2011

From Isotropic to Anisotropic Side Chain Representations: Comparison of Three Models for Residue Contact Estimation

Weitao Sun

Jing He
Old Dominion University

Original Publication Citation

Sun, W.T., & He, J. (2011). From isotropic to anisotropic side chain representations: Comparison of three models for residue contact estimation. *Plos One, 6*(4), 1-14. doi: 10.1371/journal.pone.0019238

OPEN ACCESS Freely available online

PLoS one

From Isotropic to Anisotropic Side Chain Representations: Comparison of Three Models for Residue Contact Estimation

Weitao Sun[1]*, Jing He[2]

1 Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, China, **2** Department of Computer Science, Old Dominion University, Norfolk, Virginia, United States of America

Abstract

The criterion to determine residue contact is a fundamental problem in deriving knowledge-based mean-force potential energy calculations for protein structures. A frequently used criterion is to require the side chain center-to-center distance or the C_α-to-C_α atom distance to be within a pre-determined cutoff distance. However, the spatially anisotropic nature of the side chain determines that it is challenging to identify the contact pairs. This study compares three side chain contact models: the Atom Distance criteria (ADC) model, the Isotropic Sphere Side chain (ISS) model and the Anisotropic Ellipsoid Side chain (AES) model using 424 high resolution protein structures in the Protein Data Bank. The results indicate that the ADC model is the most accurate and ISS is the worst. The AES model eliminates about 95% of the incorrectly counted contact-pairs in the ISS model. Algorithm analysis shows that AES model is the most computational intensive while ADC model has moderate computational cost. We derived a dataset of the mis-estimated contact pairs by AES model. The most misjudged pairs are Arg-Glu, Arg-Asp and Arg-Tyr. Such a dataset can be useful for developing the improved AES model by incorporating the pair-specific information for the cutoff distance.

Citation: Sun W, He J (2011) From Isotropic to Anisotropic Side Chain Representations: Comparison of Three Models for Residue Contact Estimation. PLoS ONE 6(4): e19238. doi:10.1371/journal.pone.0019238

Editor: Jörg Langowski, German Cancer Research Center, Germany

Received November 10, 2010; **Accepted** March 29, 2011; **Published** April 28, 2011

Funding: This work was supported by the Tsinghua University Initiative Scientific Research Program (20101081751), the Army High Performance Computing Center and NSF HRD-0420407. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sunwt@tsinghua.edu.cn

Introduction

The accurate identification of inter-residue contact is a crucial step in the understanding of protein structure. The residue contacts observed in crystal structures of globular proteins are generally considered the intrinsic inter-residue interactions. Based on this commonly accepted assumption, structures from the Protein Data Bank (PDB) [1] have been used to elucidate two-body residue contact and packing potentials since 1970s [2]. Miyazawa and Jernigan developed the theory of effective inter-residue energy from protein crystal structures [3,4] based on the Bethe Approximation [5,6,7,8] and quasi-chemical approximation [9,10,11,12,13]. Applying Boltzmann's law, Sippl proposed an approach to yield mean force potential of residue interactions as a function of distance [14]. In addition to the residue-distance-dependence studies [15,16], the effect of relative orientations on contact energy has been investigated [17,18]. In order to include the influence of multi-residue interactions and local environmental dependence, development of tri-residue [19,20,21], four-residue [22,23,24,25,26,27,28,29,30] and secondary structure-related energy [31] have been the focus of recent research.

For the sake of simplicity and computational efficiency, mean force potential are widely used in various applications, such as assessment of protein structures [32,33,34,35,36], folding recognition and threading [37,38,39,40,41,42,43], detection of native protein conformation [44,45,46,47,48], native topologies [49,50,51] and protein structure prediction [52,53,54,55]. Mean force potentials use reduced representations for side chains.

The contact models can be classified into two broad categories: the all atom model and the reduced representation model. In the all atom model, a pairs of residues are considered in contact if any two non-hydrogen side chain atoms (NHSA) from residues i,j are within a specified cutoff distance [56,57,58,59]. This model is expected to have accurate determination of the contact pairs [15,47,60,61]. The drawback is that it requires the knowledge of location of all the atoms on the side chains, and that is computationally expensive in structure prediction. Popular reduced representation of the side chains have been proposed through the use of C_α atom [62,63], C_β atoms, the centroid of amino acid and the centroid of side chain. Models with all atom main chain backbone and a single united atom for side chain have been proposed [53]. Another advanced model has been proposed with hydrogen bonds and flexible ellipsoidal side chains [64,65,66]. However, a more accurate description is required to capture atom-atom interactions in detail.

