Geometric cooperativity and anticooperativity of three-body interactions in native proteins

Geometric Cooperativity and Anticooperativity of Three-Body Interactions in Native Proteins

Xiang Li and Jie Liang*

Department of Bioengineering, SEO, MC-063, University of Illinois at Chicago, Chicago, Illinois

Abstract **Characterizing multibody interactions of hydrophobic, polar, and ionizable residues in protein is important for understanding the stability of protein structures. We introduce a geometric model for quantifying 3-body interactions in native proteins. With this model, empirical propensity values for many types of 3-body interactions can be reliably estimated from a database of native protein structures, despite the overwhelming presence of pairwise contacts. In addition, we define a nonadditive coefficient that characterizes cooperativity and anticooperativity of residue interactions in native proteins by measuring the deviation of 3-body interactions from 3 independent pairwise interactions. It compares the 3-body propensity value from what would be expected if only pairwise interactions were considered, and highlights the distinction of propensity and cooperativity of 3-body interaction. Based on the geometric model, and what can be inferred from statistical analysis of such a model, we find that hydrophobic interactions and hydrogen-bonding interactions make nonadditive contributions to protein stability, but the nonadditive nature depends on whether such interactions are located in the protein interior or on the protein surface. When located in the interior, many hydrophobic interactions such as those involving alkyl residues are anticooperative. Salt-bridge and regular hydrogen-bonding interactions, such as those involving ionizable residues and polar residues, are cooperative. When located on the protein surface, these salt-bridge and regular hydrogen-bonding interactions are anticooperative, and hydrophobic interactions involving alkyl residues become cooperative. We show with examples that incorporating 3-body interactions improves discrimination of protein native structures against decoy conformations. In addition, analysis of cooperative 3-body interaction may reveal spatial motifs that can suggest specific protein functions. Proteins 2005;60:46–65.**

Key words: higher order interactions; many-body interactions; 3-body potential; nonadditive effects; nonadditive coefficient; cooperativity; anticooperativity

INTRODUCTION

Interactions stabilizing proteins are often modeled by pairwise contacts at the atom or residue level.[1–3] To reflect the physicochemical nature of protein stabilizing forces, empirical potentials based on statistics of pairwise interactions have found wide use in studying protein folding, in structure prediction, and in sequence design.[1,4–8] An assumption associated with this approach is the additivity of pairwise interactions, namely, that the total energy or fitness score of a protein is the linear sum of all of its pairwise interactions.

The inadequacy of considering only pairwise interactions is widely recognized[9]: Pairwise interactions cannot explain 2-state kinetics of protein folding[10,11]; the nonadditivity effects have been clearly demonstrated in cluster formation of hydrophobic methane molecules,[11–14] and protein structure refinement likely will require higher order interactions.[3] Recognizing the necessity of including higher order interactions, 3-body contacts have been introduced in several studies,[15–18] where physical models explicitly incorporating 3-body interactions are developed. In addition, several studies of Delaunay 4-body interactions clearly showed the importance of including higher order interactions in explaining the observed frequency distribution of residue contacts, and in protein fold recognition.[19–24]

However, the overall characteristics of 3-body interactions in proteins remain poorly understood. For example, a simple system of 3 methanes has been the focus of extensive studies,[11–14] but no consensus has yet emerged from these studies on the cooperative or anticooperative nature of hydrophobic 3-body interactions. Polar interactions are also of great importance for protein stability and protein function.[25–31] The cooperative nature of polar interactions is also not well understood.

The lack of understanding of higher order interactions is due to several technical bottlenecks: It is difficult to develop a precise physical model at an atomic level for higher order interactions in proteins that can be easily computed. A simple method using distance cutoff would

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat

Grants sponsor: National Science Foundation. Grant numbers: CAREER DBI0133856 and DBI0078270). Grant sponsor: National Institute of Health; Grant number: GM68958. Grant sponsor: Office of Naval Research; Grant number: N000140310329.

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

Received 16 June 2004; Accepted 30 November 2004

Published online 22 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20438

include many noncontacting 3-bodies with no physical interactions,[7,32] thus making it difficult to isolate the lone effects of 3-body interactions. Studies based on the distribution of Delaunay tetrahedra of C_α atoms do not model physical atomic contact but rely on the cutoff threshold of length measure of tetrahedra, which depends on residue distances between C_α atoms. It is challenging to detect robustly the subtle effects of higher order interactions amid much more common pairwise interactions.

In this study, we introduce a simple geometric model for 3-body interactions and focus on interactions defined by volume overlaps. Our approach is a knowledge-based bioinformatics approach. Although it is different from a physics-based approach derived from first principles, the results obtained are expected to be applicable in bioinformatics applications where empirical statistical potential functions play important roles. These include fold recognition,[7,33,34] protein design,[35–37] modulating protein–protein interactions,[38] and ab initio structure prediction.[39] As examples, preliminary results using 3-body potential in improving discrimination of native proteins from decoys and in discovery of spatial motifs of proteins will be shown later in this work. In our model, higher order interactions are defined not based on distance cutoffs, but based on the topological structure of the protein.[40,41] Atoms have to be in physical nearest neighboring contact and must have volume overlap. Technically, pairwise contact occurs in this case if 2 atoms from nonbonded residues share a Voronoi edge, and this edge is at least partially contained in the body of the 2 atoms.[7] Three-body contact with volume overlap occurs if 3 atoms from Nonbonded residues share a Voronoi vertex, and this vertex is contained in the body of the 3 atoms. We calculate empirical statistical propensities of 3-body interaction for all 1540 possible 3-body contacts in the protein interior and on the protein surface, along with the 95% confidence intervals of these estimated parameters.[7] A nonadditivity coefficient for each type of 3-body interaction is also introduced and computed. This coefficient represents the deviation of 3-body interaction from linear sum of pairwise interactions. It compares the 3-body propensity value from what would be expected if only pairwise interactions were considered, and highlights the distinction of propensity and cooperativity of 3-body interaction. Three-body cooperativity studied here is different from the kinetic cooperativity studied in the folding of 2-state proteins, as discussed in Shimizu and Chan.[11]

Our results indicate that there are many favorable 3-body interactions with high-propensity values; that is, these 3 residues have a stronger tendency to have 3-body interactions than would be expected by random chance. However, some of these favorable 3-body interactions are anticooperative. These residues already have favorable pairwise interactions. Although they have overall favorable propensity for 3-body interactions, their propensity values may be significantly less than what would be expected if only favorable propensity for pairwise interactions were summed up and there were no additional 3-body effects.

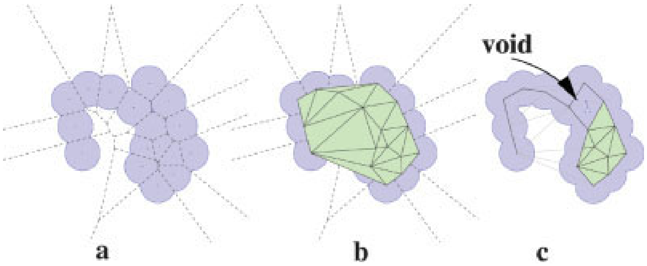

Fig. 1. Geometric constructs of a simple 2D molecule. (**a**) The molecule is formed by disks of uniform size. The dashed lines represent the Voronoi diagram, where each region contains one atom. (**b**) The Delaunay triangulation of the molecule. (**c**) The α shape of the 2D molecule is a subset of the Delaunay triangulation. It is contained within the molecule and reflects the topological and metric properties of the molecule.

We also find that there are many unfavorable 3-body interactions. These are cases where three residues have a lower tendency to have 3-body interactions than would be expected by random chance. However, some of these unfavorable 3-body interactions are positively cooperative. These residues have unfavorable pairwise interactions. Although they have overall unfavorable propensity for 3-body interactions, their propensity values for interaction may be significantly better than what would be expected if only unfavorable propensity for pairwise interactions were summed up and there were no additional 3-body effects.

Anticooperative but favorable 3-body interactions and cooperative but unfavorable 3-body interactions are examples of nonadditive effects in higher order interactions of amino acid residues. We find that such nonadditive effects depend on the geometric location of the 3-body interactions. For example, hydrophobic 3-body interactions are favorable both on the protein surface and in the protein interior. However, those located in the interior are often anticooperative, and those on surface, are cooperative. Another example is the hydrogen bond between N and O atoms, which contributes to cooperativity in the protein interior but to anticooperativity on the protein surface.

We organize this article as follows: We first discuss results of propensities and nonadditivity coefficients for interior and surface 3-body interactions. Next we show that a model empirical potential function incorporating 3-body interactions can improve the discrimination of native structure against benchmark decoy structures. We then highlight a cooperative spatial motif formed by 3-body interactions, and conclude with remarks and discussion. Details of the model and calculation of propensities and nonadditivity coefficients for 3-body interactions are presented in the next section.

MODEL AND METHODS

Alpha Contacts From Dual Simplicial Complex

Figure 1(a) shows a 2-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 2 atom centers. Each Voronoi cell contains 1 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, connect the corresponding 2 atom centers with a line segment, and for each Voronoi vertex, place a triangle spanning the 3 atom centers of the 3 Voronoi cells. Completing this for all Voronoi edges and Voronoi vertices gives a collection of line segments and triangles. Together with the vertices representing atom centers, they form the "Delaunay complex," which is the underlying structure of Delaunay triangulation.

Now we remove any Delaunay edge (or line segments) where the corresponding Voronoi edge of the 2 atoms does not intersect with the molecule [Fig. 1(c)]. When 2 atoms are spatially very close, the disks representing the 2 atoms intersect, and these 2 atoms have nonzero, 2-body volume overlap. When 3 atoms are spatially very close, they intersect and have nonzero, 3-body volume overlap. We further remove all Delaunay triangles where the corresponding Voronoi vertex of the 3 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 α *complex*. We are interested in identifying only contacting atoms that are spatial nearest neighbors.

We define 3-body interactions based on the nearest neighbor spatial relationship of atoms. In Figure 2, where a 2D molecule is shown, 2 neighboring atoms share a Voronoi edge, and 3 neighboring atoms share a Voronoi vertex. Three-body interactions therefore can be identified by examining the corresponding Voronoi vertices. However, atoms sharing a Voronoi vertex may still be distantly located and have no physical contact interactions (Fig. 2, atoms A, B, and C, and Voronoi vertex v). Atoms with volume overlap can be identified as those whose shared Voronoi vertices are contained within the body of the 3 atoms. Atom triplets satisfying this criterion are guaranteed to have physical contact interactions. Mathematically, 3-body contacts can be mapped from the 2-simplices σ_2 (or triangles) in the dual simplicial complex (or the α shape) of the protein structure.[42] We call these *3-body* α *contacts*. Similarly, we define 2-body α contact as atom pairs whose Voronoi edge is partially contained within the body of the 2 atoms. Atom pairs satisfying this criterion are guaranteed to have physical contact interactions with volume overlaps. These can be mapped from the 1-simplices σ_1s (or edges) in the simplicial complex of the protein. In 3D space, Voronoi vertices corresponding to 3 atom contacts become Voronoi edges. See Edelsbrunner and coworkers[40–43] for more details of the α shape theory.

The geometric model for the 3-body interactions is based on the same underlying shape models for molecules, namely, the union of ball model, which was first proposed by Lee and Richards.[44] It is the dominant shape model that is currently used in all studies of protein structures. This shape model assumes that each atom takes the shape of a ball, which may be assigned a specific radius (e.g., van der Waals radius). The body of the molecule takes the shape of the union of these atom balls, where spatially neighboring atom balls may intersect with each other. Another widely used surface model, namely, Connolly's surface, is combinatorially equivalent to this model.[41,42] We believe that the usual assumption that protein takes the shape of the union of balls is quite realistic and is the only workable model available. Under this assumption, our model for 3-body interactions is an exact account of common intersections involving 3 balls for a molecule of arbitrary shape. Unlike heuristics, such as grid-based approaches, there is no approximation involved in our 3-body calculations, because they are all computed from analytical formulae.

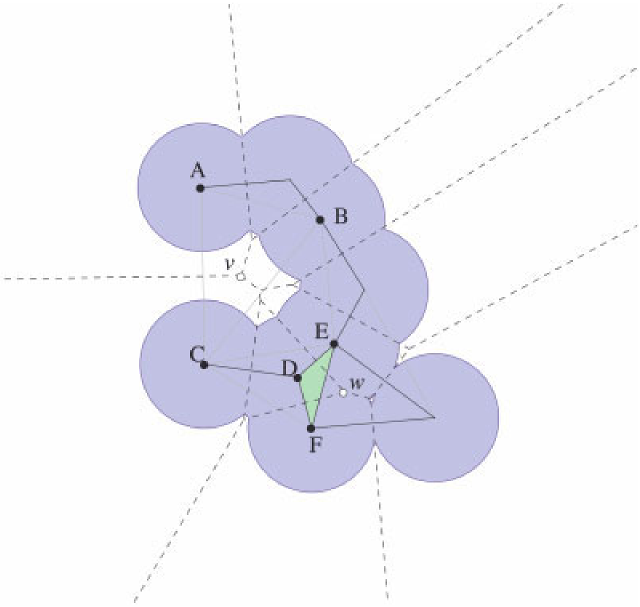

Fig. 2. Geometric model for 3-body interactions. The Voronoi diagram of a 2D simple molecule is shown by the dashed line, and the α shape (represented by solid line segments and triangle) of this molecule is also shown. Voronoi cells of atoms A, B, and C meet at Voronoi vertex *v*. The latter is an indicator that these atoms are spatial nearest neighbors. However, these atoms have no physical contact interactions (i.e., there is no volume overlap). Correspondingly, the Voronoi vertex *v* is not contained within the body of the 3 atoms, A, B, and C. Another Voronoi vertex *w* is formed by Voronoi cells of atoms D, E, and F. Unlike Voronoi vertex *v*, *w* is contained within the body of the union of atoms D, E, and F, which are in physical contact and have volume overlap. Note that the dual 2-simplex (triangle) spanning the atom centers of D, E, and F is part of the α shape of this molecule.

