PROTEINS: Structure, Function, and Genetics 53:792–805 (2003)

Simplicial Edge Representation of Protein Structures and Alpha Contact Potential with Confidence Measure

Xiang Li, Changyu Hu, and Jie Liang*
Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois

ABSTRACT **Protein representation and potential function are two important ingredients for studying protein folding, equilibrium thermodynamics, and sequence design. We introduce a novel geometric representation of protein contact interactions using the edge simplices from the alpha shape of the protein structure. This representation can eliminate implausible neighbors that are not in physical contact, and can avoid spurious contact between two residues when a third residue is between them. We developed statistical alpha contact potential using an odds-ratio model. A studentized bootstrap method was then introduced to assess the 95% confidence intervals for each of the 210 propensity parameters. We found, with confidence, that there is significant long-range propensity (>30 residues apart) for hydrophobic interactions. We tested alpha contact potential for native structure discrimination using several sets of decoy structures, and found that it often performs comparably with atom-based potentials requiring many more parameters. We also show that accurate geometric representation is important, and that alpha contact potential has better performance than potential defined by cutoff distance between geometric centers of side chains. Hierarchical clustering of alpha contact potentials reveals natural grouping of residues. To explore the relationship between shape and physicochemical representations, we tested the minimum alphabet size necessary for native structure discrimination. We found that there is no significant difference in performance of discrimination when alphabet size varies from 7 to 20, if geometry is represented accurately by alpha simplicial edges. This result suggests that the geometry of packing plays an important role, but the specific residue types are often interchangeable. Proteins 2003;53:792–805.**

Key words: simplicial edge; alpha contact; alpha shape; protein potential function; bootstrap

INTRODUCTION

Potential function plays an important role in a variety of computational studies of proteins, including prediction of protein structures, characterization of ensemble thermodynamic properties of proteins, and design of novel proteins. For example, an essential requirement for the prediction of three-dimensional (3D) structure of protein from primary sequence is a potential function that can select the native conformation from an ensemble of alternative conformations. Potential function is also often used to guide the sampling of protein conformations.[1] A variety of potential functions have been developed for these important tasks, including physical model–based potentials,[2–4] empirical statistical potentials,[5–7] and empirical potentials obtained from optimization.[8–12]

The effectiveness of potential function depends on another critically important factor, the representation of protein structures. Within this framework, we explore a new type of pairwise contact potential using a novel contact definition. We introduce a contact definition that reflects protein geometry more accurately. These contacts are based on the computation of the alpha shape, or the dual simplicial complex description of protein structures.[13,14] Here, contact occurs if atoms from nonbonded residues share a Voronoi edge, and this edge is at least partially contained in the body of the molecule; that is, atoms have to be in physical contact with their nearest neighbor. In this study, we only examine the 1-simplices, or edges in the dual simplicial complex that represent the pairwise contacts. This description is related to the work by Wernisch et al.,[15] in which the contact area between atoms is calculated as the area of intersection of the accessible atom ball around each atom and the faces of its weighted Voronoi cell.

We developed statistical contact propensities based on alpha-edge simplices and a combinatorial null model. To account for the uncertainty of estimated propensity parameter, we developed a studentized bootstrap method for estimating 95% confidence intervals (CIs) and also examined how geometric representation of the dual simplicial complex influences the effectiveness of pairwise contact potential functions. An additional goal is to understand how alphabet size of amino acid residues affects empirical pair potentials' effectiveness. This is important for protein

Grant sponsor: National Science Foundation; Grant numbers: CAREER DBI0133856, DBI0078270, and MCB998008. Grant sponsor: American Chemical Society/Petroleum Research Fund Grant number: 35616-G7.

*Correspondence to: Jie Liang, Department of Bioengineering, SEO, MC-063, University of Illinois at Chicago, 851 S Morgan St., Room 218, Chicago, IL 60607-7052. E-mail: jliang@uic.edu

Received 3 September 2002; Revised 10 January 2003; Accepted 17 February 2003

design, in which any reduction of the alphabet size of residues will result in exponentially more efficient sampling in the sequence space, therefore leading to more successful design strategies.[16,17]

This work is also motivated by the need to develop potential functions that take advantage of recent development in computational geometry and topology. The alpha-shape representation of proteins allows rapid, precise calculations of metric, topologic, and combinatorial structures of proteins.[14,18,19] These advantages can lead to improvements in void and binding-site detection,[18–20] in hierarchical representation of protein dynamic shapes at different resolution, and in conformation sampling. Recent theoretical developments suggest many intriguing applications in protein studies.[21–23] These important advances are largely unexploited and we hope this work provides a useful link by developing empirical pairwise contact potentials based on dual simplicial complex representations of proteins.

Our approach also solves a problem that cannot be satisfactorily addressed with current contact pair potentials. In these approaches, pairwise contact interactions are declared if two residues are within a specific cutoff distance. Potentials based on this contact definition have achieved considerable success. Nevertheless, contacts by distance cutoff can include many implausible, noncontacting neighbors that have no significant physical interaction.[24] Whether or not a pair of residues can make physical contact depends not only on the distance between their center positions (such as C_α or C_β, or geometric centers of sidechains) but also the size and orientations of sidechains.[24] Furthermore, two atoms close to each other may in fact be shielded from contact by other atoms. As emphasized by Taylor,[25] these contact pairs should not contribute to the assessment of pairwise contacts. By occupying the intervening space, other residues can block the interaction of a pair of residues. Inclusion of these fictitious contact interactions would be undesirable.

We organized this article as follows: We describe briefly the dual simplicial complex representation of proteins structures and discuss the probabilistic models for developing pairwise potentials, then introduce a bootstrap resampling procedure that provides CIs of estimated pairwise potential parameters. We then present the pairwise contact potential, along with experimental results, in discriminating between native and decoy structures using several benchmark data sets. We further examine how pair potentials developed from the dual simplices compare with cutoff contact definitions using sidechain centers. The effects of reduced alphabet size for amino acid residues are then described. We conclude with remarks and discussion.

MODEL AND METHODS

Alpha Contacts From Dual Simplicial Complex

Alpha shape has been successfully applied to study a number of problems of proteins, including void measurement, binding-site characterization, protein packing, electrostatic calculations, and protein hydrations.[14,18,20,26–30] Details of alpha shape have been described elsewhere; here, we only provide a brief description for completeness.

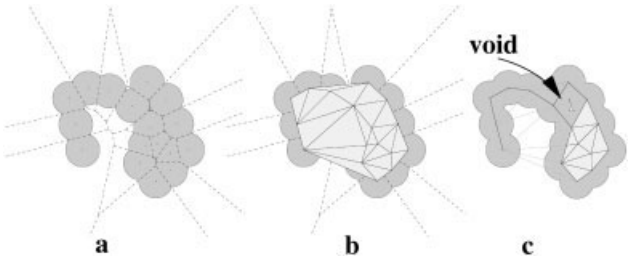

Fig. 1. Geometric constructs of a simple 2D molecule. (**a**) The molecule is formed by disks of uniform size. The dashed lines represent the Voronoi diagram, in which each region contains one atom. (**b**) The convex hull of the atom centers and the Delaunay triangulation of the convex hull. (**c**) The alpha shape of the 2D molecule is a subset of the Delaunay triangulation. It is contained within the molecule and reflects the topologic and metric properties of the molecule. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

To illustrate, Figure 1(a) shows a two-dimensional (2D) molecule formed by a collection of disks of uniform size. Each Voronoi cell is defined by its boundaries, shown as broken lines. Every Voronoi edge is a perpendicular bisector of the line between two atom centers. Each Voronoi cell contains one atom, and every point inside a Voronoi cell is closer to this atom than to any other atom. Three connected Voronoi edges meet at a Voronoi vertex. Another geometric construct, the Delaunay triangulation [Fig. 1(b)], is mathematically dual to the Voronoi diagram and can be explained by the following procedure: For each Voronoi edge, we connect the corresponding two atom centers with a line segment and, for each Voronoi vertex, place a triangle spanning the three atom centers of the three Voronoi cells. Completion of this for all Voronoi edges and vertices gives a collection of line segments and triangles. Together with the vertices representing atom centers, they form the "Delaunay complex," the underlying structure of Delaunay triangulation.

Then, we remove all Delaunay edges (or line segments) where corresponding Voronoi edges of the two atoms do not intersect with the molecule [Fig. 1(c)]. When two atoms are spatially very close, the balls representing the two atoms intersect, and these two atoms have nonzero, two-body volume overlap. When three atoms are spatially very close, they intersect and have nonzero, three-body volume overlap. We further remove all Delaunay triangles for which the corresponding Voronoi vertex of the three atoms is not contained within the molecule. The subset of the Delaunay complex formed by the remaining triangles, edges, and vertices (atom centers) is called the *dual simplicial complex,* or the *alpha complex.* We are interested in identifying only contacting atoms that are spatial nearest neighbors. These, precisely, are atoms with two-body volume overlap whose Voronoi cells intersect. By following the mathematical dual structure (i.e., the edges in the α complex), we can accurately identify all contacting nearest neighbor atom pairs. For convenience, we use a rather arbitrary criterion and declare two residues to be in alpha contact if there is at least one edge connecting these two residues.

Using the alpha shape API kindly provided by Prof. Edelsbrunner, we implemented a program, INTERFACE,

to compute contacting atoms based on precomputed Delaunay triangulation and alpha shape. The Delaunay triangulation is computed with the DELCX program, and the alpha shapes, with the MKALF program. Both can be downloaded from the website at NCSA (http://www.ncsa.uiuc.edu). The van der Waals radii of protein atoms are taken from Tsai et al.[31] We follow Singh and Thornton[32] and increment the van der Waals radii by 0.5 Å. This increment is small and comparable to the resolution of the structure. It enables the modeling of imprecisely determined atomic coordinates without introducing many spurious two-body contacts.

