect that these residues are located on the protein surface.

According to our algorithm, all residues in *H* with score value greater than *score*(*H*) are identified as a hydrophobic core. For example, as shown in Table 6, all residues in part of Trehalose repressor from *Escherichia coli* (PDB code 1BYK, residues 279–283) except the residue 281 have the score values greater than *score*(*H*). Therefore, by this parameter we expect these residues to be in hydrophobic cores and the residue 281 with the score value 0.5198 is known as a nonhydrophobic residue.

We also present a new sale to sort the amino acids corresponding to hydrophobicity properties. We compare our scale and other scales obtained by different methods. We show that there are strong agreements between them.

Acknowledgment

Changiz Eslahchi and Mahnaz Habibi are supported by Shahid Beheshti University.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Eriksson A, Baase W, Zhang XJ, et al. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*. 1992;255:178–183.
2. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein. *Natl Acad Sci U S A*. 1997;94:10172–10177.
3. Privalov PL. Intermediate states in protein folding. *J Mol Biol*. 1996; 258:707–725.
4. Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. *J Mol Biol*. 1997;82:1–14.
5. Finney JL. Volume occupation, environment and accessibility in proteins. The problem of the protein surface. *J Mol Biol*. 1975;96: 721–732.
6. Gellatly BJ, Finney J. Calculation of protein volumes: an alternative to the Voronoi procedure. *J Mol Biol*. 1982;161:305–322.
7. Richards FM. Areas, volumes, packing, and protein structures. *Annu Rev Biophys Bioeng*. 1977;6:15–76.
8. Finney JL. Volume occupation, environment and accessibility in proteins. Environment and molecular area of RNase. *J Mol Biol*. 1978; 119:415–441.
9. Gerstein M, Chothia C. Packing at the protein-water interface. *Proc Natl Acad Sci U S A*. 1996;93:10167–10172.
10. Gerstein M, Richards FM. Protein geometry: volumes, area, and distances. In: Rossman MG, Arnold E, editors. *International Tables for Crystallography*. Norwell, MA: Springer; 2001;Volume F, Chapter 22.
11. Liang J, Dill KA. Are proteins well-packed? *Biophys J*. 2001;81: 751–766.
12. Richards FM. Packing defects, cavities, volume fluctuation, and access to the interior of protein. Including some general comments on surface area and protein structure. *Carlsberg Res Commum*. 1979;44:47–63.
13. Richards FM, Lim WA. An analysis of packing in the protein folding problem. *Q Rev Biophys*. 1994;26(2):423–498.
14. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*. 1971;55:379–400.
15. Hayryan S, Hu CK, Skrivánek J, Hayryane E, Pokorný I. A new analytical method for computing solvent-accessible surface area of macromolecules and its gradients. *J Comput Chem*. 2004:334–343.
16. Masuya M, Doi J. Detection and geometric modeling of molecular surface and cavities using digital mathematical morphological operations. *J Mol Graphics*. 1995;13:331.
17. Braun W. In van Gunsteren W, Weiner P, Wilkinson T, editors. *Computer Simulation of Biomolecular Systems*. Leiden, The Netherlands: ESCOM; 1996.
18. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of cantact maps with neural networks and correlated mutations. *Protein Eng*. 2001;14:835–843.
19. Fariselli P, Olmea O, Valencia A, Casadio R. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins Suppl*. 2001;5:157–162.
20. Pollastri G, Baldi P. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*. 2002;8 Suppl 1:S62–S70.
21. Hamilton N, Burrage K, Ragan MA, Huber T. Protein contact prediction using patterns of correlation. *Proteins*. 2004;56:679–684.
22. Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. *Fold Des*. 1997;3:329–336.
23. Thomas DJ, Casari G, Sander C. The prediction of protein contacts from multiple sequence alignments. *Protein Eng*. 1996;9:941–948.
24. Diestel R. *Graph Theory*. New York, NY: Springer-Verlage Heidelberg; 2005.

25. Zhang J, Chen R, Tang C, Liang J. Origin of scaling behavior of protein packing density: a sequential Monte Carlo study of compact long chain polymers. *J Chem Phys*. 2003;118:6102–6109.
26. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci*. 1992:1–227.
27. Fariselli P, Casadio R. A neural network based on predictor of residue contacts in proteins. *Protein Eng*. 1999;12:15–21.
28. Kauzmann W. Some factors in the interpretation of protein denaturation. *Adv Prot Chem*. 1959;14:1–63.
29. Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol*. 1987;195(3):659–685.
30. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol*. 1984;179(1):125–142.
31. Engelman DM, Steitz TA, Goldman A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*. 1986;15:321–353.
32. Hopp TP, Woods KR. A computer program for predicting protein antigenic determinants. *Mol Immunol*. 1983;20(4):483–489.
33. Janin J. Surface and inside volumes in globular proteins. *Nature*. 1979;277(5696):491–492.
34. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982;157(1):105–132.
35. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. *Science*. 1985;229(4716):834–838.