Using the α shape application program interface kindly provided by Edelsbrunner and colleagues, a program INTERFACE3, has been implemented to compute atomic triplets. INTERFACE3 uses precomputed Delaunay triangulation and α shape. The Delaunay triangulation of the proteins is computed using the DELCX program,[40,45] and the α shape is computed using the MKALF program.[40,46] Both can be downloaded from the website of NCSA (http://biogeometry.cs.duke.edu/). The van der Waals radii of protein atoms are taken from Tsai et al.[47] To account for uncertainty in the precision of atomic coordinates, the van der Waals radii are incremented by 0.5 Å, following Singh and Thornton.[48] Larger increments (e.g., 1 Å) introduce spurious triplets of amino acid residues that are not packed tightly.

Probabilistic Model for 2-Body and 3-Body α Contact Propensities

The development of pairwise α contact propensities has been reported before.[7] Briefly, pairwise contact 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 $q(i,j)$ of a pairwise α contact involving both residue i and j, and the expected probability $p(i,j)$ from a null model:

$$P(i,j) \equiv \frac{q(i,j)}{p(i,j)}. \qquad (1)$$

We pool together the observed contacts from different proteins in the entire database. Details of the null model and the computation of $p(i,j)$, and $q(i,j)$ can be found in Li et al.[7] and Adamian and Liang.[49]

We use the same strategy to compute 3-body α contact propensities, which were described previously in the development of 3-body α contact potential for interhelical interactions of membrane protein.[32] The 3-body propensity $P(i,j,k)$ for residues of type i, j, and k is modeled as the odds ratio of the observed probability $q(i,j,k)$ of a 3-body (triple) atomic α contact involving residue i, j, and k, and the expected probability $p(i, j, k)$:

$$P(i,j,k) \equiv \frac{q(i,j,k)}{p(i,j,k)}. \qquad (2)$$

To compute the observed probability $q(i,j,k)$, we have:

$$q(i,j,k) = \frac{a(i,j,k)}{(\Sigma_{i',j',k'},\ a(i',j',k')}$$

where $a(i,j,k)$ is the number count of atomic contacts among residue types i, j, and k, and $\Sigma_{i',j',k'}\, a(i',j',k')$ is the total number of all-atomic 3-body contacts. We exclude 3-body contacts from 3 residues if 2 of them are sequence neighbors.

The observed probability $q(i,j,k)$ is then compared against the random probability $p(i,j,k)$ that 3 atoms are picked from a residue of type i, a residue of type j, and a residue of type k, when chosen randomly and independently from the pooled database.[49] This is similar to the reference state of composition-independent scale discussed in Skolnick et al.[50] When all 3 interacting residues are of different types (e.g., $i\ j\ k$), we have

$$\begin{aligned} p(i,j,k) &= N_iN_jN_k \\ &\times \left\{ \frac{n_in_jn_k}{n(n-n_i)(n-n_i-n_j)} + \frac{n_in_jn_k}{n(n-n_i)(n-n_i-n_k)} \right. \\ &+ \frac{n_in_jn_k}{n(n-n_j)(n-n_j-n_i)} + \frac{n_in_jn_k}{n(n-n_j)(n-n_j-n_k)} \\ &\left. + \frac{n_in_jn_k}{n(n-n_k)(n-n_k-n_i)} + \frac{n_in_jn_k}{n(n-n_k)(n-n_k-n_j)} \right\}, \end{aligned}$$

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

When 2 of the 3 interacting residues are of the same type (i.e., $i = j \neq k$), we have

$$\begin{aligned} p(i,j,k) = N_i(N_i-1)N_k \cdot &\left\{ \frac{n_in_in_k}{n(n-n_i)(n-2n_i)} \right. \\ &\left. + \frac{n_in_in_k}{n(n-n_i)(n-n_i-n_k)} + \frac{n_in_in_k}{n(n-n_k)(n-n_k-n_i)} \right\}. \end{aligned}$$

When all 3 residues are of the same type (i.e., $i = j = k$), we have

$$p(i,j,k) = N_i(N_i-1)(N_i-2) \cdot \frac{n_in_in_i}{n(n-n_i)(n-2n_i)}$$

Nonadditivity Coefficient v of 3-Body Interaction

If 3-body interaction is a simple consequence of 3 independent pairwise contact interactions, the propensity $P'(i,j,k)$ for such interaction would be

$$P'(i,j,k) = P(i,j) \cdot P(j,k) \cdot P(i,k). \qquad (3)$$

In order to investigate the cooperative and anticooperative effects of 3-body interactions, we define the following nonadditive coefficient v of 3-body interactions:

$$v(i,j,k) \equiv \frac{P(i,j,k)}{P'(i,j,k)}. \qquad (4)$$

This is similar to the definition used in Shimizu and Chan.[13] There are 3 possibilities: (1) $v \approx 1$: Interaction of a triplet type is additive in nature and can be well approximated by the sum of 3 pairwise interactions; (2) $v > 1$: 3-body interactions are cooperative and their association is more favorable than 3 independent pairwise interactions; (3) $v < 1$: 3-body interactions are anticooperative.

Empirical Potential Incorporating Higher Order Interaction

We develop the empirical potential of higher order interactions based on a geometric model. Our model is analogous to the work of Hummer[51] in terms of the use of inclusion–exclusion formula. The difference is that we are interested in the geometric properties of the molecules, whereas Hummer studies the free energy of molecular hydration, and we use a much simpler inclusion–exclusion formula. Hummer's method was developed to study atomic solvation of molecules. Because any fractional exposure of constituent groups of proteins to water cannot be accurately captured by statistical potentials derived solely from known folded protein structures, our knowledge-based method is not expected to be accurate at atomistic details in addressing questions related to protein solvation. For ease of understanding, we describe Hummer's model first. A molecule M is modeled as a set of n fused hard balls $B = \{b_1, \ldots, b_n\}$, and it creates a cavity corresponding to the union of the excluded volumes vol $(\cup_{i=1}^{n} b_i)$. The free energy $E(M)$ of the hydration of the n balls forming this cavity is modeled as[51]

$$E(M) = w \cdot \mathrm{vol}(\bigcup_{i=1}^{n} b_i),$$

where w is a coefficient constant.

This is different from the simple volume summation of individual balls, because the latter neglects all volume overlaps. For molecules with chemical bonds, straightforward expansion of potential of mean force for $E(M)$ is inaccurate without explicit consideration of pair and higher order volume overlaps.[51] We can explicitly correct this by expanding the free energy associated with volume overlaps. Following the inclusion–exclusion principle, the corrected solvation energy $E(M)$ can be written as

$$\begin{aligned} E(M) = w \cdot \Big(& \sum_{i} \mathrm{vol}(b_i) - \sum_{i,j} \mathrm{vol}(b_i \cap b_j) \\ & + \sum_{i,j,k} \mathrm{vol}\,(b_i \cap b_j \cap b_k) \\ & - \sum_{i,j,k,l} \mathrm{vol}\,(b_i \cap b_j \cap b_k \cap b_l) + \ldots \Big) \\ & = w \cdot \sum_{\substack{\mathrm{vol}(\cap T) > 0 \\ T \subset B}} (-1)^{\dim(T)-1} \mathrm{vol}(\cap T), \end{aligned}$$

where vol $(b_i \cap b_j, \ldots$ represents volume overlap of various degree, $T \subset B$ is a subset of balls with nonzero volume overlap: vol$(\cap T) > 0$.

In principle, there can be very high degrees of overlap, and simulation studies showed that the volume overlap can go up to 7–8.[52,53] However, if a special expansion is used, the exact volume of the molecule can be obtained using terms with 4° at most in 3D space.[42] This requires that the 2-body, 3-body, and 4-body terms enter the formula if, and only if, the corresponding edge σ_{ij} connecting the 2 balls (1-simplex), triangles σ_{ijk} spanning the 3 balls (2-simplex), and tetrahedron σ_{ijk} cornered on the 4 balls (3- simplex) all exist in the dual simplicial complex κ or the α shape of the molecule.[41,42,54] Atoms represented by these atoms will all have volume overlaps. In this case, we have the exact expansion:

$$\begin{aligned} E(M) = w \cdot \Big(& \sum_{\sigma_i \in \kappa} \mathrm{vol}(b_i) - \sum_{\sigma_{ijk} \in \kappa} \mathrm{vol}(b_i \cap b_j) \\ & + \sum_{\sigma_{ijk} \in \kappa} \mathrm{vol}\,(b_i \cap b_j \cap b_k) - \sum_{\sigma_{ijkl} \in \kappa} \mathrm{vol}(b_i \cap b_j \cap b_k \cap b_l). \end{aligned}$$

We now generalize to account for different types of contact interactions. Allowing the n hard balls to be of different physicochemical types, we empirically introduce weight coefficients w_i, w_{ij}, w_{ijk}, and w_{ijkl} that are specific for each of the types of interaction:

$$\begin{aligned} E(M) = & \sum_{\sigma_i \in \kappa} w_i \cdot f(b_i) - \sum_{\sigma_{ij} \in} \\ & \kappa w_{ij} \cdot f(b_i \cap b_j) + \sum_{\sigma_{ijk} \in \kappa} w_{ijk} \cdot f(b_i \cap b_j \cap b_k) \\ & - \sum_{\sigma_{ijkl} \in \kappa} w_{ijkl} \cdot f(b_i \cap b_j \cap b_k \cap b_l), \end{aligned}$$

where $f(\cdot)$ is now a function that generalizes the volume measurement of the overlap. In the simplest form, $f(\cdot)$ can be a simple identity function $\mathbf{I}(\cdot)$ for the number count of various types of contact interactions.[4,5]

When considering only 1-body terms, we have $E(M) = \Sigma_{\mathrm{i}}\, \omega_{\mathrm{i}} \cdot f\,(b_i)$, where $\mathrm{f}(\mathrm{b_i})$ can be the solvent accessible surface area measurement, and w_i can be the solvation parameter for residue type i. This is the commonly used solvation model based on surface area.[55] When incorporating also the 2-body term, we have $E(M) = \Sigma_{\sigma_{\mathrm{i}}} w_i \cdot f(b_i) - \Sigma_{\sigma ij} \in \kappa\, w_{ij}) \cdot f(b_i \cap b_j)$. If $f(b_i \cap b_j)$ is chosen to be the number count of pairwise contacts, the model is similar to the grand canonical model developed in Sun et al.[56]

An alternative choice is to remove 1-body terms and use only 2-body terms but weight them with pairwise statistical contact potential derived from protein database.[4,5] In this study, we can empirically take w_{ij} as the statistical pairwise α contact potential, defined as $w_{ij} \equiv \ln P(i,j)$ in kT unit for residue i and j[7]: $E(M) = -\Sigma \sigma_{\mathrm{ij}} \in \kappa\, w_{ij} \cdot f(b_i \cap b_j)$. The contribution from solvation in this case is encoded implicitly, because buried residue will have more pairwise contacts. When adding the 3-body term, we follow the formulation based on the inclusion–exclusion principle and have

$$\begin{aligned} E(M) = & - \sum_{\sigma_{ij} \in \kappa} w_{ij} \cdot f(b_i \cap b_j) \\ & + \sum_{\sigma_{ijk} \in \kappa} w_{ijk} \cdot f(b_i \cap b_j \cap b_j \cap b_k). \end{aligned}$$

The 3-body α contact potential w_{ijk} in kT unit for residue i, j, and k is empirically taken as the nonadditive term: $w_{ijk} = -\ln \nu(i,j,k)$. Because the 2-body term is already included explicitly, it is inappropriate to take $\ln P(i,j,k)$ as w_{ijk}, since we would be overcounting. In all cases, we use simple number count of simplices [i.e., we use identity function $\mathbf{I}(\sigma)$ as $f(\cdot)$

$$\begin{aligned} E(M) = & - \sum_{\sigma_{ij} \in \kappa} w_{ij} \cdot \cap\, \mathbf{I}(\sigma_{\mathrm{ij}}) + \sum_{\sigma_{\mathrm{ijk}} \in \kappa} w_{ijk} \cdot \mathbf{I}(\sigma_{\mathrm{ijk}}) \\ & = - \sum_{\sigma_{ij} \in \kappa} \ln P(i,j) - \sum_{\sigma_{ijk} \in \kappa} \ln \nu(\mathrm{i,j,k}). \quad (5) \end{aligned}$$

Our focus here is the 3-body term $\Sigma_{\sigma_{\mathrm{ijk}}} \in \kappa \ln v(i,j,k)$. A further expansion can include 4-body body terms $-\Sigma_{\sigma_{\mathrm{ijkl}}} \in \kappa\; w_{\mathrm{ijkl}} \cdot \mathbf{I}(\sigma_{\mathrm{ijkl}})$, where $\mathrm{w_{ijkl}}$ is defined as $w_{\mathrm{ijkl}} \equiv \ln[P(i,j,k,l)/ P(i,j,k) \cdot P(i,j,l) \cdot P(i,k,l) \cdot P(j,k,l)]$. We neglect 4-body contacts in this study.