Probabilistic Model for Pairwise Alpha Contact Propensity

The propensity $P(i, j)$ for residue of type i interacting with residue of type j is modeled as the odds ratio of the observed probability $q(i, j)$ and the expected probability $p(i, j)$ of a pairwise alpha contact involving both residues i and j:

$$P(i, j) = \frac{q(i, j)}{p(i, j)}. \quad (1)$$

To compute $p(i, j)$ and $q(i, j)$, we chose a simple null model, in which the observed contacts of different proteins in the entire database are pooled together, and the expected contact numbers are calculated. This is the same as the reference state of composition-independent scale discussed in the literature:[33]

$$q(i, j) = \frac{a(i, j)}{\sum_{i', j'} a(i', j')}. \quad (2)$$

Here, $a(i, j)$ is the number count of atomic contacts between residue types i and j, and $\Sigma_{i', j'} a(i', j')$ is the total number of all atomic contacts. The observed probability is then compared with the random probability $p(i, j)$ that a pair of contacting atoms is picked from residues of types i and j, when chosen randomly and independently.[34]

$$p(i, j) = N_i N_j \cdot \left(\frac{n_i n_j}{n(n - n_i)} + \frac{n_i n_j}{n(n - n_j)} \right), \text{ when } i \neq j \quad (4)$$

$$p(i, j) = N_i N_{i-1} \cdot \frac{n_i n_i}{n(n - n_i)}, \text{ when } i = j, \quad (5)$$

where N_i is the number of interacting residues of type i, n_i is the number of atoms with residue type i, and n is the total number of interacting atoms.

The alpha contact potential between residues i and j is obtained from the propensity value $P(i, j)$ as $-\ln P(i, j)$, and the overall energy of a protein is calculated as

$$E = -\sum_{i,j} \ln P(i, j). \quad (6)$$

Cys-Cys Interactions

In principle, only 210 parameters are necessary for 20 amino acid residues. However, Cys-Cys contact requires special attention. Its propensity value is the largest compared to the other 209 contacts, because a Cys residue tends to form a disulfide bond with another Cys residue. Nevertheless, many Cys-Cys residue pairs in close spatial proximity do not form a disulfide bond. As a result, a misclassification of a nondisulfide Cys–Cys contact as a disulfide bond will affect the overall score considerably, especially for small proteins with abundant Cys residues. The problem associated with the misclassification of Cys–Cys contact has already been discussed in literature.[10]

To avoid assigning the same score to the two different types of Cys–Cys contacts, we introduce a slightly more detailed propensity score for Cys–Cys interactions. Because contacts between C:O, C:N, C:C, and O:O atoms never appear in disulfide bonds, a Cys–Cys contact pair lacking these atomic interactions is classified as a disulfide bond Cys–Cys contact if, in addition, the distance between their SG atoms is less than 2.5 Å. All other Cys–Cys interactions are classified as nonbonded Cys–Cys contact. The propensity values estimated for these two types of Cys–Cys contacts are listed in Table I.

Bootstrap Resampling

Because the sample size of 1045 proteins in PDBSELECT is limited, statistical modeling with approximations may be prone to errors. It is therefore essential to assess reliability of estimated contact potentials. Here, we apply bootstrap techniques to calculate CIs from simulated data sets.[35,36] For alpha contact potential of a specific residue pair (e.g., Trp–Trp), we denote the true value of the contact potential as θ. Our probabilistic model [Eq. (1)] can be regarded as an estimator T that gives the estimated value t from the finite amount of data for θ. Our goal is to calculate a 95% CI for θ.

We resample 1045 proteins independently R times from the set of PDBSELECT proteins, with replacement allowed. We have a simulated data set of $Y_1^*, \ldots Y_R^*$, each containing 1045 proteins. Some structures in the original PDBSELECT set appear multiple times, others appear once, and still others never appear. We estimated the pair contact value θ from each of the R samples and obtained $t_1^*, \ldots, t_R^*$. For an equitailed 95% CI (95% = 1 − 2α, α = 2.5%), we have the *basic bootstrap confidence limits:*

$$(t^*_{(R+1)(1-\alpha)}, t^*_{(R+1)\alpha}). \quad (7)$$

The accuracy of these limits depend on R and how well the distribution of $t^* - t$ agrees with that of $T - \theta$. Perfect agreement occurs only when the distribution of $T - \theta$ does not depend on any unknown variables.

To reduce possible errors due to unknown variables, we use the *studentized bootstrap.* For the rth bootstrapped sample,

$$z_r^* = \frac{t_r^* - t}{\upsilon_r^{*1/2}}. \quad (8)$$

To obtain υ, we bootstrap with replacement again M times the rth sample of the original bootstrap.

$$\upsilon_r^* = \frac{1}{M - 1} \sum_{m=1}^{M} (t_m^* - \bar{t}^*)^2, \quad (9)$$

TABLE I. Alpha Contact Propensity of Pairwise Interactions of Amino Acid Residues

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE
ALA	2.1	1.0	0.9	0.8	1.7	1.0	0.8	1.4	1.0	1.7
ARG	0.9–1.0	0.7	0.8	1.2	0.8	0.9	1.3	0.9	0.8	0.8
ASN	0.9–1.1	0.7–0.8	1.0	0.9	0.8	0.9	0.7	1.0	0.8	0.7
ASP	0.8–0.9	1.1–1.3	0.9–1.0	0.6	0.7	0.8	0.5	0.8	0.8	0.6
CYS	1.5–1.9	0.7–0.9	0.7–0.9	0.6–0.9	[a]15.2	0.8	0.6	1.5	1.3	1.5
GLN	1.0–1.1	0.8–0.9	0.8–1.0	0.7–0.8	0.7–0.9	0.9	0.8	0.9	0.7	0.8
GLU	0.8–0.9	1.2–1.4	0.6–0.7	0.4–0.5	0.5–0.7	0.7–0.8	0.5	0.6	0.7	0.7
GLY	1.3–1.5	0.9–1.0	1.0–1.1	0.8–0.9	1.3–1.7	0.8–1.0	0.6–0.7	1.5	0.8	1.1
HIS	0.9–1.1	0.7–0.9	0.7–0.8	0.8–0.9	1.1–1.5	0.6–0.8	0.6–0.7	0.7–0.9	1.0	0.9
ILE	1.5–1.9	0.7–0.8	0.7–0.8	0.6–0.6	1.4–1.7	0.7–0.9	0.6–0.7	1.0–1.2	0.8–0.9	2.1
LEU	1.5–1.6	0.8–0.9	0.7–0.8	0.6–0.7	1.4–1.7	0.9–1.0	0.7–0.7	1.0–1.2	1.0–1.1	1.9–2.0
LYS	0.8–0.9	0.4–0.5	0.7–0.9	1.2–1.3	0.5–0.7	0.7–0.8	1.3–1.4	0.7–0.8	0.5–0.7	0.7–0.8
MET	1.5–1.7	0.8–1.0	0.7–0.9	0.6–0.8	1.5–2.2	0.9–1.1	0.7–0.8	1.1–1.4	0.9–1.2	1.7–2.1
PHE	1.2–1.4	0.8–1.0	0.7–0.9	0.5–0.6	1.5–1.9	0.8–0.9	0.6–0.6	1.0–1.2	0.9–1.1	1.5–1.7
PRO	0.8–0.9	0.7–0.8	0.5–0.7	0.4–0.5	1.0–1.3	0.7–0.8	0.5–0.6	0.8–0.9	0.7–0.9	0.7–0.8
SER	1.1–1.2	0.7–0.8	0.9–1.0	1.0–1.1	1.1–1.5	0.8–1.0	0.8–0.9	1.1–1.2	0.9–1.1	0.8–0.9
THR	1.1–1.3	0.8–0.9	0.9–1.0	0.8–1.0	0.9–1.2	0.9–1.1	0.8–0.9	1.0–1.2	0.8–1.0	1.0–1.1
TRP	1.0–1.2	1.2–1.5	0.8–1.1	0.5–0.7	1.4–2.1	0.9–1.2	0.6–0.8	1.1–1.4	1.2–1.6	1.2–1.5
TYR	1.0–1.2	1.0–1.1	0.7–0.8	0.7–0.7	1.2–1.5	0.8–1.0	0.6–0.7	1.0–1.2	1.0–1.3	1.2–1.4
VAL	1.5–1.7	0.7–0.8	0.7–0.8	0.5–0.6	1.4–1.7	0.7–0.8	0.6–0.7	1.1–1.2	0.8–0.9	1.8–1.9
[b]$CI_{(i,i)}$	1.9–2.3	0.7–0.8	0.9–1.1	0.5–0.7	13.2–17.3	0.8–1.0	0.5–0.5	1.4–1.6	0.8–1.2	2.0–2.2