Database Selection

In this study, we use protein structures from the PDBSELECT database obtained from http://www.cmbi.kun.nl/

TABEL I. Propensity Values for 3-Body α Contact Interactions

Interaction Composition	95% Confidence interval	Average propensity	N^a
3 × same charged	0.05–0.45	0.21	124
2 × same charged + alkyl	0.16–0.47	0.29	1388
2 × same charged + aromatic	0.19–0.54	0.34	1094
2 × same charged + polar	0.27–0.82	0.52	962
2 × same charged + small	0.27–0.81	0.51	1447
2 × aromatic + ionizable	0.32–0.77	0.52	2538
Ionizable + alkyl + aromatic	0.35–0.72	0.52	7962
Ionizable + polar + alkyl	0.36–0.81	0.57	5790
2 × alkyl + ionizable	0.40–0.81	0.59	6218
2 × polar + aromatic	0.42–1.14	0.75	1829
Polar + alkyl + aromatic	0.42–0.92	0.65	5435
2 × aromatic + polar	0.43–1.06	0.72	1915
3 × aromatic	0.44–1.07	0.73	960
Ionizable + polar + aromatic	0.44–1.00	0.69	5109
Ionizable + small + aromatic	0.45–1.00	0.70	7062
Ionizable + small + alkyl	0.46–1.03	0.72	11,467
2 × small + aromatic	0.90–2.02	1.41	5365
2 × opposite charged + polar	1.00–1.78	1.36	2395
2 × small + polar	1.05–2.58	1.74	5635
2 × opposite charged + small	1.06–1.86	1.44	5739
2 × small + alkyl	1.15–2.39	1.72	11,901
2 × alkyl + aromatic	1.16–1.88	1.50	11,446
3 × opposite charged[b]	1.17–1.98	1.55	3319
2 × alkyl + small	1.30–2.52	1.86	17,420
3 × alkyl	1.59–2.30	1.93	14,183
3 × small	1.80–4.15	2.88	4639

Here a reduced alphabet set is used, and the average propensity values of the triplet types that can be represented by the same reduced triplet type are listed. Triplets formed by like-charged ionizable residues are listed as "same charged"; triplets formed by at least two opposite charged residues are listed as "opposite charged." Only 3-body interactions with both ends of bootstrap 95% confidence intervals for propensity < 1.1 or > 0.9 are listed.

[a] *N*: number of triplets corresponding to different group of residues, along with the corresponding confidence intervals.

[b] "3 × opposite charged": triplet types consist of 2 opposite-charged residues and another ionizable residue (e.g., EEK, DER).

swift/pdbsel.[57] PDBSELECT contains 1045 proteins selected from the Protein Data Bank (PDB). The sequence identity between any pair of proteins in this set of PDBSELECT data is smaller than 25%. We only include 3-body interactions from residues that are not contiguous in primary sequence. All 1540 possible 3-body α contacts occur in the PDBSELECT data set, with an average occurrence of 1200. Residues with the corresponding confidence intervals are also listed. The number count of triplets corresponding to different groups of residue (with the corresponding confidence intervals) are listed in Table I.

Confidence Intervals

Because the sample size of 1045 proteins in PDBSELECT is limited for assessing a large number of parameters for higher order interactions, statistical modeling with approximations may be prone to errors. It is therefore essential to assess reliability of estimated parameters for 3-body interactions. Here we use bootstrap technique to calculate confidence intervals of estimated 3-body propensity values and nonadditivity coefficients at 95% level from simulated data sets (Fig. 3).[58,59] We generate 1,000 × 50 nested bootstrapped samples for calculating confident interval, which takes about 45 min on a Pentium III (866 MH). Details of bootstrap calculation can be found in Li et al.,[7] Adamian et al.,[32] and Stitziel et al.[60]

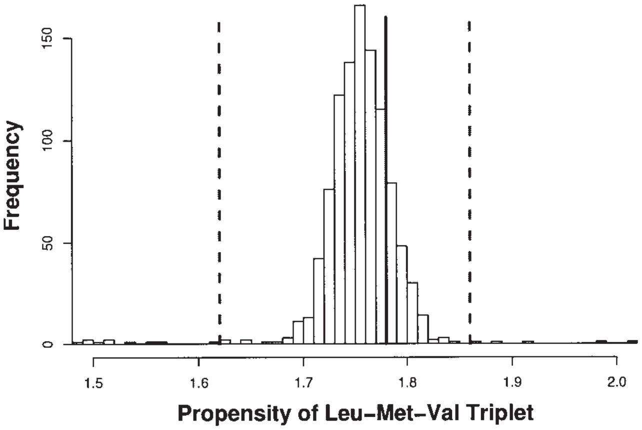

Fig. 3. Distribution of 1000 bootstrapped 3-body α contact propensity values of triplet LMV. Two dashed lines indicate the 95% confidence interval. The solid line represents the estimated 3-body α contact propensity value.

RESULTS

3-Body Contact Propensities

The distribution of the overall estimated 1540 3-body α contact propensities is shown in Figure 4(a), and their values are listed in the Supplementary Material. The majority of these triplet contacts have a propensity close to 1.0. Of the triplet contacts, 111 types (7.2%) have propensity > 2.0 and are strongly favorable. Eighty-one types of triplet contacts have propensity < 0.5 and are unfavorable. Among the favorable contacts, triplets containing C-C (Cys-Cys) contacts have the highest propensity (e.g., 14.6 for the CCG triplet) because of the tendency to form disulfide-bonds between Cys residues.

To facilitate interpretation and discussion, we use a reduced alphabet set to summarize the results of the propensities of the remaining 1520 = 1540 − 20 triplet contacts that do not contain C-C residues. This alphabet has 5 residue types: ionizable residues = {(R, K), (E,D)}; polar residues = {H, D, N, S, T, C,}; small residues = {G, A, P}; alkyl residues = {V, I, L, M}; and aromatic residues = {F, Y, W}. The use of a reduced alphabet is justified by a well-known fact that the natural alphabet of 20 amino acid residues has a large degree of built-in redundancy.[61–64] A standard test set of decoy conformations of proteins can also be detected using empirical pair potential based on reduced alphabets with 9 residue types.[7] Table I lists a summary of the 3-body contact propensities using this reduced alphabet. These triplet types are identified with

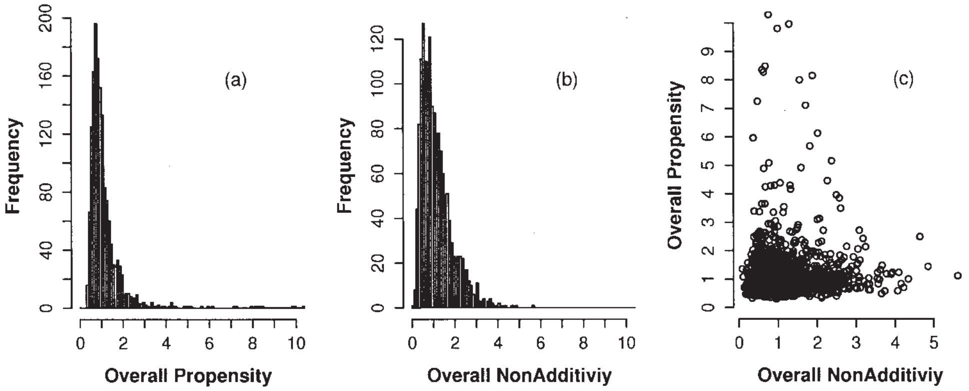

Fig. 4. Distribution of estimated overall propensity values and nonadditivity coefficients of 1540 types of 3-body interactions. (**a**) Propensity values for triplets from the whole protein regardless of geometric location. (**b**) Nonadditivity coefficient values for triplets from the whole protein regardless of geometric location. (**c**) Scatter-plot of overall propensity values versus overall nonadditivity values. No correlation is observed.

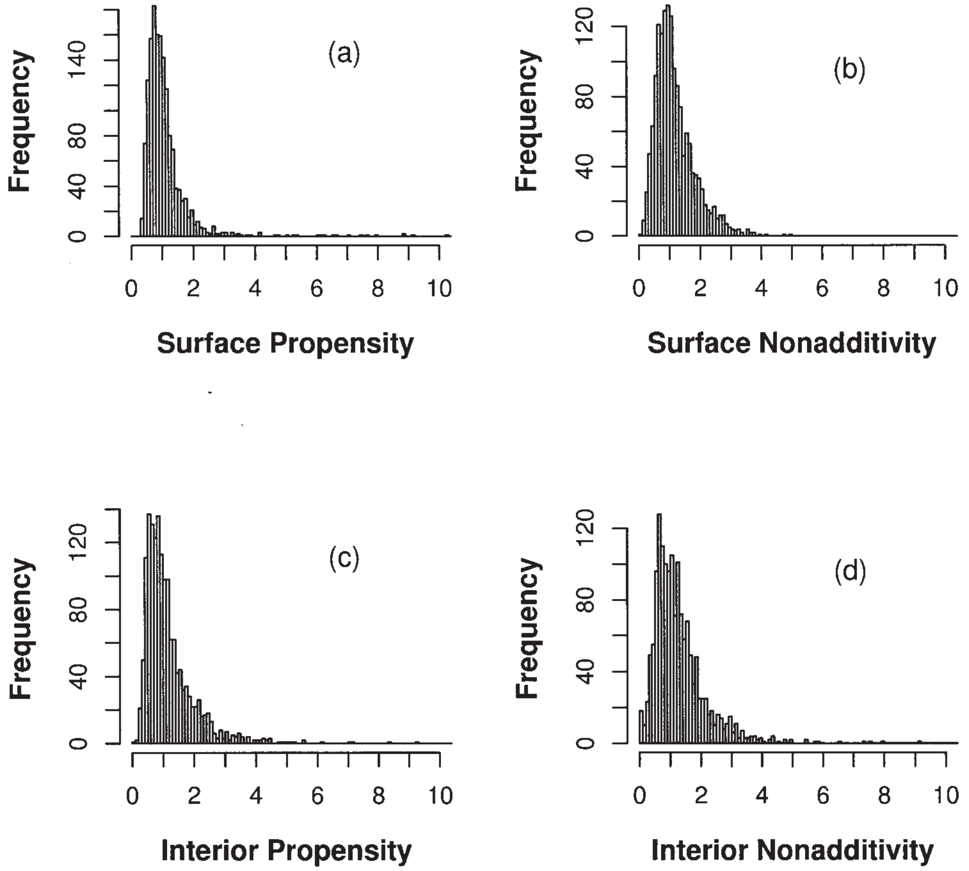

Fig. 5. Distribution of estimated propensity values and nonadditivity coefficients of 1540 types of 3-body interactions. (**a**) Propensity values for triplets located on the protein surface. (**b**) Propensity values for triplets located in the protein interior. (**c**) Nonadditivity coefficient values for triplets located on the protein surface. (**d**) Nonadditivity coefficient values for triplets located in the protein interior.

the additional criterion that both ends of the 95% confidence intervals are either ≥ 0.9 for favorable triplets, or ≤ 1.1 for unfavorable triplets.

Favorable 3-body contacts are often from triplets containing ionizable residues with opposite charges (e.g., DRS, DRY, DRH), triplets with hydrophobic residues, and triplets with small-size residues (e.g., AAA, GGG, AGG, AAM). Unfavorable 3-body α contacts are mostly composed of at least 2 ionizable residues of the same charge (e.g., KKK, EEE, KKM, and DDL), or mixture of hydrophobic residue and ionizable residue (e.g., EFV, RIF).

An important structural property of a residue is whether it is exposed to solvent. Residues located on the protein surface have large solvent-accessible surface (SAS) area, and buried residues have small SAS area. To understand whether solvent accessibility affects propensity of 3-body interactions, we classify 3-body contact interactions into the 2 groups of "interior buried contacts" and "exterior surface contacts." We consider a 3-body contact as "interior" if all 3 residues are buried; otherwise, a 3-body contact is considered to be located on the "surface." A residue is buried if its SAS area is less than 15% of its total surface area in the model of a tripeptide Gly-X-Gly.[65] Using this threshold, the total 3-body contacts in the data set are divided into interior and exterior contacts with about equal frequency. This is similar to the approach of Simons et al.[6] where pairwise surface and interior potentials are developed.

The distribution of the estimated 1540 3-body contact propensities in the protein interior and on the protein surface are shown in Figure 5(a) and Figure 5(b), respectively. Their values are listed in the Supplementary Material. For the triplet types of CHH and those containing residues AA(AA–{A,K,Q,D,E}), the propensities of them on the protein surface are about 1.0 unit higher than the same triplet types in the protein interior. Conversely, for the triplet types of KPP and those containing salt bridges (DDK, RRE, and REK), or abundant hydrogen bonds (GHS, NDS, HHH), or G-G (GG-{G,M,S}), the propensities of them in the protein interior are $\geq$2.0 units higher than the same triplet types on the protein surface.

Anticooperative and Cooperative Three-Body Interactions: Interior and Surface of Protein

It is important to distinguish between the concepts of propensity and cooperativity. Three residues may have strong propensity for interactions, but this can be a simple consequence of the high propensity of 3 independent pairwise interactions.[13] The nonadditivity coefficient υ defined in Eq. (3) provides a quantitative measure of the deviation of propensity from the combination of 3 independent pairwise interactions.

The estimated υ values for all 1540 triplet types are shown in Figure 4(b), and their values are listed in the Supplementary Material. The majority of triplet contacts have a nonadditivity coefficient υ close to 1.0. For interior residues, there are 482 triplet types whose 3-body propensity significantly deviates from linear additivity (i.e., the log values of 3-body contact propensities differ significantly from linear sum of log values of pairwise propensities. We use the criterion that both ends of the 95% confidence interval of the nonadditivity coefficients υ have to be > 0.8 for cooperative triplets, and < 1.2 for anticooperative triplets. For residues located in the protein interior, we found that 119 triplet types (7.7% of 1540) are strongly cooperative with $\upsilon > 2.0$, and 124 triplet types (8.1%) are strongly anticooperative with $\nu < 0.5$. For residues located on the protein surface, there are 619 triplet types whose 3-body propensity significantly deviates from linear additivity, of which 131 (8.5%) have $\upsilon > 2.0$, and 137 (8.9%) have $\nu < 0.5$.

For buried interior residues (Table II), the cooperative triplet types consist of those with at least 2 polar residues [e.g., ν(GNQ) = 2.56], at least 2 small residues [e.g., ν(AGG) = 3.92], or 2 ionizable residues with opposite charge [e.g., υ(RRE) = 2.83]. The anticooperative triplet types consist of those with at least 2 aromatic residues [e.g., υ(WWW) = 0.13], or 2 alkyl residues [e.g., ν(ILM) = 0.35]. Although the propensity values of 3-body contacts of alkyl residues are favorable in protein interior (Table II), the significant anticooperativity of average $\nu = 0.65$ of all alkyl triplets suggests that hydrophobic interactions are weaker than the sum of 3 independent pairwise hydrophobic interactions.

For protein surface residues (Table II), the cooperative triplet types include those consist of at least 2 small residues [e.g., ν(AGG) = 3.13; ν(GGI) = 4.17], or 2 alkyl residues [e.g., υ(IMV) = 3.20; υ(VVV) = 13.58, υ(IIF) = 3.17]. Although the propensity values of 3-body contacts between 2 alkyl and 1 aromatic residues are unfavorable on protein surface ($\bar{p}$ = 0.64; Table II), the significant positive cooperativity of average $\bar{\nu}$ = 2.05 suggests that hydrophobic interactions between 2 alkyl and 1 aromatic residues are stronger than the sum of 3 independent pairwise hydrophobic interactions. The anticooperative triplet types consist of at least 2 aromatic residues [e.g., υ(WWW) = 0.22], or 2 ionizable residues [e.g., υ(KKW) = 0.10; υ(ERS) = 0.67].

Several triplet types formed by 3 ionizable residues of the same charge have a large confidence interval [e.g., υ(DDD) = 1.32, with 95% confidence interval from 0.09 to 3.44]. Strong cooperativity is occasionally observed during the process of bootstrap resampling, depending on the subset of sample structures chosen. This may be due to the instances where 3 ionizable residues in some proteins are coordinated to bind to metal ions (e.g., Fe, Mn, Ca, Zn).

Altogether, there are 208 triplet types whose υ values both in interior and on surface can be reliably assessed. Among these, the nonadditive coefficients of 115 types of triplets do not change sign whether they are located on the surface or in the interior. Fifty-one types of triplets are cooperative both in the protein interior and on the protein surface (these are enclosed by "[ ]" in Table II). Most of these are triplets composed of 1 or more small residues [e.g., υ(GGV) = 2.40 on the surface and 2.55 in the interior, υ(GSS) = 2.58 on the surface and 2.54 in the interior). Sixty-four types of triplets are anticooperative both in the protein interior and on the protein surface [enclosed by "( )" in Table II]. Most of these are triplets composed of 1 or more aromatic residues, or 2 same-charged residues [e.g., υ(WWW) = 0.22 on surface, and 0.11 in interior, υ(DDQ) = 0.29 on the surface, and 0.27 in the interior].

For 93 types of triplets, the sign of nonadditive coefficient are different depending whether the triplet is on the surface on in the interior of the protein. Seventeen types of triplets are cooperative in the protein interior but are anticooperative on the protein surface [enclosed by "{ }" in Table II]. Most of these are triplets composed of ionizable or polar residues [e.g., υ(DEK) = 0.58 on the surface, but 3.92 in the interior; υ(DEH) = 0.35 on the surface but 5.47 in the interior]. Seventy-six types of triplets are cooperative on the protein surface but are anticooperative in the

TABLE II. Nonadditivity Coefficient v Values for 3-Body Interactions in the Protein Interior and on the Protein Surface

	Interior		Surface	
Triplet type	$v(i,j,k)$	$p(i,j,k)$	$v(i,j,k)$	$p(i,j,k)$
(2 × same charged + aromatic)	0.27 (0.05–0.63)	0.39	0.30(0.12–0.48)	0.51
(2 × same charged + alkyl)	0.32 (0.07–0.77)	0.61	0.39(0.16–0.70)	0.75
(2 × aromatic + ionizable)	0.41 (0.14–0.81)	0.44	0.35(0.17–0.60)	0.48
(3 × aromatic)	0.30 (0.15–0.53)	0.34	0.32(0.12–0.55)	0.37
(2 × aromatic + polar)	0.41 (0.16–0.80)	0.47	0.42(0.17–0.77)	0.53
2 × alkyl + ionizable	0.48 (0.23–0.81)	0.79	—[a]	1.07
Polar + alkyl + aromatic	0.47 (0.23–0.79)	0.51	—	0.65
‖2 × alkyl + aromatic‖	0.57 (0.40–0.79)	0.64	2.05(1.38–2.93)	0.64
‖Small + alkyl + aromatic‖	0.71 (0.45–1.04)	0.77	2.10(1.24–3.24)	1.12
‖3 × alkyl‖	0.65 (0.48–0.87)	1.30	2.03(1.42–2.82)	2.56
2 × polar + small	2.55 (0.97–4.17)	2.04	—	1.50
{2 × polar + ionizable}	2.99 (0.96–6.32)	—	0.50(0.24–0.84)	0.71
{2 × opposite charged + alkyl}	1.98 (1.00–3.31)	1.65	0.76(0.57–0.99)	1.14
{2 × opposite charged + polar}	2.77 (1.27–5.00)	1.71	0.78(0.49–1.15)	0.95
{2 × opposite charged + aromatic}	2.58 (1.36–4.34)	—	0.60(0.38–0.88)	0.78
{2 × opposite charged + small}	3.64 (1.63–6.64)	2.08	0.82(0.58–1.06)	1.30
{3 × opposite charged}	3.09 (1.69–5.36)	2.72	0.53(0.32–0.79)	1.13
3 × polar	2.91 (1.02–6.19)	—	—	—
[2 × small + polar]	1.95 (0.88–3.57)	2.39	2.74(1.19–5.03)	1.86
[2 × small + alkyl]	2.28 (1.06–4.02)	2.04	2.36(1.31–3.73)	1.73
[2 × small + ionizable]	3.23 (1.11–6.02)	2.07	2.63(1.09–4.94)	1.62
[Ionizable + polar + small]	3.79 (1.20–8.11)	1.78	1.84(0.94–3.25)	1.20
[3 × small]	3.25 (1.68–5.46)	3.62	3.10(1.47–5.12)	2.67

Only 3-body interactions in the protein interior with both ends of bootstrap 95% confidence intervals for nonadditivity $v < 1.2$ or > 0.8 are listed. The triplet types enclosed by parentheses "()" are anticooperative both in the protein interior and on the protein surface. The triplet types enclosed by brackets "[]" are cooperative both in the protein interior and on the protein surface. The triplet types enclosed by braces "{}" are cooperative in the protein interior but anticooperative on the protein surface. The triplet types enclosed by vertical bars "‖" are cooperative on the protein surface but anticooperative in the protein interior.

[a] Values are not listed because both ends of bootstrap 95% confidence intervals are not < 1.2 or > 0.8.

protein interior [enclosed by "‖ ‖" in Table II]. Most of these are triplets composed of 3 hydrophobic residues [e.g., v(IMV) = 3.20 on the surface but 0.75 in the interior: v(GMM) = 3.63 on the surface and 0.52 in the interior).

Anticooperative and Cooperative Three-Body Interactions on the Protein–Protein Interface

In the computation of the nonadditivity of 3-body contacts on the protein surface, we do not distinguish different surface regions. Therefore, the nonadditivity coefficients listed in Table II are averaged over all surface regions. However, protein-binding surfaces are different from the rest of the surfaces in both chemical characteristics and solvent accessibility: the average protein–protein interfaces are more hydrophobic than the rest of the surface[66–68] and are less hydrated after complexation. Can these differences result in different nonadditivity effects on the protein–protein interface from the rest of the protein surface? To answer this question, we use the DIMMER2 data set collected by Lu et al.,[69] which contains 768 nonredundant biological protein–protein interfaces. We select the interfacial 3-body contacts, of which at least 1 residue must come from a different polypeptide chain from that of the other 2 residues. Following the same procedure as described in the Model and Methods section, we computed the nonadditivity coefficients for the interfacial 3-body contacts. Due to the small size of available data, we can only obtain 15 clustered types of triplets with a good confidence interval, which contains 319 types of 3-body residual contacts on protein–protein interface (Table III). On the rest of protein surface, there are even fewer triplets with a good confidence interval. Thus, we are unable to carry out a comprehensive study on the difference between the 3-body nonadditivities on protein–protein interfaces and those on the rest of protein surfaces.

The cooperative triplets listed in Table III are formed by at least 2 hydrophilic residues [e.g., v(EKL) = 1.75 on the protein–protein interface but 0.32 on the overall surface; v(DHK) = 1.83 on the protein–protein interface but 0.45 on the overall surface]. Hydrophilic interactions are important for protein–protein interactions.[70,71] Results from continuum electrostatics calculation suggest that a hydrophilic bridge across the protein–protein interface often provides significant stability to protein binding,[71,72] although probably at a cost of destabilizing protein folding.[73] Experimentally, polar and charged residues are frequently hot spots, and mutations to alanine often result in the destabilization of the protein–protein complexes.[74] Both computational and experimental studies suggest that clustered hydrophilic residues on the protein–protein binding

TABLE III. Nonadditivity Coefficient *v* Values for 3-Body Interactions on the Protein–Protein Interface

Interaction composition	95% Confidence interval	Average nonadditivity
2 × aromatic + alkyl	0.16–0.74	0.39
2 × aromatic + small	0.19–0.82	0.42
Polar + small + aromatic	0.21–1.10	0.53
Polar + alkyl + aromatic	0.22–1.09	0.54
2 × alkyl + aromatic	0.24–0.80	0.46
Ionizable + polar + aromatic	0.26–1.07	0.56
Small + alkyl + aromatic	0.36–1.03	0.63
2 × same charged + alkyl	0.87–4.56	2.24
2 × same charged + small	0.91–5.17	2.44
2 × small + alkyl	0.93–2.85	1.69
2 × small + ionizable	0.95–3.10	1.78
Polar + small + alkyl	0.95–3.42	1.88
2 × opposite charged + alkyl	0.97–2.75	1.67
2 × opposite charged + small	1.01–2.76	1.75
Ionizable + small + alkyl	1.11–3.62	2.10

Only three-body interactions in protein interior with both ends of bootstrap 95% confidence intervals for nonadditivity v less than 1.2 or greater than 0.8 are listed.

interface may enhance the stability of the complex, where a desolvation penalty is partially compensated in interfaces through the formation of networks of ion pairs and hydrogen bonds. Our results indicate that these interactions are cooperative (i.e., they are more favorable than pairwise contacts alone). In many cases, these 3-body interactions are formed by at least 2 same-charged residues. The mechanistic origin of such cooperativity should be an interesting subject for further studies.