	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	1.6	0.9	1.6	1.2	0.8	1.2	1.2	1.1	1.1	1.6
ARG	0.9	0.5	0.9	0.9	0.7	0.8	0.8	1.3	1.1	0.7
ASN	0.7	0.8	0.8	0.8	0.6	1.0	0.9	1.0	0.8	0.7
ASP	0.6	1.3	0.7	0.5	0.5	1.0	0.9	0.6	0.7	0.6
CYS	1.6	0.6	1.8	1.7	1.1	1.3	1.1	1.8	1.4	1.5
GLN	1.0	0.8	1.0	0.8	0.8	0.9	1.0	1.0	0.9	0.8
GLU	0.7	1.3	0.7	0.6	0.6	0.8	0.8	0.7	0.6	0.6
GLY	1.1	0.8	1.2	1.1	0.9	1.2	1.1	1.2	1.1	1.2
HIS	1.0	0.6	1.1	1.0	0.8	1.0	0.9	1.4	1.1	0.9
ILE	1.9	0.7	1.9	1.6	0.7	0.9	1.0	1.3	1.3	1.8
LEU	2.0	0.7	1.8	1.7	0.7	0.9	1.0	1.6	1.3	1.7
LYS	0.7–0.7	0.5	0.7	0.7	0.5	0.8	0.7	0.9	0.9	0.7
MET	1.7–1.9	0.7–0.8	2.4	1.9	0.9	1.0	1.1	1.6	1.5	1.6
PHE	1.7–1.8	0.7–0.8	1.7–2.0	2.0	1.0	1.0	0.8	1.7	1.4	1.6
PRO	0.7–0.8	0.5–0.6	0.8–1.0	0.9–1.0	0.7	0.7	0.7	1.4	1.2	0.8
SER	0.9–1.0	0.7–0.8	0.9–1.2	0.9–1.0	0.6–0.8	1.1	1.0	1.0	0.9	0.9
THR	0.9–1.0	0.6–0.7	1.0–1.2	0.8–0.9	0.6–0.7	1.0–1.1	1.1	0.8	0.8	1.0
TRP	1.5–1.7	0.8–1.0	1.4–1.9	1.5–1.8	1.3–1.5	0.9–1.1	0.8–0.9	1.8	1.4	1.4
TYR	1.3–1.4	0.9–1.0	1.4–1.6	1.3–1.5	1.1–1.3	0.8–0.9	0.8–0.9	1.3–1.5	1.2	1.2
VAL	1.7–1.8	0.6–0.7	1.5–1.7	1.5–1.6	0.8–0.9	0.8–0.9	1.0–1.1	1.3–1.5	1.1–1.3	1.8
$CI_{(i,i)}$	1.9–2.1	0.5–0.5	2.1–2.7	1.8–2.1	0.6–0.8	1.1–1.3	1.0–1.2	1.5–2.1	1.1–1.3	1.7–1.9
[c]P(CC)		$P_{\text{nondisulfide bond}} = 1.8, CI = (1.7, 1.9)$					$P_{\text{disulfide bond}} = 13.3, CI = (12.2, 15.3)$			

The upper triangle of the table lists the propensity values, the lower triangle lists the 95% confidence intervals. The 95% confidence intervals for the diagonal entries are listed separately. The propensity values for the two different types of Cys-Cys contacts are also listed.
[a]The alpha contact propensity of Cys-Cys, if all Cys-Cys conformations are classified as one type.
[b]95% confidence interval of alpha contact propensities between self-pair of amino acid residues.
[c]$P_{\text{nondisulfide bond}}$ is the propensity of Cys-Cys without a disulfide bond; $P_{\text{disulfide bond}}$ is the propensity of Cys-Cys with a disulfide bond.

where $t_1^*, \ldots t_M^*$ are calculated from the second bootstrap sampling. We then use the $(R+1)\cdot\alpha$th order statistic of the simulated values $z_1^*, \ldots z_R^*$, or $z^*_{(R+1)\alpha}$ to obtain the studentized bootstrap CI for θ:

$$(t - \upsilon^{1/2} z^*_{(R+1)(1-\alpha)}, t - \upsilon^{1/2} z^*_{(R+1)\alpha}) \qquad (10)$$

Because M bootstrap samples from the rth sample are needed to obtain υ, the total number of nested bootstrap samples is $(M + 1)\cdot R$. We chose $R = 1000$ and $M = 50$. Altogether, we generated $1001 \times 50 = 50{,}050$ bootstrap samples to calculate the CIs for each of the 209 + 2 pairwise alpha contact propensities.

Database Selection

In this study, we use PDBSELECT from http://www.cmbi.kun.nl /swift/pdbsel,[37,38] which contains 1045 proteins selected from the Protein Data Bank (PDB). The sequence identity between any pair of proteins in PDBSELECT is less than 25%.

RESULTS

Pairwise Alpha Contact Potentials

The pairwise alpha contact propensities are listed in Table I. These are calculated for all residue contacts at least 3 residues away in primary sequence. As expected, Cys–Cys has the highest propensities for contact interactions. Other residue pairs with the highest propensities for contact interactions ($P(i, j) = 1.4–2.5$) are pairs of hydrophobic residues (e.g., Met-Met, Ala-Ala, Ile-Ile, Phe-Phe, Ile-Leu, and Ile-Met). The group of residue pairs with the second highest propensities ($P(i, j) = 1.2–1.3$) are ionizable residues with opposite charges (e.g., Arg-Asp, Arg-Glu, Asp-Lys, and Gly-Lys). Residue pairs with lowest alpha contact propensities ($P(i, j) = 0.4–0.6$) are dominated by pairs of ionizable residues of the same charge (e.g., Arg-Lys, Asp-Glu, Lys-Lys, and Glu-Glu). The group of residue pairs with the second lowest alpha contact propensities ($P(i, j) = 0.5–0.7$) are between ionizable residues and hydrophobic residues (e.g., Asp-Phe, Asp-Ile, Asp-Leu, and Glu-Val). Noticeably, pairs of Pro and ionizable/polar residues also have very low propensity for contacting interactions, probably due to the lack of a backbone-NH for H-bonding interactions.

The CIs of these propensity values given by the studentized bootstrap procedure indicate that most of them are estimated accurately. Among the 209 parameters, excluding Cys–Cys interaction, 95% CIs for 153 contact propensities are <0.2, a very tight interval. The CIs of 36 contact propensities are <0.3. Contact propensities with the largest CIs, around 0.6, are Trp-Trp, Met-Met, Cys-Met, and Cys-Trp.

Correlating and Clustering Similar Amino Acid Residues

The overall behavior of pairwise contact interactions for a specific residue type is determined by its 20 pairwise contact propensity values. These values represent a profile of contact interactions specific for the residue type and can be represented as a 20-D vector x.

We group the 20 types of amino acid residues by the Euclidean distance between the 20 vectors.[34] Figure 2 shows the grouping of the 20 amino acid residues obtained by hierarchical clustering. As an exploratory tool for data analysis, hierarchical clustering can discover interesting and informative grouping patterns in data.[39] In Figure 2, residues that have close contact propensity values to the 20 residue types are grouped together.

The pattern of residue groupings clearly reflects the physical and geometric characteristics of the amino acid residues. Cys residue is different from all other residues, because of its propensity to form disulfide bonds. The rest of the 19 residues can be broadly divided into two well-defined branches of hydrophobic and hydrophilic residues.

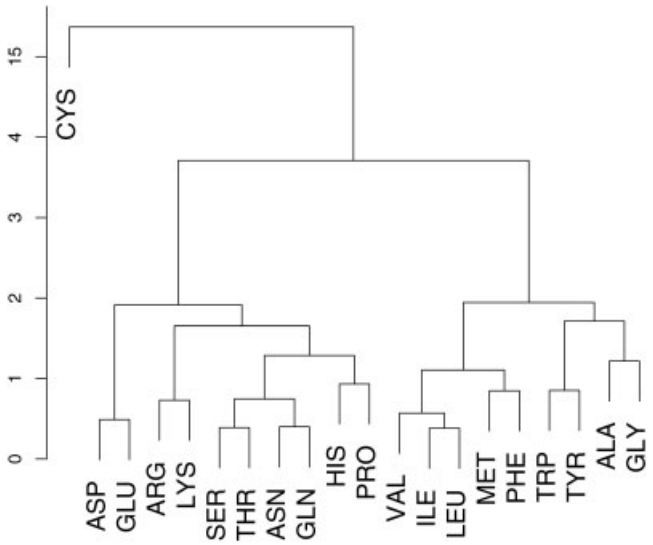

Fig. 2. Grouping of the 20 types of amino acid residues by hierarchical clustering, with complete linkage of the Euclidian distance between their 20D- propensity vectors.

Among the hydrophilic residues, ionizable residues with positive and negative charge are grouped into two small clusters. Hydroxyl-containing residues (Ser and Thr) and amide residues (Asn and Gln) are also clustered into two small clusters. These residues are all capable of forming sidechain hydrogen bonds, and their clusters are neighbors with each other, forming a larger cluster of Ser, Thr, Asn, and Gln. Among the hydrophobic residues, the branched residues (Val, Ile, and Leu) and small residues (Ala and Gly) are grouped into their own clusters. Aromatic residues Trp and Tyr also are grouped into one cluster. Phe has strong hydrophobicity and is grouped with other strongly hydrophobic amino acid residues, rather than clustering with the other two aromatic residues. Pro and His are grouped together, probably because both do not form strong favorable contacts with either hydrophilic or hydrophobic residues.

The clustering pattern of residues by alpha contact propensity also resembles to a certain extent the clustering pattern derived from mutation matrix BLOSUM50,[40] as reported by Murphy et al.[41] For example, Arg with Lys, Asn with Gln, Ser with Thr, and Glu with Asp are all clustered tightly with each other. In addition to the distinct grouping of Cys in our clustering result, a notable difference is that, by alpha contact propensity residues, Ser, Thr, and Pro are grouped with hydrophobic residues instead of hydrophilic residues.

Discriminating Between Native Structures and Decoys

An important method to determine the effectiveness of contact potential functions is to evaluate their success and failure in distinguishing native protein structures from incorrectly folded decoy structures. We use three decoy data sets to assess alpha contact potential.

TABLE II. Discriminating Between Native Structures and Decoys in PROSTAR Data Sets Using Alpha Contact Potential (Alpha)

	Misfold	ifu	Asilomar	pdberr and spga
RAPDF	25/25	32/44	39/42	5/5
CDF	19/25	20/44	36/42	5/5
GC	25/25	21/44	32/42	4/5
MJ	25/25	21/44	34/42	5/5
BT	25/25	22/44	35/42	4/5
Alpha	25/25	24/44	37/42	4/5

Results for each decoy subset are compared with those of other potentials, including RAPDF (residue-specific all-atom conditional probability discriminatory function),[6] CDF (contact discriminatory function),[6] MJ (Miyazawa–Jernigan potential),[5] BT (Betancourt–Thirumalai potential),[55] and GC (geometry center–based potential). The first number in each cell is correctly identified native proteins, and the second is the total number of proteins in the subset. The first row lists the names of the decoy subsets. Data for RAPDF and CDF are taken from Samudrala and Moult (Fig. 1b[6]).