Atomic Origin of Nonadditivity

What are the physical bases of cooperativity and anticooperativity of 3-body interactions? Is cooperativity associated with specific types of atomic interactions? We examine the details of atomic contacts in residue triplet types with $v > 1$ and triplet types with $v < 1$. For a specific type of 3-body interactions (e.g., AGG), we collect all AGG triplets found in the dataset and examine the atomic pairwise contacts (Fig. 6). These are divided into the groups of hydrophobic interactions (C—C), regular hydrogen-bonding interactions (N—H· · ·O), weak hydrogen-bonding interactions (Cα—H· · ·O),[75–77] and polar-hydrophobic interactions (C with N, O, and S). We calculate the percentage of each types of atomic interaction in all atomic contacts found for a triplet type (e.g., AGG). For ease of interpretation, we use the same reduced alphabet of amino acid residues as before to describe the triplet types.

For triplets located in the protein interior, Figure 7(a) shows that the distributions of the fraction of hydrophobic atomic contacts for cooperative triplets and for anticooperative triplets are different: Interior anticooperative triplets have significantly more hydrophobic atomic contacts compared to cooperative triplets (p-value $= 2.2 \times 10^{-16}$ for 1-tailed t-test on the values of individual triplet types). Figure 7(c and e) shows that the majority of interior cooperative 3-body interactions have significantly more regular hydrogen bonds and weak hydrogen bonds compared to anticooperative triplets ($p = 2.2 \times 10^{-16}$ and $p = 6.6 \times 10^{-12}$ for 1-tailed t-test, respectively). For anticooperative 3-body interactions, the percentage of either of the 2 types of hydrogen bonds is low (an exception is 3-body interactions formed by 3 ionizable residues with the same charges, e.g., DDE). These observations suggest that in the protein interior, hydrogen bonding contributes significantly to 3-body cooperativity, while hydrophobic interactions contribute to anticooperativity.

For triplets located on the protein surface, Figure 7(b) shows that the distributions of fraction of hydrophobic atomic contacts for cooperative triplets and for anticooperative triplets are also different. In contrast to the pattern observed for interior residues, cooperative triplets have more hydrophobic atomic contacts compared to anticooperative triplets ($p = 1.4 \times 10^{-6}$ for 1-tailed t-test). For example, the frequency of hydrophobic atomic interactions counting for 80–100% of all atomic contacts is higher for cooperative triplets. Figure 8(d) shows that many anticooperative 3-body interactions have significantly more regular hydrogen bonds compared to cooperative triplets ($p = 5.8 \times 10^{-5}$). The distribution of atomic contacts involving weak hydrogen bond are similar for both exposed and buried triplets. These observations suggest that on protein surface, hydrophobic interaction contribute to cooperative 3-body interactions.

Another view of the atomic origin of cooperativity and anticooperativity of 3-body interactions is presented in Figure 8. Here the relative contributions of regular H-bond, weak H- bond, and hydrophobic interactions to different types of cooperative and anticooperative triplets are shown. We select 3-body interactions of the following 5 classes: those containing 2 or more of each of the 5 categories of aliphatic residues, aromatic residues, opposite-charged residues, polar residues, or small residues. These triplet types are chosen for ease of interpretation because they have simple composition and show clear patterns of cooperativity or anticooperativity.

We find that triplets with at least 2 aliphatic residues are anticooperative in the protein interior but cooperative on the protein surface, and triplets with at least 2 polar residues or 2 opposite-charged residues are cooperative in the protein interior but anticooperative on the protein surface. Triplets with at least 2 small residues are always cooperative, and triplets with at least 2 aromatic residues are always anticooperative. Figure 8 further shows that regular H-bonds contribute to cooperativity in the protein interior but mostly to anticooperativity on the protein surface. In contrast, hydrophobic interactions contribute to anticooperativity in the protein interior, but to cooperativity on the protein surface. Weak H-bonds, which can be attributed to small residues such as Gly, Ala, and Pro, seems to contribute consistently to cooperative interactions both in the protein interior and on the protein surface.

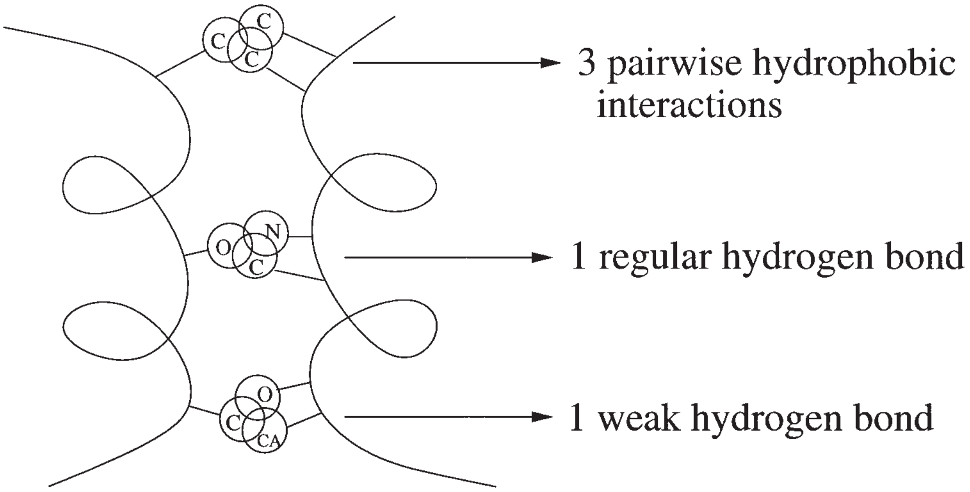

Fig. 6. Three-body atomic contacts can be divided into 4 types: hydrophobic, regular hydrogen bond, weak hydrogen bond, and polar–hydrophobic interactions (not shown). Here CA represents C_α, and C represents the rest of the carbon atoms.

Decoy Discrimination With Three-Body Potential

A useful test to assess the effectiveness of potential function is to evaluate its ability to distinguish native protein structures from incorrectly folded decoy structures. In this study, we carry out a limited test using 3 decoy sets, as chosen by Fain et al.[78] These decoy sets are the *lattice_ssift, 4-state_reduced,* and *lmds* sets from the Decoys "R' Us website (Table IV) (http://dd.stanford.edu).

The *4 state_reduced* decoy test set contains native and near-native conformations of 7 sequences, along with about 650 misfolded structures for each sequence. The positions of $C\alpha$ of these decoys were generated 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 score by a variety of scoring functions, and have low root-mean-deviation (RMSD) to the native structure (Table IV).[79] The *lattice_ssfit* set contains conformations for 8 small proteins generated by ab initio protein structure prediction methods.[80,81] 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. Secondary structures were fitted to these conformations using a 4-state model.[79] The 10,000 conformations were further scored using a combination of an all-atom scoring function,[82] a hydrophobic compactness function, and a 1-point per residue scoring function.[83] The 2000 best scoring conformations for each protein are selected as decoys for this data set. The local minima decoy set (*lmds*) contains decoys that were derived from the experimentally obtained secondary structures of 10 small proteins that belong to diverse structural classes. Each decoy is a local minimum of a "handmade" energy function.[84–87] Ten thousand initial conformations were generated for each protein by randomizing the torsion angles of the loop regions.[88] 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.

We use the following energy function for decoy discrimination. If we consider only pairwise contacts, the energy function $E(M)$ of a protein molecule M in kT unit is

$$E(M) = -\sum_{\sigma_{ij}\in \mathrm{int}\ \kappa} \ln P^{\mathrm{int}}(i,j) - \sum_{\sigma_{ij}\in \mathrm{bd}\ \kappa} \ln P^{\mathrm{bd}}(i,j),$$

where $\sigma_{ij} \in$ int κ represents pairwise α contacts or 1-simplices that are in protein interior, and $\sigma_{ij} \in$ bd κ represents 1-simplices that are on protein boundary surface. When we incorporate nonadditive effects of 3-body interactions following Eq. (5), the energy function $E(M)$ of a protein M in kT unit becomes

$$E(M) = -\sum_{\sigma_{ij}\in \mathrm{int}\ \kappa} \ln P^{\mathrm{int}}\,(i,j) - \sum_{\sigma_{ijk}\in \mathrm{int}\ \kappa} \ln \upsilon^{\mathrm{int}}(\mathrm{i,j,k})$$
$$- \sum_{\sigma_{\mathrm{ij}}\in \mathrm{bd}\ \kappa} \ln P^{\mathrm{bd}}(i,j) - \sum_{\sigma_{ijk}\in \mathrm{bd}\ \kappa} \ln \upsilon^{\mathrm{bd}}(i,j,k).$$

Table V summarizes the results of decoy discrimination. We found that the inclusion of 3-body interactions improves the performance of decoy discrimination in all decoy sets in the test data: There are less discrimination errors in general, and native structures have higher Z-value when evaluated with potential including 3-body contacts.

Incorporation of the cooperativity of 3-body contacts does not improve the performance of decoy discrimination for 4 sets of decoys in *lmds* set. 1b0n-B (Sini protein subunit) 1bba (pancreatic hormone) and 1fc2-C (fragment B of protein A) are small proteins or protein fragment with less than 45 residues. There are not enough 3-body contacts in these 3 proteins (Table V), thus, incorporation of 3-body cooperativity does not have any significant effects

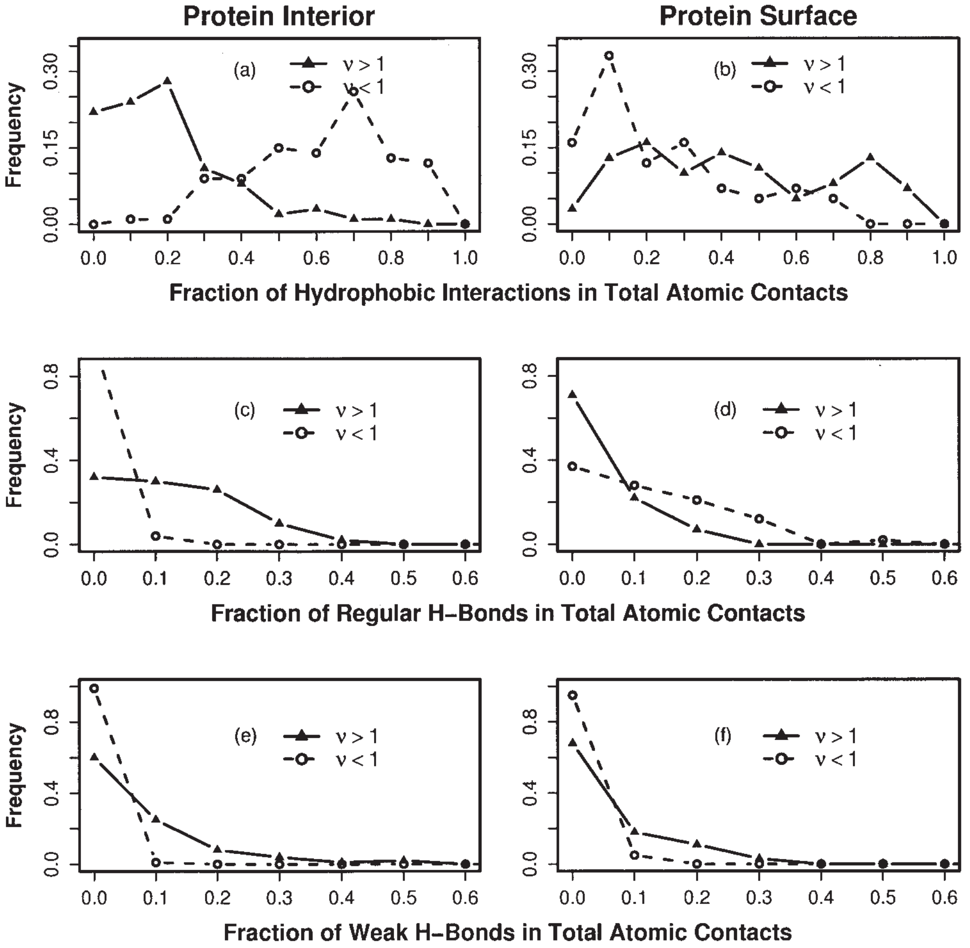

Fig. 7. The distributions of fractions of various types of atomic interactions in total atomic contacts for cooperative triplet types (solid line) and for anticooperative triplet types (dashed line): (**a**) hydrophobic contacts in the interior; (**b**) hydrophobic contacts on the surface; (**c**) regular hydrogen bond in the interior; (**d**) regular hydrogen bond on the surface; (**e**) weak hydrogen bond in the interior; and (**f**) weak hydrogen bond on the surface. In the protein interior, anticooperative triplets have significantly more hydrophobic atomic contacts compared to cooperative triplets. The majority of cooperative 3-body interactions in the protein interior have significantly more regular hydrogen bonds and weak hydrogen bonds compared to anticooperative triplets. On the protein surface, anticooperative triplets have fewer hydrophobic atomic contacts compared to cooperative triplets. Cooperative 3-body interactions on the protein surface have more hydrophobic interactions. Many anticooperative 3-body interactions on the protein surface have significantly more regular hydrogen bonds compared to cooperative triplets. The distribution of atomic con tacts involving weak hydrogen bond is similar for both exposed and buried triplets.

on the performance discrimination. For the set of decoys associated with 1dtk (Dendrotoxin K), incorporating 3-body cooperativity leads to deteriorated discrimination. Noticeably, Dendrotoxin K contains 3 disulfide bonds in its native structure, which are critical to stabilize the tertiary structure of this small protein. The improvement of discrimination for this small proteins requires first the incorporation of disulfide bonds interactions[7] before any higher order interactions is considered.