ProStar

The decoy sets in the *ProStar* database[6] contain several subsets. The MISFOLD subset contains conformations of 25 sequences, which are obtained by placing these sequences on structures of different folds with the same number of residues. The conformations of the sidechains are obtained by Monte Carlo sampling.[42] Alpha contact potential succeeded in identifying all 25 native structures correctly (Table II).

For the ASILOMAR subset, alpha contact potential failed to identify 5 native structures out of 42 proteins. For the IFU subset, alpha contact potential failed to identify 20 out of 44 native structures. Alpha contact potential belongs to the class of residue-based potentials, similar to the Miyazowa–potential (MJ), the Betaneourt–Thirumalai potential (BT), and the contact discriminating function (CDF).[6] As pointed out by Lu and Skolnick,[7] decoys in the IFU subset are especially challenging for residue-based potentials, because they are conformations of short loops and have only a small number of residue contact interactions.

In the subsets PDBERR and SGPA, the decoys are structures determined by diffraction but contain serious errors, or are structures generated by molecular dynamics (MD) simulations starting from experimental conformations. Alpha contact potential missed one out of the five native structures.

Park and Levitt Set

This decoy test set contains native and near-native conformations of seven sequences, along with about 650 misfolded structures for each sequence. Park and Levitt generated the positions of C_α in these decoys by exhaustively enumerating 10 selectively chosen residues in each protein, using a 4-state off-lattice model. All other residues were assigned the ϕ/ψ value based on the best fit of a 4-state model to the native chain. Conformations in the decoy sets all have low scores by a variety of scoring functions, and low root-mean square differences (RMSDs) compared to the native structure (Table III).[43]

The results of discrimination test are listed in Table IV and plotted in Figure 3. For five of the seven proteins, the native structures have lowest energy by alpha contact potential. For proteins 3icb and 4rxn, the native structures have the 5th and 51st lowest energy values, respectively. For all proteins, decoys with the lowest energy are within 2.5 Å RMSD of the native structure.

The protein 3icb is a vitamin D-dependent calcium-binding protein. Although the energy of the native structure ranks as the fifth lowest energy, the RMSDs of the first four lowest energy decoys are all within 2.0 Å RMSD of the native structure. For a higher resolution structure (4icb at 1.60 Å) of this protein, the energy of 4icb by alpha contact potential is lower than any of the decoys. It is possible that this misclassification might be due to the lower resolution of structure of 3icb.

Rubredoxin (4rxn) is an iron–sulfur protein. A Fe(III) ion is covalently bound in this structure with four Cys sulfur atoms, preventing them from forming two possible disulfide bonds. Because the protein description and the contact potential do not contain any information about the important covalent bonds of Cys with Fe-S cluster, it is reasonable to expect that native structure will not be of the lowest energy if there is no accounting made for these important covalent bond interactions. It is likely that the structure of rubredoxin might be different without the Fe-S cluster. The decoys at the lowest energy states form one or two fictitious disulfide bridges, and all are near-native structures, with RMSDs around 2 Å of the native structure. MJ and BT potentials work better on 4rxn, because they classify the four Cys–Cys contacts as disulfide bonds.

LATTICE_SSFIT Set

The LATTICE_SSFIT set contains conformations for eight small proteins generated by ab initio protein structure prediction methods.[44,45] The conformational space of a sequence was exhaustively enumerated on a tetrahedral lattice. A lattice-based scoring function was used to select the 10,000 best-scoring conformations. Park and Levitt fitted secondary structures to these conformations using a 4-state model.[43] The 10,000 conformations were further scored with a combination of an all-atom scoring function,[6] a hydrophobic compactness function, and a one-point-per-residue scoring function.[46] The 2000 best-scoring conformations for each protein were selected as decoys for this data set.

Results (Table IV) indicate that for six out of eight proteins, no decoy structures scored better than the native structure. The exceptions are 1fca and 1trl. Similar to 4rxn in the Park and Levitt decoy set, ferrodoxin 1fca contains a Fe-S cluster. Its four Cys residues form four covalent bond with the four Fe(III) irons, instead of two disulfide bridges. These critical contacts, again, are unaccounted for in the alpha contact potential; therefore, the native structure of this protein was not identified successfully. 1trlA is aNMR solution structure of the C-terminal fragment (255–316) of thermolysin. NMR structures are far more difficult to recognize, as discussed in detail by Bastolla et al.[10] They are usually represented as an ensemble of conformations.

TABLE III. Description of Proteins in the 4-STATE-REDUCED, LATTICE-SSFIT, and LMDS Decoy Sets

Decoy set	Protein	Description	N_{res}	N_{decoy}	cRMSD range
4STATE_REDUCED	lctf	C-terminal domain of the ribosomal protein L7/L12	68	630	2.16–10.16
	1r69	N-terminal domain of phage 434 repressor	63	675	2.28–9.50
	1sn3	Scorpion toxin variant 3	65	660	2.50–10.46
	2cro	phage 434 Cro protein	65	674	2.05–9.72
	3icb	Vitamin D–dependent calcium-binding protein	75	653	1.81–10.74
	4pti	trypsin inhibitor	58	687	2.83–10.79
	4rxn	rubredoxin	54	677	2.58–9.28
LATTICE_SSFIT	1beo	β-Cryptogein	98	2000	7.00–15.61
	1ctf	(see above)	68	2000	5.45–12.81
	1dkt-A	Human cyclin–dependent kinase subunit, Type 1	72	2000	6.69–14.05
	1fca	Ferredoxin	55	2000	5.14–11.39
	1nkl	Nk-Lysin	78	2000	5.27–13.64
	1pgb	B1 immunoglobulin-binding domain of streptococcal protein G	56	2000	5.81–12.91
	1trl-A	NMR solution structure of the C-terminal fragment 255–316 of thermolysin	62	2000	5.38–12.52
	4icb	Calcium-binding protein	76	2000	4.74–12.92
LMDS	1b0n-B	Sini protein subunit	39	497	2.45–6.03
	1bba	Pancreatic hormone (AVE. NMR)	36	500	2.78–8.91
	1ctf	(see above)	68	497	3.59–12.53
	1dtk	Dendrotoxin K (NMR)	57	215	4.32–12.58
	1fc2-C	Fragment B of protein A (complexed to immunoglobin Fc)	43	500	4.00–8.45
	1igd	3rd IgG-binding domain from streptococcal protein G	61	500	3.11–12.56
	1shf-A	Fyn Proto-Oncogene Tyrosine Kinase subunit (SH3 domain)	59	437	4.39–12.35
	2cro	(see above)	65	500	3.87–13.48
	2ovo	3rd domain of silver pheasart ovomucoid	56	347	4.38–13.38
	4pti	(see above)	58	343	4.94–13.18

The contact energies of conformations in the ensemble can be substantially different. It is conceivable that an energy function valid for crystal structures cannot reliably recognize native NMR structures.[10] In addition, the structure 1trlA occurs in a dimeric state in the original PDB file. There is substantial interaction between the two chains. Because of this, it is unclear whether a single subunit of 1trlA in a monomeric state would retain the same conformation.

The decoy structures in this data set are generated by ab initio methods. None are near-native, and all have (cRMSD) to the native structure greater than 4.7 Å (Table III). When decoys are so different from the native conformation, energy evaluated with alpha contact potential shows little correlation with the RMSD. The lowest energy decoys in this data set all have large RMSDs, similar to results reported by Samudraia and Levitt.[47] (data not shown).

LMDS Set

The local minima decoy set (LMDS) contains decoys derived from the experimentally obtained secondary structures of 10 small proteins belonging to diverse structural classes. Each decoy is a local minimum of a "handmade" energy function.[48–51] The authors generated ten thousand initial conformations for each protein by randomizing the torsion angles of the loop regions.[52] The adjacent local minima were found by truncated Newton–Raphson minimization in torsion space. Each protein is represented in the decoy set by its 500 lowest energy local minima.

The alpha contact energy function works fairly well in the recognition of 1b0n-B, 1ctf, 1dtk, 2cro, and 2ovo. 1dtk (Dendrotoxin K) contains three disulfide bonds in its native structure. However, in most of its 215 decoys, six Cys residues are spatially arranged together and form on average seven Cys-Cys contacts. For some decoys, up to 15 Cys–Cys contacts by distance cutoff can be found. The ability of discriminating disulfide bonded versus nondisulfide-bonded Cys–Cys contacts probably makes the alpha contact potential discriminate better than the MJ and the BT potentials. The native structure of 1igd (immunoglobulin G–binding domain from streptococcal protein G) ranks first by the MJ and the BT potential, but ranks ninth by alpha contact potential. The recognition of 1bba and 1fc2-C in this set failed for all residue-based contact potentials. 1bba is an atypical structure of a small protein, determined by NMR, which forms a helix with random coil. 1fc2-C is a fragment of protein complexed with an immunoglobulin molecule. It possible that this protein may not maintain the same conformation without the complexed immunoglobulin.

By the criterion of the ranking of native protein, with the exception of the ion-sulfur proteins 4rxn and 1fca, the overall results shown in Tables II and IV indicate that the performance of alpha contact potential in discriminating

TABLE IV. Discriminating Native Structures by Alpha Contact Potential H_α and Potential by Cutoff Distance Between Geometric Centers of Sidechains H_{gc}

		H_α				H_{gc}		H_{MJ}		H_{BT}	
Decoy set	PDB	[a]Rank	[b]Z	[c]$\bar{y}$	[d]f/1000	Rank	Z	Rank	Z	Rank	Z
4STATE_REDUCED	1ctf	1	3.08	1.0	1000	1	3.42	1	3.73	1	3.86
	1r69	1	3.33	1.0	1000	8	2.34	1	4.11	1	4.47
	1sn3	1	3.10	1.0	1000	8	2.49	2	3.17	6	2.97
	2cro	1	3.00	1.0	1000	2	2.91	1	4.29	1	3.92
	3icb	5	2.19	3.4	52	10	2.14	2	2.80	1	2.83
	4pti	1	2.30	1.0	1000	11	2.28	3	3.16	5	2.65
	4rxn	51	1.22	43.0	0	1	2.75	1	3.09	1	3.01
LATTICE_SSFIT	1beo	1	4.74	1.1	923	2	3.69	1	4.74	1	7.29
	1ctf	1	4.62	1.0	1000	1	5.09	1	5.35	1	6.99
	1dkt-A	1	4.33	1.0	1000	15	2.38	32	2.41	5	3.49
	1fca	40	2.01	32.0	0	254	1.18	5	3.40	2	3.92q
	1nkl	1	5.21	1.0	1000	1	7.20	1	5.09	1	7.28
	1pgb	1	3.31	1.0	964	32	2.18	3	3.78	2	3.82
	1trl-A	5	3.35	6.2	0	504	0.63	4	2.91	2	3.82
	4icb	1	4.59	1.0	1000	1	4.11	1	3.67	1	5.07
LMDS	1b0n-B	2	3.13	1.5	525	99	0.85	1	2.65	2	2.50
	1bba	217	0.03	340.8	0	441	−1.11	364	−0.64	234	0.04
	1ctf	1	3.12	1.0	1000	74	1.09	1	3.86	1	3.15
	1dtk	2	2.13	2.2	234	173	−0.92	13	1.71	122	−0.08
	1fc2-C	500	−3.68	500.0	0	480	−1.63	501	−6.24	501	−5.11
	1igd	9	2.43	14.0	0	138	0.61	1	3.25	1	3.76
	1shf-A	17	1.46	8.2	0	322	−0.57	11	1.30	16	1.06
	2cro	1	4.36	1.0	999	159	0.44	1	5.07	1	4.01
	2ovo	3	3.07	5.2	29	326	−1.34	2	3.25	31	1.29
	4pti	9	2.23	7.0	0	242	−0.49	4	2.53	117	0.42

[a]Rank of native structures.
[b]$z = \bar{E} - E_{native}/\sigma$; $\bar{E}$ and σ are the mean and standard deviation of the energy values of conformations, respectively.
[c]Average ranking of native structures in energy evaluated with 1000 bootstrapped potential values.
[d]The number of times of a native structure is ranked to have the lowest energy.

native protein from decoys is better than that of MJ and BT potentials for MISFOLD, IFU, ASILOMAR, 4STATE_REDUCED, and LATTICE_SSFIT sets, and has comparable results for the LMDS set, PDBERR, and SPGA sets.

DISCUSSION

Contact Definition

The alpha contacts introduced in this work are different than contacts by cutoff distances. Atoms in alpha contacts are all within a distance that depends on the identities of the two atoms. Here, this distance is equal to the sum of the van der Waals radii of the two atoms, plus 2×0.5 Å. Unlike contacts by distance cutoff, this distance is not a single, fixed constant but depends on the atom types. Another important distinction of alpha contact is that only a subset of atoms satisfying the distance criterion will be counted as physical nearest neighbors, because we have an additional criterion: Contacting atoms must have intersecting Voronoi cells. Alpha contacts represent the geometry more accurately and can capture contact interactions due to sidechain size and orientation.[25] In addition, no fictitious contacts are introduced between two atoms when there is a third intervening atom.[24] Perhaps this is the reason that alpha contact potential is sensitive to the presence of Fe-S clusters and other hetero atoms, which can be potentially exploited to determine whether a protein structure should contain hetero atoms.

For the 1045 proteins in the PDBSELECT data set, we compared contacts identified by distance cutoff with the threshold of two van der Waals atom radii plus 2×0.5 Å and contacts identified by the alpha shape. We found that about 30–50% of atom contacts detected by distance cutoffs are blocked by a third atom and hence do not have physical interactions. As a result, 3–6% of residue contacts detected by distance cutoffs do not interact physically. Inclusion of these fictitious contact is problematic, especially in developing all-atom contact potentials, as well as in future studies in which higher order interactions in the form of three or four body contacts are incorporated.

Evaluating Discrimination of Alpha Contact Propensities by Bootstraping

How robust are the results of decoy discrimination to the specific values of alpha contact potentials and the specific choices of the structures in the database? We further make use of the bootstrap resampling technique to evaluate the reliability of the discrimination results. As discussed earlier, we resampled 1045 proteins in PDBSELECT independently $R = 1000$ times, with replacement allowed, and

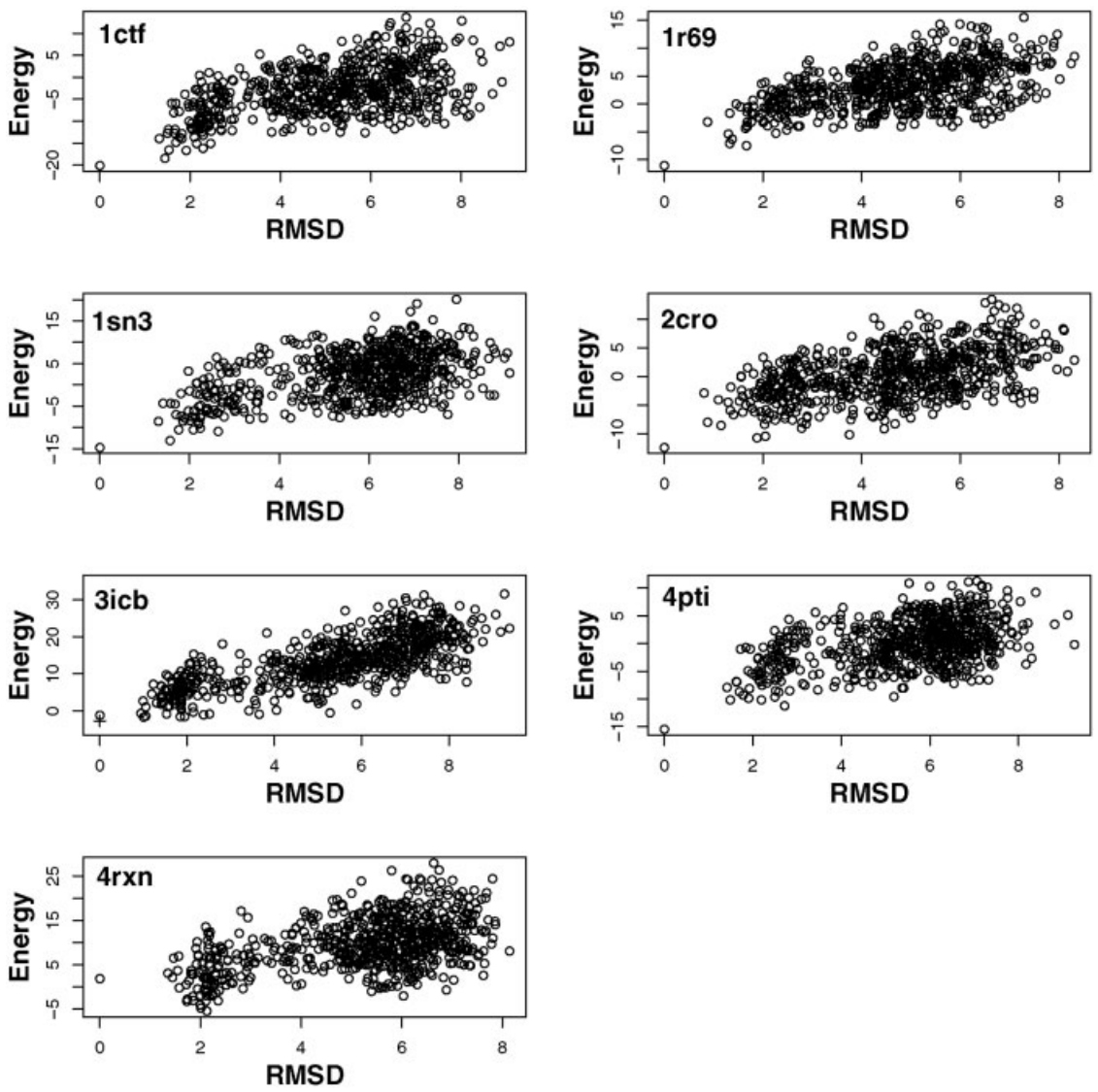

Fig. 3. Energy evaluated by alpha contact potential plotted against the RMSD to native structures for conformations in Park and Levitt decoy set. The alphabet of residues has 20 types of amino acids. For vitamin D–dependent calcium-binding protein (3icb), a structure with better resolution (4icb) has the lowest energy (denoted by +).

obtained 1000 contact propensity matrices. Each was then used to discriminate the decoys in Table III. We use two parameters: $\bar{r}$, the average ranking of the native structure, and *f,* the times a native structure was ranked as having the lowest energy. Table IV indicates that for many decoy sets (e.g., 1ctf, 1r69, 1sn3, 2cro, and 4pti in the 4STATE_REDUCED set), the native structure always ranks first in the 1000 bootstrapped energy evaluations. The performance with 1000 different sets of "bootstrapped" potential values validates the robustness of the method deriving the alpha contact potential and the informativeness of the underlying protein structure database.

Comparison With Contact by Geometric Centers

Contact definition by distance cutoff is widely used in the development of many potential functions. Here, we compare potentials obtained by alpha contact, denoted as H_α, and by contact defined by cutoff distance between geometric centers of sidechains, denoted as H_{gc}. As in the work of Tobi and Elber,[53] two residues are declared to be in contact if the distance d between the geometric centers of their sidechains is $2\text{Å} < d < 6.5$ Å. Geometric center–based contact potential H_{gc} is developed with the same PDBSELECT data, with the same null model as that of alpha contact potential, and, similarly, counting only contact between residues that are three or more residues apart. Therefore, any difference between H_α and H_{gc} is solely due to different geometric representation.