Overall, we have shown that incorporation of 3-body effects improves decoy discriminations for the *4-state-reduced* decoy set, the *Lattice-ssift* decoy set, as well as decoys associated with proteins in the *lmds* set that are not small or do not possessing unaccounted disulfide bond interactions. In this study, we only consider 3-body interactions defined by contact interactions. The decoy data set used for discrimination is limited. Our goal is to understand the dominant factors contributing to the effects of many-body interactions for the stability of native proteins. Discrimination of decoy in this case is used here only for the purpose of assessing the importance of 3-body interactions. Empirical potentials that are distance dependent have been shown to have better discrimination of native structures than contact potentials.[2,82,89] To further improve decoy discrimination, it will be useful to consider in addition longer range 3-body interactions by incorporating additional Delaunay triangles that are excluded in the α shape computation. This can be done in a hierarchical fashion by progressively adding triangles with increasing size measurement. Such an analysis can be the subject of future study and should be straightforward to implement with the approach developed here.

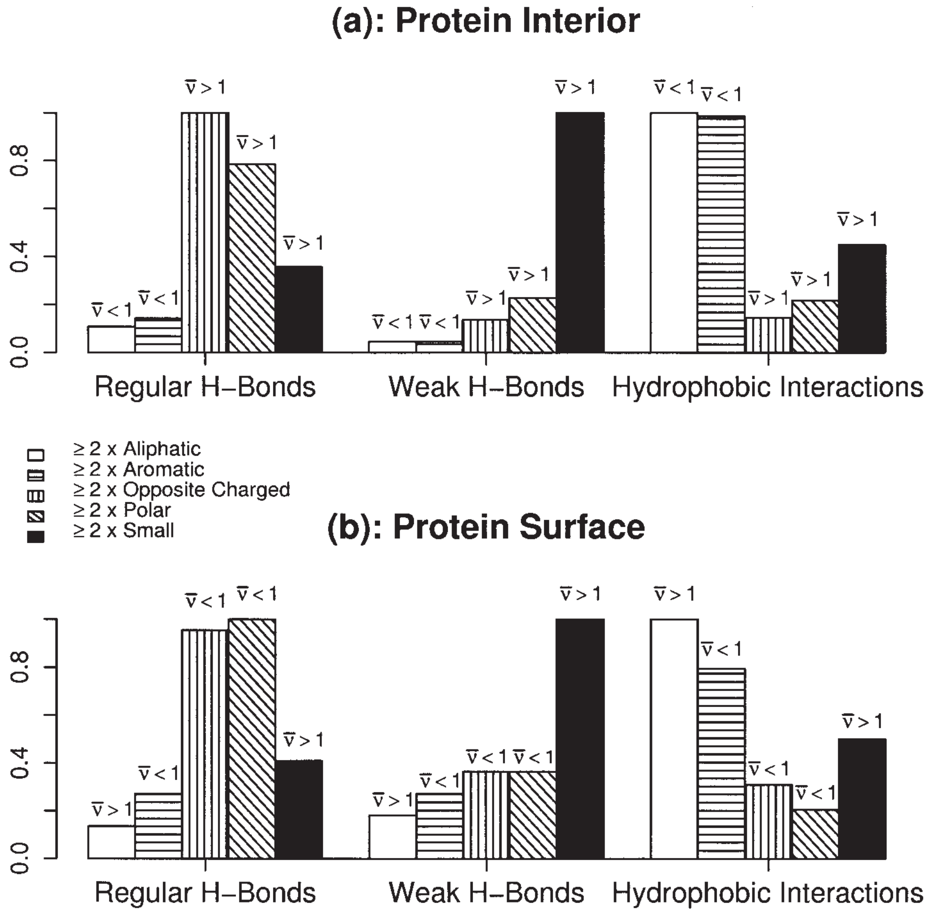

Fig. 8. Relative contributions of regular H-bonds, weak H-bonds, and hydrophobic interactions as fractions of total atomic contact interaction for 5 types of 3-body residue contacts. The contributions are normalized such that the type of interaction that contributes most has a *Y*- value of 1.0. (**a**) In the protein interior, cooperativity is positively correlated with the amount of regular and weak hydrogen bonds, and is negatively correlated with hydrophobic interactions. (**b**) On the protein surface, cooperativity is negatively correlated with the amount of regular hydrogen bonding, and positively correlated with the amount of hydrophobic interactions (except aromatic triplets).

Spatial Motifs of Three-Body Interactions

How can propensities and nonadditivity coefficients of 3-body interactions in proteins be used in practical applications? We show 2 possibilities. First, the incorporation of 3-body interactions helps in discriminating native protein conformations from decoy conformations. Second, the characterization of 3-body interactions can reveal spatial motifs that are characteristic of specific functional and structural properties of proteins.

The spatial conformations of contact triplets formed by a specific set of 3 residues often can be clustered into a small number of spatial motifs. This is analogous to the side-chain rotamers when only a single amino acid residue is considered. For example, triplet EEH is highly cooperative when located in protein interior ($v = 2.82$, propensity = 1.15). Its spatial conformations can be roughly divided into 5–6 spatial motifs (Fig. 9). One spatial motif is shown in details in the lower part of Figure 9. This motif is found in neighboring antiparallel helices and, in one case, neighboring parallel β-strands. Here the OD1 and OD2 atoms from a Glu residues and the NE2 atom from the His residues form network of hydrogen bonds. This motif is found in a diverse set of proteins, including Ni-Fe hydrogenase (1h2r), bacterial nonheme iron hydroxylase (1mmo), ribonucleotide reductase R2F protein (2r2f), ferritin (1bg7), rubrerythrin (1ryt), ribonucleoside-diphosphate reductase (1xik), and a designed 4-helix bundle modeled after metallo proteins (1ec5). Many of them are involved in redox reactions where electron transfer occurs. Six of them are found to bind to irons such as Fe and Ca (not drawn for clarity). This example indicates that 3-body interactions can suggest biologically interesting spatial motifs.

DISCUSSION

A variety of knowledge-based potentials has been developed for proteins. An important class of empirical potentials is based on statistical analysis of databases of protein structures.[2,4,82,90] For pairwise potential, the interaction between 2 residues is estimated from its relative frequency in a database when compared with a reference state or a

TABLE IV. Description of Proteins in the *4-state-reduced* Decoy Set, *lattice-ssfit* Decoy Set, and *lmds* Decoy Set (Number of Decoy Structures and cRMSD Ranges are Listed)

Decoy set	Protein	Description	N_{res}	N_{decoy}	cRMSD range
4-state_reduced	1ctf	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
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
	1nkl	Nk-lysin	78	2000	5.27–13.64
	1pgb	B1 immunoglobulin-binding domain of streptococcal protein G	56	2000	5.81–12.91
	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 (avg. 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 pheasant ovomucoid	56	347	4.38–13.38
	4pti	(See above)	58	343	4.94–13.18

null model. This approach has been successfully applied in fold recognition, in threading, and in many other studies.[1,2,6,7,82,91–94] Another class of empirical potential is derived by optimization. In this case, the parameters of the potentials are obtained by optimization following some optimal criterion [e.g., maximized Z-score difference between native conformation and a set of alternative (or decoy) conformations[95–103]. This approach is very effective in obtaining potentials that recognize native structures from alternative conformations or decoys.[103] However, the derivation of potential by optimization requires extensive training with a very large number of decoys to ensure that enough challenging decoys are included and a precise decision boundary can be obtained.[98,100] This often requires the use of millions of decoys when they are generated by gapless threading, or requires explicit generation of very challenging decoy conformations.[98–100] Because 3-body interactions raise the number of parameters from 210 to 1540, the requirement for decoy generation is likely to be more demanding than pair potential. We therefore choose the first approach and derive 3-body potential based on database statistical analysis. To ensure the reliability of estimated 3-body potential, we use nested bootstrap to obtain confidence intervals at 95% level, and report only parameters that can be estimated reliably.

The nonadditive nature of 3-body interactions is well recognized and was the subject of several studies based on the methods of molecular dynamics (MD) and Monte Carlo simulations.[13,104–106] In this study, we introduce a geometric model for analyzing higher order interactions in proteins. We define 3-body interactions through the aid of a simple geometric model, where 3 atoms are considered to be in 3-body interactions only when their 3-body volume overlap is nonzero, and their corresponding Voronoi vertex is contained in the protein. This definition is consistent with a long line of research where geometry-inspired models and parameters (e.g., SAS, molecular surface, and packing) facilitated fruitful insights into the nature of protein structure and protein stability.[44,107–109] The empirical energy model is similar to the volume-based hydrophobic force field developed in Hummer,[51] where higher degree expansion based on volume overlaps are incorporated. Unlike pioneering work using Delaunay triangulation of C_α atoms for assessing 4-body interactions,[19–23] a geometric model based on volume overlaps has direct physical interpretation and corresponds to precise atomic contact interactions. Without the aid of computed α shape, the method of Delaunay triangulation relies on size-related empirical cutoff values to prune long or skinny tetrahedra, which are judged empirically to be too large for physical contact. Such a heuristic method cannot decide exactly when volume overlaps occur at atomic contact level. In addition, 4-body potential involves a far larger number of parameters, and the estimated parameters are more likely to be subject to problems associated with small sample size.

Our model is simple and does not involve metric calculation of the volume overlap. Although volume calculation of 3-body and 4-body overlap can be found using an analytical

TABLE V. Discriminating Between Native Protein Conformation and Decoy Conformations

		$H_{2\text{-}body}^{in,ex}$		$H_{2\text{-}body,3\text{-}body}^{in,ex}$		
Decoy set	PDB	Native rank	Z	Native rank	Z	$N_{triplets}$
4-state_reduced	1ctf	1	2.59	1	3.38	85
	1r69	2	2.24	1	3.16	94
	1sn3	1	1.78	1	3.01	100
	2cro	2	2.32	1	3.04	99
	3icb	5	1.90	2	3.18	105
	4pti	28	1.34	2	3.15	84
lattice_ssift	1beo	12	3.12	1	5.14	127
	1ctf	1	4.58	1	4.83	85
	1dkt	7	2.87	1	4.12	99
	1nkl	1	3.34	1	4.23	93
	1pgb	10	2.69	1	3.37	60
	1trl-A	23	2.34	7	2.95	81
	4icb	3	3.24	1	4.82	106
lmds	1b0n-B	3	3.12	3	3.12	0
	1bba	312	−0.02	318	−0.02	10
	1ctf	1	2.87	1	3.43	85
	1dtk	56	1.56	89	1.20	82
	1fc2-C	478	−3.45	501	3.72	40
	1igd	7	2.50	5	2.52	67
	1shf-A	8	2.20	6	2.87	83
	2cro	1	4.56	1	5.69	99
	2ovo	5	3.45	1	4.56	85
	4pti	13	2.45	4	3.31	84

The ranking and Z-value of the native conformation by pairwise α contact potential $H_{2\text{-}body}^{in,ex}$ alone, and by combined pairwise and 3-body contact potential $H_{2=body,3=body}^{in,ex}$ are listed. Here Z-value is defined as $Z = (\bar{E} - E_{native})/\sigma$, where σ is the standard deviation of the evaluated energy value, $\bar{E}$ is the average energy value of all conformations, and E_{native} is the evaluated energy of the native conformation. $N_{triplets}$ in the table records the number of 3-body contact interactions in the native protein.

formula,[54] we only count the number of 3-body volume overlaps as recorded by the number of α triangle simplices. Our goal is to provide a description of the topological events of 3-body volume overlap and to investigate how 3-body interactions are different from what would be expected by considering 2-body interactions alone.

The main results of this work are the estimated propensities and additivity coefficients for 3-body interactions. These results are based on geometric patterns of 3-body volume overlaps emerging from the statistical analysis of protein structures. The only geometric model assumption is that each atom takes the shape of a ball. We find that both hydrophobic interactions and hydrogen bonding contribute differently to 3-body cooperativity, depending on whether they are located on the protein surface or in the interior. Although pairwise hydrophobic interactions are always favorable, they are strongly associated with anticooperativity when located in the protein interior, and regular hydrogen-bonding is strongly associated with cooperativity in the interior. When located on the protein surface, regular hydrogen bonding is associated with anticooperativity, and hydrophobic interactions with cooperativity. For both locations, weak hydrogen bondings are likely contributors to cooperativity.