The log values of the parameters of H_α and H_{gc} have an overall correlation coefficient of $p = 0.77$. However, the contact maps of individual proteins obtained by these two different contact definitions are often substantially different. For example, alpha-shape dual simplicial complex gives significantly more contacts than cutoff distance by geometric centers of sidechains for protein 1abe (Fig. 4).

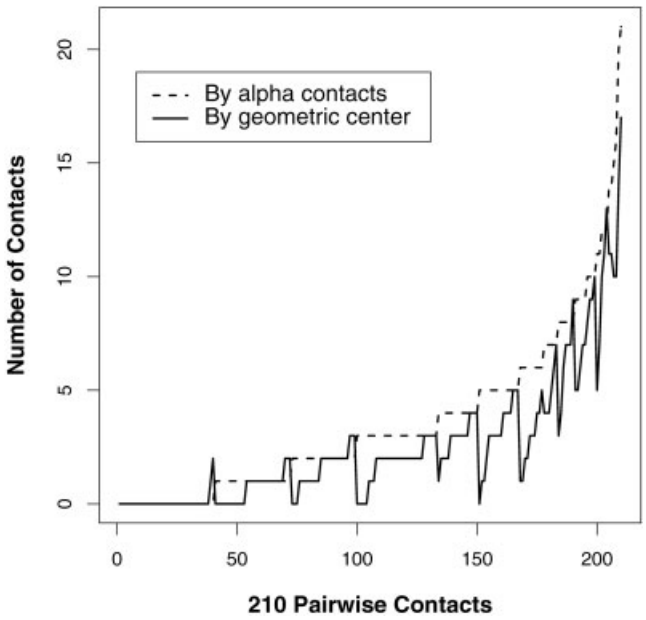

Fig. 4. Difference in contact histograms between the two definitions of alpha contact and contact by geometric centers for protein structure 1abe. Data along x axis are arranged in ascending order by the frequency of contact pairs in the alpha-shape contact model. There are frequently significantly fewer contacts when contact defined by distance between geometric centers is used.

Figure 5 illustrates why such a discrepancy exists between these two contact models. The strong correlation between H_α and H_{gc} is deceptive, and this is reflected in another aspect. Although the correlation coefficient is high, the pairwise contact potentials may have very different values for H_α and H_{gc}. H_{gc} categorically gives much higher propensity values for interactions between small residues. For example, H_{gc} for Gly-Gly and Ala-Gly are 4.55 and 3.04, respectively, but H_α is only 1.48 and 1.39, respectively. Gly-Pro interaction is strongly favorable by H_{gc} [$P(Gly, Pro) = 1.84$] but is unfavorable by H_α (0.87). On the other hand, H_{gc} gives much lower propensity values for interactions between large residues. For example, H_{gc} for Trp-Trp and Phe-Trp is 0.39 and 0.56, respectively, but H_α is 1.75 and 1.65, respectively. In addition, many pair contact interactions between Trp and another residue with large sidechains (such as Arg, Tyr, Phe, His, Leu, Ile, and Met) are unfavorable (<1.0) by H_{gc} but favorable by H_α.

These differences lead to different discrimination in identifying native and near-native protein structures from decoy structures. Because it is impossible to define refined potential for Cys-Cys contacts in the H_{gc} model, as in the alpha contact model, we excluded proteins containing Cys–Cys contacts to avoid complication for comparison. Table IV shows that H_{gc} can only recognize two native structures out of nine proteins in the union set of the 4-STATE decoy set and the LATTICE-SSFIT decoy set. In addition, the z scores for native structures are generally higher for H_α than for H_{gc}. For the 4-STATE Decoy set, the correlation coefficient p by H_α is also higher. These results indicate that geometric description of protein structures is important, and contact model by alpha shape is more

Fig. 5. Alpha contact maps provide accurate geometric descriptions. (**a**) Two Val residues are nearly parallel to each other. Their geometric centers are close enough, but no physical contacts occur between them. (**b**) Two Lys residence are oriented linearly in opposite directions. Their geometric centers are far away from each other, but a hydrogen bond forms between O from one Lys and N from the other. Alpha contact definition correctly identifies (a) as "not in contact" and (b) as "in contact." Contact model by distance cutoff of geometric centers of sidechains gives different contact assignment in both cases.

accurate with more discriminative information for identifying native-like structures. However, the cutoff distance approach may be more convenient to implement, and it is possible to gain further improvement by setting the values of the cutoff threshold as a variable, depending on the type of contacting residue pairs.

Comparison With Other Potentials

We further compare alpha contact potential with several previously developed residue contact potentials, including MJ,[54] BT,[55], SK,[33] and TD[9] potentials. We took the values of MJ potential from Table 3 in Miyazawa and Jernigan.[54] Following the author's recommendation, the average hydrophobicity ($e_{rr} = -2.55$) was subtracted from the potential, and a pair of residues was declared to be in contact if the geometric centers of their sidechains were within an interval of 2.0–6.5 Å. The values of BT potential taken from Zhang et al.[55] were obtained by rescaling MJ potentials with Thr as the reference solvent. The correlation coefficients p and dispersions[55] between each pair of potentials are shown in table Table V. The correlation coefficients p for alpha contact potential log $P(i, j)$ and these residue contact potentials are $p = 0.66$, 0.80, 0.61, and 0.66 for MJ, BT, SK, and TD potentials, respectively. The dispersions as defined (p. 363, Formula 4)[55] are 1.45, 0.28, 0.51, and 0.39, respectively. Because the contact map obtained by alpha edges can be substantially different from other contact definitions (Fig. 4), the absolute value of energy by alpha contact potential and by other potentials for the same structure can be substantially different.

Long-Range Interactions

Interactions between residues with large sequence distance d are relatively rare. We found that they more likely to occur in the interior of a protein than on the surface (data not shown). Identifying such interactions are of particular interest, because they result in significant reduction of conformational entropy. Prediction of protein structures seem to be most difficult for proteins with large contact order,[56] namely, those with significant interactions between residues with large sequence distances.

The bootstrap procedure introduced here provides a reliable method to identify the contact pairs of long sequence separation whose propensity values can be confidently assessed (see Supplementary Online Material for tables of alpha contact potential for $d \geq 30$). Among all

TABLE V. Correlation Coefficients and Dispersions[55] Between Each Pair of Potentials

	Alpha	MJ	BT	GC	SK	TD
Alpha	1/0	0.66/1.45	0.80/0.28	0.77/0.41	0.61/0.51	0.66/0.39
MJ	0.66/1.45	1/0	0.66/1.43	0.37/1.60	0.73/1.29	0.67/1.35
BT	0.80/0.25	0.66/1.43	1/0	0.49/0.56	0.76/0.41	0.63/0.37
GC	0.77/0.41	0.37/1.60	0.49/0.56	1/0	0.15/0.81	0.43/0.63
SK	0.61/0.55	0.76/1.29	0.82/0.41	0.15/0.81	1/0	0.64/0.52
TD	0.66/0.39	0.67/1.35	0.63/0.37	0.43/063	0.64/0.52	1/0

The first number of each cell is the correlation coefficient between each pair of potentials. The second number is the dispersion between each pair of potentials.

possible 210 pairwise interactions, nine contact pairs with high propensity (lower value of 95% CI >1.5) can be reliably assessed for $d > 30$. In addition to Cys–Cys, they include hydrophobic–hydrophobic interactions (Gly-Gly, Met-Met, Ile-Ile, Phe-Phe, Val-Val, Met-Phe), salt–bridge interactions (Arg-Asp, Asp-Lys), and Pro-Trp.

Some long-range interactions are clearly associated with specific secondary structures. After correction for prior probability of being in a particular secondary structure, we found that Met-Met contact has a high propensity to occur between two helices (h) or two β-strands (s) [$P(Met, Met)_{hh} = 2.4$, $P(Met, Met)_{ss} = 2.3$], and a low propensity to occur between either a helix and a coil (c) [$P(Met, Met)_{hc} = 0.65$], or a strand and a coil [$P(Met, Met)_{sc} = 0.56$]. Similarly, because Gly is a helix breaker, long-range Gly-Gly contact has a high propensity to occur between two coils [$P(Gly, Gly)_{cc} = 3.2$], and a low propensity to occur between two helices [$P(Gly, Gly)_{hh} = 0.53$].

Reduced Alphabet for Amino Acid Residues

The clustering of pairwise alpha contact potentials shown in Figure 2 suggests that many residues behave similarly in contact interactions. This points to possible degeneracy of the amino acid alphabet.[16,17] Reduced alphabet is important, because a smaller size alphabet will lead to exponentially more efficient sampling methods in sequence design and protein engineering.[57–61] Many experimental and computational studies have already suggested that a minimum number of amino acid residue types, far less than 20, may be adequate for protein folding.[40,62–64] Wang and Wang examined different ways to reduce the MJ interaction matrix and concluded that by minimizing mismatches, a reduced alphabet of just five amino acid residue types can be used to construct sequences with good foldability and kinetic accessibility.[65] The reduced five-alphabet set coincides with the same alphabet set reported in the work of Riddle et al.[64] in which fully functional constructs for a small, 57-residue β-barrel protein were experimentally obtained when residues in 38 out of 40 selected amino residues were drawn from the alphabet set of I, K, E, A, and G. Murphy et al. further examined reduced alphabets based on BLOSUM50 substitution matrix.[41] When using a variety of reduced alphabets, with size ranging from 10 to 20, they found that there is little loss of the information necessary to select structural homologs in a database of representative protein sequences using dynamic programming–based global alignment.

We continued investigation in this direction and studied the capability of various reduced alphabet sets in discriminating native proteins from decoys. Figure 2 provides a natural way to reduce the residue alphabet set, similar to the approach used by Murphy et al.[41] By placing a horizontal line at different vertical heights, we can obtain a reduced residue alphabet that is determined by the heights of the branching points in the dendrogram from the hierarchical clustering. For example, we can have an alphabet of seven residue types at a height of about 1.5: A = {D,E}, B = {R,K}, C = {S,T, N, Q, H, P}, D = {V, I, L, M, F, W}, E = {W, Y}, F = {A, G}, and G = {C}. An alphabet of two residue types would take Cys as a residue type, grouping everything else into another residue type. An alphabet of three residue types would have Cys, polar residues, and hydrophobic residues.

Does a reduced alphabet still capture the basic information of protein contact interactions? We used Eq. (**1**) to estimate pairwise alpha contact potentials for a reduced alphabet size of 2, 3, 5, 7, 8, 9, 10, 11, 15, and 20, and tested their effectiveness in selecting native structure from decoys in the Park and Levitt[43] data set. Figure 6 shows the results when an alphabet set of nine residue types, plus the two types of Cys-Cys contacts, is used. Remarkably, the discrimination of native conformation from decoys is almost as good as when there are 20 different residue types.

Figure 7 shows the detailed results of discriminating native structure of 3icb from decoys by using potentials derived from different alphabet sets. The average RMSDs of n decoys with lowest energy by potentials of different alphabets are calculated. Smaller average RMSD of these lowest energy decoys indicates that a large fraction of them are near-native structures. This would suggest good discrimination. The top $n = 100$ decoys of lowest energy are all found to have average RMSD very similar to the native structure, regardless of the size of the alphabet. This suggests that alphabets with just a few residues have respectable results in decoy discrimination. Table VI shows the correlation coefficient of energy and RMSDs for all proteins in the Park and Levitt data set[43] using different alphabets. The results indicate that an alphabet of seven residues would perform very similarly to an alphabet with 20 residues in discriminating decoys. Our results extended earlier work in which subjectively defined alphabet sets were used to extract contact residue potentials by an iterative optimization method.[9] We found that potential derived

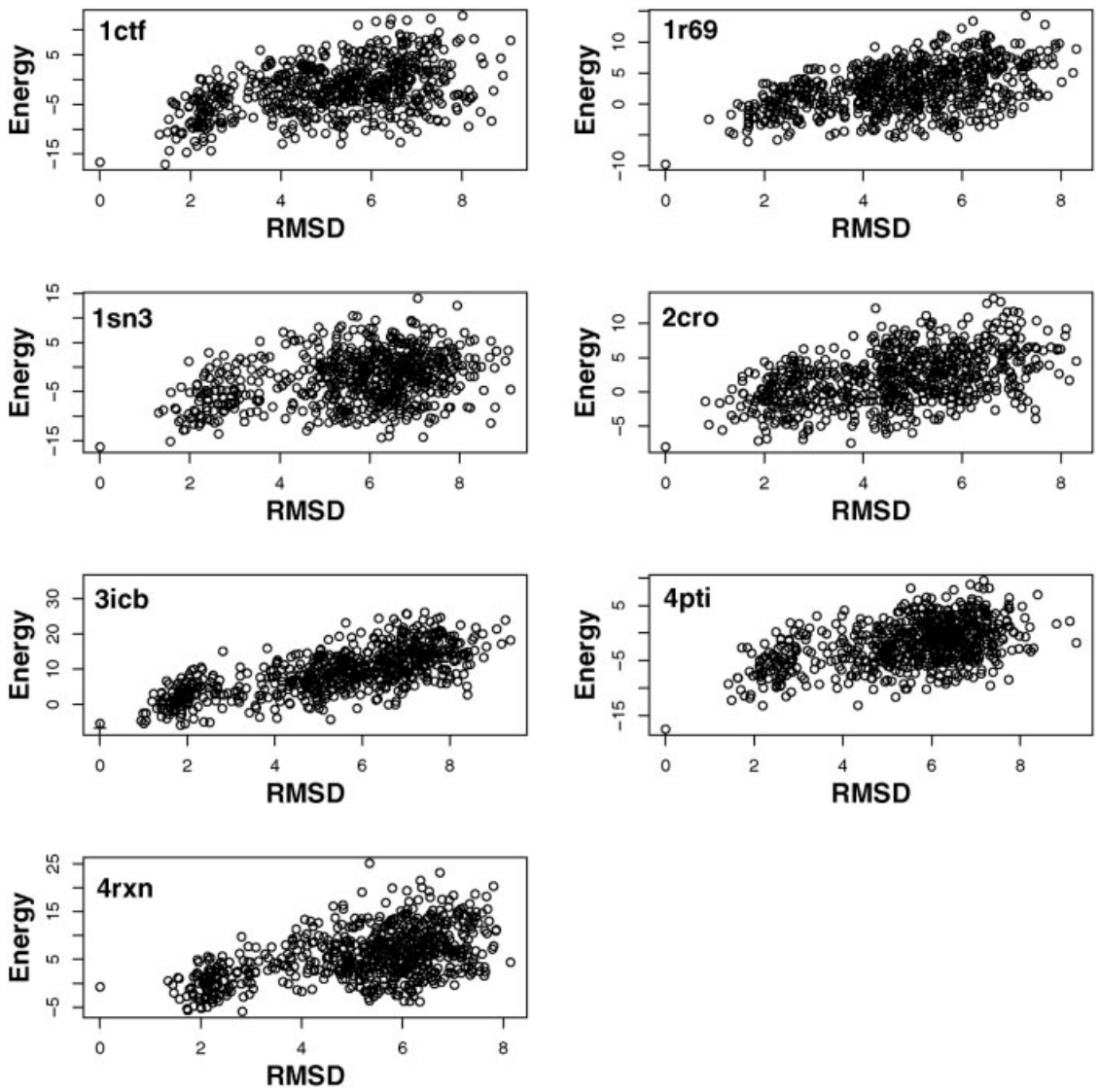

Fig. 6. Energy evaluated by alpha contact potential plotted against the RMSD to native structures for conformations in Park and Levitt decoy set. The alphabet of residues has nine types of amino acids. The discrimination is similar to that shown in Figure 3. 4icb is denoted by + and has the lowest energy.

using such reduced alphabets was effective in discriminating decoys generated by gapless threading. Here, we showed that a similar conclusion can be drawn for statistical potential using alphabet sets derived from natural clustering of residues, in discriminating natives against more stringent compact decoy conformations generated by an off-lattice model.[43] Our conclusion is also consistent with a recent study, which revealed that much of the information in pairwise contact potential is related to just a few variables, such as hydropathy, charge, disulfide bonding, and residue burial.[66]

Although the alphabets we used have different numbers of residues, they were all developed with one aspect in common: The contacts were all derived from dual simplicial complexes, which provide a faithful representation of the geometry. This suggests that as long as the same space-filling pattern is conserved, the specific residue types are not critical in many cases. It seems that packing geometry plays a very important role, but the specific residue types are often replaceable. This observation is consistent with experimental results in which it is well known that proteins are robust against many mutations.

Edge and Tetrahedron Simplices

Pairwise alpha contact potential only considers the edge simplices, or 1-simplice in the dual simplicial complex. There have been several studies of statistical potential based on 3-simplices or tetrahedra.[67–69] In the work of Tropsha et al., 3-simplices are obtained from unweighted Delaunay triangulation.[67–68] In these studies, all residues are treated as balls of equal size located at C_α or C_β positions, and a cutoff distance is used to remove tetrahedra that are considered too large. Our approach is differ-

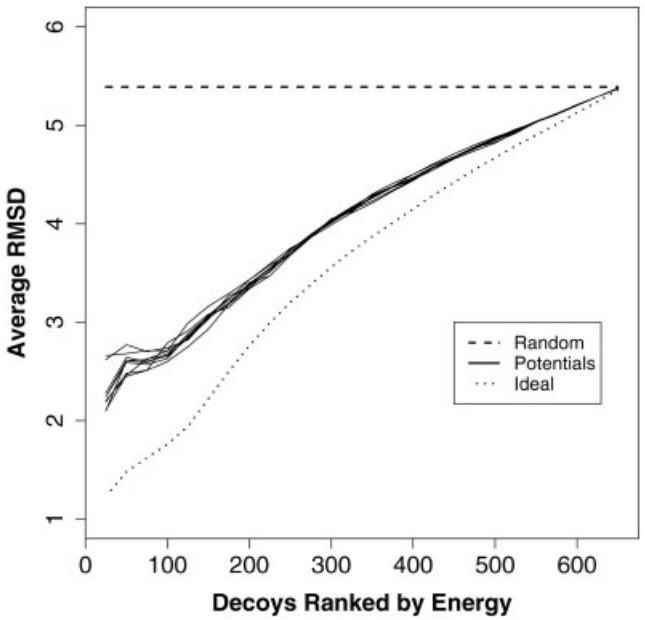

Fig. 7. Discriminating native structure of 3icb from decoys that use different alphabets. For each alphabet, we rank the decoys by their energy. The average RMSD to the native structure is then calculated for the top *n* decoys at different *n* values. Top dashed line represents the average RMSD if the decoys are chosen randomly; dotted line represents the ideal case that *n* decoys with smallest RMSD are chosen by an ideal function; the other lines represent the average RMSD of decoys chosen by potentials of an alphabet with 2–20 amino acid types.

ent. Our model is instead based on the weighted dual simplicial complex, a different simplicial complex formed by a subset of the simplices from the weighted Delaunay triangulation of all atoms in the molecule. The dual simplicial complex or the alpha shape allows modeling at the atomic level. Therefore, in our approach, contacts can be defined by the full sidechain and main-chain atoms. Additionally, atoms are assigned with appropriate nonuniform van der Waals radii.[31] Finally, because the dual simplicial complex reflects the precise contact geometry, we avoid the use of heuristic cutoff thresholds necessary to eliminate a subset of the simplices from the Delaunay triangulation. Our contacts represents accurately geometry of the structure. We discussed earlier the differences between the alpha contact and contact by distance cutoff between geometric centers of sidechains. It is conceivable that similar difference will result between alpha contact and the approach described by Singh et al.[67] and Zheng et al.[68]

SUMMARY

In this work, we introduced a novel representation of protein structures using edge simplices of the alpha shape, or the dual simplicial complex of the protein structure. By describing pairwise contact interactions with simplicial edges, we developed alpha contact potential based on the statistics of edge simplices. We also developed a bootstrap model that provides confidence interval estimations, including those of long-range interactions. We found that alpha contact potential performs well in decoy structure discrimination. In comparison with alternative contact potential, we conclude that geometric representation of contact interaction is important, but the specific residue types are often interchangeable.