In the protein interior, cooperative interactions are often associated with polar residues capable of forming regular hydrogen bonds, with small residues capable of forming weak hydrogen bonds, and with ionizable residues capable of forming salt bridges. On the protein surface, cooperative 3-body interactions are characterized by hydrophobic residues and small residues capable of forming weak hydrogen bonds. In the protein interior, anticooperative 3-body hydrophobic interactions are often due to long-chain alkyl residues. On the protein surface, anticooperative 3-body interactions are associated with aromatic residues and triplets containing 2 ionizable residues.

In some cases, these observed nonadditivity effects can be rationalized intuitively. Cooperative interactions in the protein interior are usually associated with salt-bridges, regular hydrogen bonds, and weak hydrogen bonds. This is consistent with experimental studies that suggest that the collective strength of hydrogen bonds is stronger when buried in the interior than when exposed to solvent.[25,26] A well-known fact about protein packing is that protein surface regions are packed more loosely than the protein interior.[110,111] Since residues on the surface and in the interior have different packing, their overall distances are different. It is possible that differences in packing are correlated to the nonadditive effects. For example, long-chain alkyl and aromatic residues are anticooperative in a tightly packed interior, but when located on a loosely packed protein surface, they become cooperative. How-

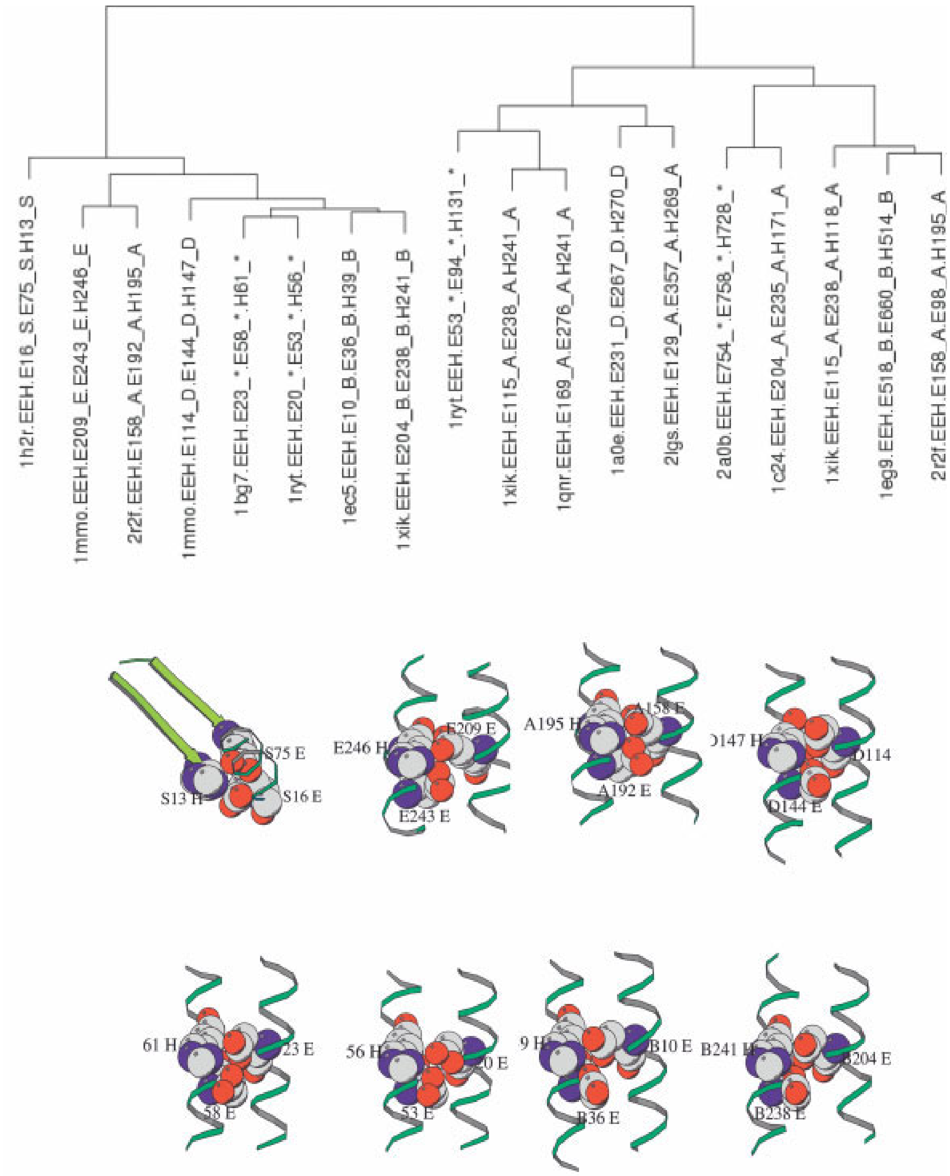
1h2r.EEH.E16_S.E75_S.H13_S
1mmo.EEH.E209_E.E243_E.H246_E
2r2f.EEH.E158_A.E192_A.H195_A
1mmo.EEH.E114_D.E144_D.H147_D
1bg7.EEH.E23_*.E58_*.H61_*
1ryt.EEH.E20_*.E53_*.H56_*
1ec5.EEH.E10_B.E36_B.H39_B
1xik.EEH.E204_B.E238_B.H241_B
1ryt.EEH.E53_*.E94_*.H131_*
1xik.EEH.E115_A.E238_A.H241_A
1qnr.EEH.E169_A.E276_A.H241_A
1a0e.EEH.E231_D.E267_D.H270_D
2lgs.EEH.E129_A.E357_A.H269_A
2a0b.EEH.E754_*.E758_*.H728_*
1c24.EEH.E204_A.E235_A.H171_A
1xik.EEH.E115_A.E238_A.H118_A
1eg9.EEH.E518_B.E660_B.H514_B
2r2f.EEH.E158_A.E98_A.H195_A
S75 E
S13 H
S16 E
E246 H
E209 E
E243 E
A195 H
A158 E
A192 E
D147 H
D114
D144 E
61 H
23 E
58 E
56 H
0 E
53 E
9 H
B10 E
B36 E
B241 H
B204 E
B238 E

Fig. 9. Spatial conformations associated with EEH triplet can be clustered into a small number of groups. Here they are clustered by RMSD using hierarchical clustering. The conformations of the 8 EEH triplets in the leftmost cluster are shown in detail. These come from proteins involved in electron transfer. Six of the 8 conformations have bound metal ions (e.g., Fe; not shown for clarity).

ever, the observed change of sign of 3-body cooperativity anticooperativity cannot be attributed to the difference in packing density alone. For example, the nonadditivity of certain ionizable triplets change from anticooperative on a loosely packed surface to cooperative in a tightly packed interior, but the nonadditivity of hydrophobic triplets changes sign in the opposite direction, from cooperative on the surface to anticooperative in the interior. The effects of packing on nonadditivity requires further study.

There are many additional 3-body contact interactions that contribute differently to cooperativity, depending on the environment where these contacts are located. In addition to difference in packing, the surface regions of soluble proteins in an aqueous environment have higher dielectric constant compared to the buried interior core. Polar residues on protein surface can form hydrogen bonding with solvent molecules. Hydrophobic residues, on the other hand, tend to aggregate and become dehydrated. In the protein interior, where the overall environment is hydrophobic[112,113] and the dielectric constant is low, polar residues and ionizable residues tend to form hydrogen bonds and salt bridge.[114] This is reflected by the favorable propensity and cooperativity of 3-body interactions of polar and ionizable residues.

The role of weak hydrogen bond in proteins has been the subject of several recent studies.[75–77,115] Our results indicate that weak hydrogen bonds seem to contribute consistently to cooperative interactions both in the protein interior and on the protein surface (Fig. 8). Triplets containing 2 or more small residues account for most weak hydrogen bonds. The contributions of weak H-bonds to cooperativity may be of geometric nature. For example, Pro residues frequently introduces bends in helices, which bring distant residues into closer contacts. They are found at N-capping,[116] β-turn,[117] and γ-turn[118] positions. Similarly, Gly and Ala are small residues. The lack of sizable side-chains allows wide access for contact interactions and tight packing, hence, the increased propensity of these residues for weak hydrogen bond interactions involving main-chain atoms.[119]

The difference in physicochemical properties for buried residues and surface residues is studied in details in a recent work of bioinformatics analysis of protein binding.[120] In this study, the matrix of pairwise binding contact potential derived from Z-score optimization is decomposed by principal component analysis. The principal components are regarded as idealized "canonical" residues, with specific physicochemical properties associated. The vectors of residue composition at different layers of protein structures have different coefficients when represented by the first 3 canonical residues. In addition, it was found that surface and buried residues provide different recognition signal for protein–protein interactions. It is clear from the present study that such differences for buried and surface residues are further reflected in different nonadditivity coefficients of 3-body interactions.

It is difficult to explain the origin of nonadditivity of some triplet types. There are many triplet types that are not listed in Table II. They have wide confidence interval of v across the neutral $v = 1$ value (i.e., from less than 0.8 to greater than 1.2). The signs of nonadditivity coefficients for these triplet types are uncertain, and it is unclear whether such triplet interaction is cooperative or anticooperative.

A number of factors can contribute to the nonadditivity of 3-body interactions. These include hydrophobic interactions, polar interactions, H-bondings, and size of side-chains. It is difficult to quantify the contribution from each of these factors. One possible scenario is that when a triplet has only 1 aromatics residue, all residues are relatively accessible, and the physicochemical nature of the atomic contacts (e.g., hydrophobic interactions, H-bondings, and polar interactions) are the main factors determining the nonadditivity. When a triplet contains 2 or 3 aromatic residues, it is possible that the geometric packing effect determined by side-chain size and shape dominates in determining the nature of nonadditivity. In addition, metal ions can have significant influence on the nonadditivity of nearby triplets (data not shown).

Our study is based on statistical analysis of protein structures in the PDB. Our approach follows that of previous studies, where much has been learned about the nature of stabilizing interactions of proteins.[1,4–8] However, database-derived empirical potential does not provide a mechanistic picture of protein folding. Although some patterns of 3-body interactions can be understood intuitively, as discussed above, patterns of many 3-body interactions, as shown in Table II, are not easily interpreted. In contrast, studies using MD or Monte Carlo simulation on simple systems can provide detailed pictures that illustrate the physical basis of nonadditivity.[11,13,104,105] It is also interesting to consider recent simulation results,[11,105] where it is shown that depending on how close 2 methane molecules are, and where the third molecule is located, interaction with a third methane molecule can be either cooperative or anticooperative. The results reported in this study should therefore be complemented by mechanistic insight gained from further simulation studies.

In simple 3-methane systems, nonadditivity largely disappears beyond the first hydration shell.[105] However, long-range higher order interactions play critical roles in hydrophobic collapse of long-chain polymers, and it is possible that such long-range cooperative effects are likely to be important for protein folding.[121] The geometric and topological characterizations derived from a simplicial complex of protein structures describe only short-range spatial interactions. They do not provide direct information about long-range spatial interactions between residues. Therefore, the 3-body potential developed here may be relevant for modeling contact interactions that are spatially local during the folding process.

Three-body contact potential developed here may also be useful for studying protein–peptide binding interactions. For tasks such as virtual screening of a large compound database, computational approaches based on MD or Monte Carlo simulations are prohibitively expensive. An attractive alternative approach is to approximate binding affinity, such as hydrophobic interaction by empirical force field. Incorporation of many-body or higher order interactions is important for such tasks.[51] An illustrative example of how many-body interactions helps in such important tasks is given in the study of energetic mapping of binding surface of the N-peptide coiled coil of gp14 protein of HIV-1 virus.[122] The 3-body potential developed in this study may be also applicable for modeling protein–peptide binding interactions. In our preliminary study, we showed that 3-body contact potential can be helpful for fold recognition. Its utility and effectiveness require further detailed studies.

In summary, we have introduced a geometric model for 3-body interactions. We show that with this model, both propensity and nonadditivity coefficients for many 3-body interactions can be reliably estimated despite the overwhelming presence of pairwise contacts. The different contributions of hydrophobic and polar interactions to cooperativity on the surface and in the interior point to a more detailed picture of physical interactions contributing to protein stability. Since higher order interactions in many cases strongly depend on the local environment, it is likely that these interactions play important roles in binding events, such as protein–protein and protein–DNA interactions. Unlike studies of protein stability, where 3-body interactions for solvent-exposed and buried interior residues often are of different signs and cancel each other, understanding of higher order interactions may bring us additional insights about binding events and other molecular recognition processes.

ACKNOWLEDGMENT

We thank Drs. Larisa Adamian, Bosco Ho, Bob Jernigan, Chao Tang, and Yaoqi Zhou for helpful discussions.