TABLE VI. Correlation Coefficients ρ Between Energy Evaluated With Alpha Contact Potential and RMSD to the Native Structures for Decoys in the Park and Levitt set

$^*\|\Sigma\|$	1ctf	1r69	1sn3	2cro	3icb	4pti	4rxn
2	0.31	0.35	0.33	−0.15	0.13	0.40	0.17
3	0.39	0.44	0.48	0.34	0.73	0.44	0.39
5	0.41	0.47	0.47	0.38	0.71	0.47	0.51
7	0.45	0.44	0.48	0.46	0.71	0.49	0.48
8	0.54	0.46	0.42	0.48	0.68	0.52	0.50
9	0.47	0.48	0.47	0.47	0.72	0.47	0.49
10	0.47	0.49	0.47	0.48	0.71	0.48	0.51
11	0.47	0.50	0.49	0.47	0.72	0.49	0.51
15	0.49	0.50	0.49	0.50	0.71	0.47	0.49
20	0.49	0.52	0.49	0.50	0.74	0.46	0.53

Alpha contact potentials with alphabet containing different number of residue classes are used.
$^*|\Sigma|$ denotes the number of residue classes in the alphabet Σ as obtained from Figure 2.

ACKNOWLEDGMENT

We thank Dr. Hui Lu for very helpful discussions.

REFERENCES

1. Skolnick J, Kolinski A. Monte Carlo methods in chemical physics. Adv Chem Phys 1999;105:203–242.
2. Jorgensen W, Tirado-Rives J. The OPLS potential function for proteins: Energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc 1988;110:1657–1666.
3. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 1999;288:477–487.
4. Gatchell D, Dennis S, Vajda S. Discrimination of near-native protein structures from misfolded models by empirical free energy functions. Proteins 2000;41:518–534.
5. Miyazawa S, Jernigan R. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 1985;18:534–552.
6. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
7. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001; 44:223–232.
8. Goldstein R, Luthey-Schulten Z, Wolynes P. Protein tertiary structure recognition using optimized hamiltonians with local interactions. Proc Natl Acad Sci U S A 1992;89:9029–9033.
9. Thomas P, Dill K. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci U S A 1996;93:11628–11633.
10. Bastolla U, Farwer J, Knapp E, Vendruscolo M. How to guarantee optimal stability for most representative structurs in the Protein Data Bank. Proteins 2001;44:79–96.
11. Dima R, Banavar J, Cieplak M, Maritan A. Scoring functions in protein folding and design. Protein Sci 2000;9:812–819.
12. Micheletti C, Seno F, Banavar J, Maritan A. Learning effective amino acid interactions through iterative stochastic techniques. Proteins 2001;42:422–431.
13. Edelsbrunner H, Mücke E. Three-dimensional alpha shapes. ACM Trans Graphics 1994;13:43–72.
14. Liang J, Edelsbrunner H, Fu P, Sudhakar P, Subramaniam S. Analytical shape computing of macromolecules I: Molecular area and volume through alpha-shape. Proteins 1998;33:1–17.
15. Wernisch L, Hunting M, Wodak S. Identification of structural domains in proteins by a graph heuristic. Proteins 1999;35:338–352.

16. Riddle D, Santiago J, Grantcharova V, Baker D. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 1997;4:805–809.
17. Fink T, Ball R. How many conformations can a protein remember? Phys Rev Lett 2001;87:198103-1–198103-4.
18. Liang J, Edelsbrunner H, Fu P, Sudhakar P, Subramaniam S. Analytical shape computing of macromolecules II: Identification and computation of inaccessible cavities inside proteins. Proteins 1998;33:18–29.
19. Edeslbrunner H, Facello M, Liang J. On the definition and the construction of pockets in macromolecules. Disc Appl Math 1998; 88:18–29.
20. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. Protein Sci 1998;7:1884–1897.
21. Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. In: Proceedings of the 41st Annual IEEE Symposium Found Computer Science; 2000. pp 454–463.
22. Edelsbrunner H, Zomorodian A. Computing linking numbers in a filtration. In: Algorithms in bioinformatics (LNCS 2149). Berlin: Springer; 2001. pp 112–127.
23. Edelsbrunner H, Harer J, Zomorodian A. Hierarchical Morse complexes for piecewise linear 2-manifolds. In: Proceedings of the 17th Annual ACM Symposium Computer Geometry; 2001. p 70–79.
24. Bienkowska J, Rogers R, Smith T. Filtered neighbors threading. Proteins 1999;37:346–359.
25. Taylor W. Multiple sequence threading: An analysis of alignment quality and stability. J Mol Biol 1997;269:902–943.
26. Edelsbrunner H, Facello M, Fu P, Liang J. Measuring proteins and voids in proteins. In: Proceedings of the 28th Annual Hawaii International Conference System Sciences. Vol. 5. Los Alamitos, CA: IEEE Computer Scociety Press; 1995. p 256–264.
27. Peters K, Fauck J, Frömmel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
28. Liang J, Dill K. Are proteins well-packed? Biophys J 2001;81:751–766.
29. Liang J, Subranmaniam S. Computation of molecular electrostatics with boundary element methods. Biophys J 1997;73:1830–1841.
30. Liang J, McGee M. Mechanisms of coagulation factor Xa inhibition by antithrombin: Correlation between hydration structure and water transfer during reactive loop insertion. Biopys J 1998;75:573–582.
31. Tsai J, Taylor R, Chothia C, Gerstein M. The packing density in proteins: Standard radii and volumes. J Mol Biol 1999;290:253–266.
32. Singh J, Thornton J. Atlas of protein side-chain interactions. 2 vols. Oxford: IRL Press; 1992.
33. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.
34. Adamian L, Liang J. Helix–helix packing and interfacial pairwise interactions of residues in membrane proteins. J Mol Biol 2001;311: 891–907.
35. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
36. Davison A, Hinkley D. Bootstrap methods and their applications. Cambridge, UK: Cambridge University Press;1997.
37. Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci 1992;1:409–417.
38. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
39. Kaufman L, Rousseeuw P. Finding groups in data. New York: John Wiley & Sons; 1990.
40. Henikoff S, Henikoff J. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89:10915–10919.
41. Murphy L, Wallqvist A, Levy R. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 2000;13:149–152.
42. Holm L, Sander C. Evaluation of protein models by atomic solvation preference. J Mol Biol 1992;225:93–105.
43. Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
44. Samudrala R, Xia Y, Levitt M, Huang E. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. Pac Symp Biocomput 1999;.
45. Xia Y, Levitt M. Extracting knowledge-based energy functions from protein structures by error rate minimization: Comparison of methods using lattice model. J Chem Phys 2000;113:9318–9330.
46. Park B, Huang E, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
47. Samudrala R, Levitt M. Decoys 'R' us: A database of incorrect conformations to improved protein structure prediction. Protein Sci 2000;9:1399–1401.
48. Levitt M, Lifson S. Refinement of protein conformations using a macromolecular energy minimization procedure. J Mol Biol 1969; 46:269–279.
49. Levitt M. Energy refinement of hen egg-white lysozyme. J Mol Biol 1974;82:392–420.
50. Levitt M. Molecular dynamics of native protein: I. Computer simulation of trajectories. J Mol Biol 1983;168:595–620.
51. Levitt M, Hirshberg M, Sharon R, Daggett V. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comp Phys Comm 1995;91:215–231.
52. Fletcher R. A new approach to variable metric algorithms. Comput J 1970;13:317–322.
53. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: Results from linear optimization. Proteins 2000;41:40–46.
54. Miyazawa S, Jernigan R. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term. J Mol Biol 1996;256:623–644.
55. Betancourt M, Thirumalai D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;8:361–369.
56. 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.
57. Yue K, Dill K. Inverse protein folding problem: Designing polymer sequences. Proc Natl Acad Sci U S A 1992;89:4163–4167.
58. Shakhnovich E, Gutin A. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci U S A 1993;90: 7195–7199.
59. Deutsch J, Kurosky T. New algorithm for protein design. Phys Rev Lett 1996;76:323–326.
60. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science 1996;273: 666–669.
61. Mélin R, Li H, Wingreen N, Tang C. Designability, thermodynamic stability, and dynamics in protein folding: A lattice model study. J Chem Phys 1999;110:1252–1262.
62. Sander C, Schulz G. Degeneracy of the information contained in amino acid sequences: Evidence from overlaid genes. J Mol Evol 1979;13:245–252.
63. Heinz D, Baase W, Matthews B. Folding and function of a t4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence. Proc Natl Acad Sci U S A 1992;89:3751–3755.
64. Riddle D, Santiago J, Bray-Hall S, Doshi N, Grantcharova V, Yi Q, Baker D. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 1997;4:805–809.
65. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 1999;6:1033–1038.
66. Cline M, Karplus K, Lathrop R, Smith T, Rogers R Jr: Information-theoretical dissection of pairwise contact potentials. Proteins 2002;49:7–14.
67. Singh R, Tropsha A, Vaisman I. Delaunay tessellation of proteins: Four body nearest-neighbor propensities of amino-acid residues. J Comp Bio 1996;3:213–221.
68. Zheng W, Cho S, Vaisman I, Tropsha A. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. In: Altman R, Dunker A, Hunter L, Klein T, editors. Pacific Symposium on Biocomputing' 97. Singapore: World Scientific; 1997. p 486–497.
69. Carter C Jr, LeFebvre B, Cammer S, Tropsha A, Edgell M. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol 2001;311:625–638.