REFERENCES

1. Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term. J Mol Biol 1996;256:623–644.
2. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001; 44:223–232.
3. Betancourt MR, 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.
4. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
5. Sippl M. Boltzmann principle, knowledge-based mean fields and protein folding—an approach to the the computational determination of protein structures. J Comput Aided Mol Des 1993;7:473–501.
6. Simons KT, Ruczinski I, Kooperberg C, Fox B, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
7. Li X, Hu C, Liang J. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. Proteins 2003;53:792–805.
8. Bonneau R, Baker D. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct 2001;30: 173–189.
9. Dill KA. Additivity principles in biochemistry. J Biol Chem 1997;272:701–704.
10. Kaya H, Chan HS. Towards a consistent modeling of protein thermodynamic and kinetic cooperativity: how applicable is the transition state picture to folding and unfolding? J Mol Biol 2002;315:899–909.
11. Shimizu S, Chan HS. Anti-cooperativity and cooperativity in hydrophobic interactions: three-body free energy landscapes and comparison with implicit-solvent potential functions for proteins. Proteins 2002;48:15–30.
12. Rank JA, Baker D. A desolvation barrier to hydrophobic cluster formation may contribute to the rate-limiting step in protein folding. Protein Sci 1997;6:347–354.
13. Shimizu S, Chan HS. Anti-cooperativity in hydrophobic interactions: a simulation study of spatial dependence of three-body effects and beyond. J Chem Phys 2001;115:1414–1421.
14. Czaplewski C, Rodziewicz-Motowidlo S, Liwo A, Ripoll DR, Wawak RJ, Scheraga HA. Molecular simulation study of cooperativity in hydrophobic association. Protein Sci 2000;9:1235–1245.
15. Eastwood MP, Wolynes PG. Role of explicitly cooperative interactions in protein folding funnels: a simulation study. J Chem Phys 2001;114:4702–4716.
16. Rossi A, Micheletti C, Seno F, Maritan A. A self-consistent knowledge-based approach to protein design. Biophys J 2001;80: 480–490.
17. Godzik A, Skolnick J. Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc Natl Acad Sci USA 1992;89:12098–12102.
18. Godzik A, Kolinski A, Skolnick J. Topology fingerprint approach to the inverse protein folding problem. J Mol Biol 1992;227:227–238.
19. Krishnamoorthy B, Tropsha A. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. Bioinformatics 2003;19:1540–1548.
20. Carter CW, LeFebvre BC, Cammer SA, Tropsha A, Edgell MH. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol 2001;311:625–638.
21. Gan HH, Tropsha A, Schlick T. Lattice protein folding with two and four-body statistical potentials. Proteins 2001;43:161–174.
22. Zheng W, Cho SJ, Vaisman II, Tropsha A. A new approach to protein fold recognition based on delaunay tessellation of protein structure. Pac Symp Biocomput 1997;486–497.
23. Singh RK, Tropsha A, Vaisman II. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol 1996;3:213–221.
24. Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potential based on delaunay tessellation and their application in sequence–structure alignment. Protein Sci 1997;6: 1467–1481.
25. Pace CN. Polar group burial contributes more to protein stability than nonpolar group burial. Biochemistry 2001;40:310–313.
26. Myers JK, Pace CN. Hydrogen bonding stabilizes globular proteins. Biophys J 1996;71: 2033–2039.
27. Zhou Y, Linhananta A. Role of hydrophilic and hydrophobic contacts in folding of the second β-hairpin fragment of protein G: molecular dynamics simulation studies of an all-atom model. Proteins 2002;47:154–162.
28. Honda S, Kobayashi N, Munekata E. Thermodynamics of a β-hairpin structure: evidence for cooperative formation of folding nucleus. J Mol Biol 2000;295:269–278.
29. Kobayashi N, Honda S, Yoshii H, Munekata E. Role of side-chains in the cooperative β-hairpin folding of the short c-terminal fragment derived from streptococcal protein G. Biochemistry 2000;39:6564–6571.
30. Kaya H, Chan HS. Energetic components of cooperative protein folding. Phys Rev Lett 2000;85:4823–4826.
31. Liang J. Experimental and computational studies of determinants of membrane–protein folding. Curr Opin Chem Biol 2002; 6:878–884.
32. Adamian L, Jackup R Jr, Binkowski A, Liang J. Higher order interhelical spatial interactions in membrane proteins. J Mol Biol 2003;327:251–272.
33. Xu Y, Xu D, Crawford OH, Einstein Jr. R, Larimer F, Uberbacher E, Unseren MA, Zhang G. Protein threading by PROSPECT: a prediction experiment in CASP3. Protein Eng 1999;12:899–907.
34. Kihara D, Lu H, Kolinski A, Skolnick J. Touchstone: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci USA 2001;98: 10125–10130.
35. Koehl P, Levitt M. *De novo* protein design: I. In search of stability and specificity. J Mol Biol 1999;293:1161–1181.
36. Koehl P, Levitt M. *De novo* protein design: II. Plasticity of protein sequence. J Mol Biol 1999;293:1183–1193.
37. Hu C, Li X, Liang J. Developing optimal non-linear scoring function for protein design. Bioinformatics 2004;20:3080–3098.
38. Li X, Liang J. Computational design of combinatorial peptide library for modulating protein–protein interactions. Pac Symp Biocomput 2005;28–39.
39. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;*Suppl.* 3:171–176.
40. Edelsbrunner H, Mücke EP. Three-dimensional alpha shapes. ACM Trans Graphics 1994;13:43–72.
41. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. Analytical shape computing of macromolecules: I. Molecular area and volume through alpha-shape. Proteins 1998;33:1–17.
42. Edelsbrunner H. The union of balls and its dual shape. Discrete Comput Geom 1995;13:415–440.
43. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. Analytical shape computing of macromolecules: II. Identification and computation of inaccessible cavities inside proteins. Proteins 1998;33:18–29.
44. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
45. Edelsbrunner H, Shah N. Incremental topological flipping works for regular triangulations. Algorithmica 1996;15:223–241.
46. Facello MA. Implememntation of a randomized algorithm for delaunay and regular triangulations in three dimensions. Comput Aided Geomet Des 1995;12:349–370.
47. Tsai J, Taylor R, Chothia C, Gerstein M. The packing density in proteins: standard radii and volumes. J Mol Biol 1999;290:253–266.

48. Singh J, Thornton JM. Atlas of protein side-chain interactions. 2 Vols. Oxford, UK: IRL Press; 1992.
49. Adamian L, Liang J. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. J Mol Biol 2001;311:891–907.
50. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.
51. Hummer G. Hydrophobic force field as a molecular alternative to surface-area models. J Am Chem Soc 1999;121:6299–6305.
52. Petitjean M. On the analytical calculation of van der Waals surfaces and volumes: some numerical aspects. J Comput Chem 1994;15:507–523.
53. Kratky KW. Intersecting disks (and spheres) and statistical mechanics: I. Mathematical basis. J Stat Phys 1981;25:619–634.
54. Edelsbrunner H, Facello M, Fu P, Liang J. Measuring proteins and voids in proteins. In: Proc. 28th Ann. Hawaii Intl Conf Syst Sci. 1995;5:256–264.
55. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
56. Sun S, Thomas PD, Dill KA. A simple protein folding algorithm using a binary code and secondary structure constraints. Protein Eng 1995;8:769–778.
57. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
58. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
59. Davison AC, Hinkley DV. Bootstrap methods and their applications. Cambridge, UK: Cambridge University Press; 1997.
60. Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J. Structural location of disease-associated single-nucleotide polymorphisms. J Mol Biol 2003;327:1021–1030.
61. Riddle D, Santiago J, Grantcharova V, Baker D. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 1997;4:805–809.
62. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 1999;6:1033–1038.
63. Fan K, Wang W. What is the minimum number of letters required to fold a protein? J Mol Biol 2003;328:921–926.
64. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 2000;13:149–152.
65. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science 1985;229:834–838.
66. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;931:13–20.
67. Conte LL, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;2855:2177–2198.
68. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 1997;61:53–64.
69. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein–protein interactions. Biophys J 2003;843:1895–1901.
70. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein–protein interactions. Curr Opin Struct Biol 2000;102:153–159.
71. Xu D, Lin SL, Nussinov R. Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. J Mol Biol 1997;2651:68–84.
72. Sheinerman FB, Honig B. On the role of electrostatic interactions in the design of protein–protein interfaces. J Mol Biol 2002;3181: 161–177.
73. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;2685214:1144–1149.
74. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;2801:1–9.
75. Derewenda ZS, Lee L, Derewenda U. The occurrence of C—H. . .O hydrogen bonds in proteins. J Mol Biol 1995;252:248–262.
76. Bella J, Berman HM. Crystallographic evidence for C_α—H. . .O═C hydrogen bonds in a collagen triple helix. J Mol Biol 1996;264: 734–742.
77. Senes A, Ubarretxena-Belandia I, Engelman DM. The C_α—H. . .O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. Proc Natl Acad Sci USA 2001;98:9056–9061.
78. Fain B, Xia Y, Levitt M. Design of an optimal Chebyshev-expanded discrimination function for globular proteins. Protein Sci 2002;11:2010–2021.
79. Park BH, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
80. Samudrala R, Xia Y, Levitt M, Huang ES. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. Pac Symp Biocomput 1999; 505–516.
81. 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.
82. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
83. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
84. Levitt M, Lifson S. Rifinement of protein conformations using a macromolecular energy minimization procedure. J Mol Biol 1969;46:269–279.
85. Levitt M. Energy refinement of hen egg-white lysozyme. J Mol Biol 1974;82:392–420.
86. Levitt M. Molecular dynamics of native protein: I. Computer simulation of trajectories. J Mol Biol 1983;168:595–620.
87. 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.
88. Fletcher R. A new approach to variable metric algorithms. Comput J 1970;13:317–322.
89. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
90. Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 1976;9:945–950.
91. Wodak SJ, Rooman MJ. Generating and testing protein folds. Curr Opin Struct Biol 1993;3:247–259.
92. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
93. Lemer CMR, Rooman MJ, Wodak SJ. Protein–structure prediction by threading methods—evaluation of current techniques. Proteins 1995;23:337–355.
94. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
95. Goldstein R, Luthey-Schulten ZA, Wolynes PG. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc Natl Acad Sci USA 1992;89:9029–9033.
96. Maiorov VN, Crippen GM. Contact potential that pecognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
97. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci USA 1996;93:11628–1633.
98. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins 2000;40:71–85.
99. Vendruscolo M, Domanyi E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:11101–11108.
100. Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins 2000;38:134–148.
101. Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the Protein Data Bank. Proteins 2001;44:79–96.
102. Dima R. I. Banavar JR, Maritan A. Scoring functions in protein folding and design. Protein Sci 2000;9:812–819.
103. Micheletti C, Seno F, Banavar JR, Maritan A. Learning effective amino acid interactions through iterative stochastic techniques. Proteins 2001;42:422–431.
104. Czaplewski C, Rodziewicz-Motowidlo S, Dabal M, Liwo A, Ripoll DR, Scheraga HA. Molecular simulation study of cooperativity in

hydrophobic association: clusters of four hydrophobic particles. Biophys Chem 2003;105:339–359.

105. Ghosh T, García AE, S. G. Water-mediated three-particle interactions between hydrophobic solutes: size, pressure, and salt effects. J Phys Chem B 2003;107:612–617.

106. Shimizu S, Chan HS. Reply to “comment on ”anti-cooperativity in hydrophobic interac-tions: a simulation study of spatial dependence of three-body effects and beyond.’” J Chem Phys 2002;116: 2668–2669.

107. Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 1974;82:1–14.

108. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.

109. Banavar JR, Maritan A, Micheletti C, Trovato A. Geometry and physics of proteins. Proteins 2002;47:315–322.

110. Gerstein M, Richards FM. Protein geometry: distance, areas, and volumes. In: Rossmann M, Arnold E, editors. International tables for crystallography (Vol F). Dordrecht, The Netherlands; 2001. p 531–539.

111. Liang J, Dill KA. Are proteins well-packed? Biophys J 2001;81: 751–766.

112. Kauzmann W. Some factors in the interpretation of protein denaturation. Adv Protein Chem 1959;14:1–63.

113. Dill KA. Dominant forces in protein folding. Biochemistry 1990; 29:7133–7155.

114. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238:777–793.

115. Jiang L, Lai L. CH. . .O hydrogen bonds at protein–protein interfaces. J Biol Chem 2002;277:37792–37740.

116. Richardson JS, Richardson DC. Amino acid preferences for specific locations at the ends of alpha helices. Science 1988;240: 1648–1652.

117. Smith J, Pease LG. Reverse turns in peptides and proteins. CRC Crit Rev Biochem 1980;8:315–399.

118. Deber CM, Glibowicka M, Woolley GA. Conformations of proline residues in membrane environments. Biopolymers 1990;29:149–157.

119. Chakrabarti P, Chakrabarti S, C—H. . .O hydrogen bond involving proline residues in α. . . helices. J Mol Biol 1998;284:867–873.

120. Papoian GA, Ulander J, Wolynes PG. Role of water mediated interactions in protein–protein recognition landscapes. J Am Chem Soc 2003;125:9170–9178.

121. ten Wolde PR, Chandler D. Drying-induced hydrophobic polymer collapse. Proc Natl Acad Sci USA 2002;99:6539–6543.

122. Siebert X, Hummer G. Hydrophobicity maps of the N-peptide coiled coil of HIV-1 gp41. Biochemistry 2002;41:2956–2961.