Two residue side chains are considered to be in contact if the side chain center or the C_α atom distance is less than a specified, pre-determined threshold distance [2,3,4,14,25]. The influence of a residue over surrounding medium can be effectively characterized at a limit distance [67,68,69,70]. A cutoff distance of 8.0 Å has been used in multi-body potentials [27], folding rate of proteins [71] and protein stability [72,73]. Other various

contacting distances have also been used for protein-folding studies. A commonly used side chain center distance threshold is 6.5 Å [3,4,18]. Spatial contact is considered to exist if C_α atom pair or C_β atom pair distance is less than 7.0 Å [74,75,76,77].

In order to avoid the drawbacks of arbitrary cutoff distance, Yang, et. al. proposed the parameter-free elastic network model (pfENM) to improve the estimation of B- factor [78]. Although the artificial cutoff distance are not as perfect as we expected, these convenient criteria still find their applications in many fields, especially in protein structure networks [79,80,81]. Cutoff distance is crucial to the contact degree distribution function describing the network behavior.

It is challenging to find a cutoff distance due to the variation in sizes, the preferred orientations and the anisotropy nature of the side chains. However, it can be more and more accurate as the mechanism of the contact is more and more understood. This study compares the following three models: the surface-to-surface model based on the side chain Atom Distance Criteria (ADC), the Isotropic Sphere Side chain (ISS) model and the Anisotropic Ellipsoid Side chain (AES) model using a dataset of 424 high resolution proteins from the PDB. We derive a dataset and illustrate the pairs that were wrongly estimated using the AES model for future study to improve the AES model.

Results and Discussion

It is known that side chains tend to have preferred orientations and exist as certain energetically favorable rotamers [82,83,84,85, 86,87,88,89]. This anisotropic nature of the side chains proposes challenges in the determination of the contact. In general, there are side chain overlaps (as defined by van der Waals radii) in experimentally determined NMR and crystallographic protein structures [90,91]. But the number of steric clashes is low. Side chain overlaps defined by covalent radii are even less. We investigated the overlaps in the high resolution PDB structures using three side chain models. The dataset was used in our previous work [49] and it includes 424 protein structures with single-chain, higher than 1.5 Å in resolution, less than 30% sequence identity structures from the PDB that are determined using X-ray crystallography technique [1]. Some PDBs with missed NHSAs are excluded.

Residue-contact distribution for the ADC model

In the ADC model, two amino acids overlap if any pair of atoms, one from each amino acid, is within the overlap cutoff distance. Two non-overlapping amino acids are in contact if any pair of atoms, one from each amino acid, is within the contact distance.

We calculated the overall residue contact degree $\lambda(n_r)$ and the overlap degree $\lambda_{ovl}(n_r)$ for 424 high-resolution PDBs based on ADC model. $\lambda(n_r)$ denotes the total number of contacts among all n_r residues in a protein. The overlap degree $\lambda_{ovl}(n_r)$ means the total number of side chain overlaps for a protein with n_r residues. Since it is not expected to find large number of residue overlaps in the test dataset, the lower the $\lambda_{ovl}(n_r)$, the more accurate contact model. The total number of residues falling within the contact distance of residue i is recorded as the contact degree $n_{cnt}(i)$.

$$n_{cnt}(i)=\sum_{j=1}^{n_r} A_{ij}, \tag{1}$$

$$A_{ij}=\begin{cases}1 & \text{if residue } i \text{ and } j \text{ are in contact}\\ 0 & \text{otherwise}\end{cases}. \tag{2}$$

Here n_r is the total number of residues in the protein. Residue i and its surrounding neighbors form a residue-contact cluster. This cluster is related to residue i and contains $n_{cls}(i)=n_{cnt}(i)+1$ residues, in which the residue immediately before and after i on the protein sequence are excluded.

An overall residue contact degree $\lambda(n_r)$ can be described by the size of the contact network.

$$\lambda(n_r)=\sum_{i=1}^{n_r}\sum_{j=i+1}^{n_r} A_{ij}. \tag{3}$$

$\lambda(n_r)$ provides an intuitive understanding of the compactness and overall connectivity. In the same way, overlap degree $\lambda_{ovl}(n_r)$ can be defined to describe the side chain overlaps.

$$\lambda_{ovl}(n_r)=\sum_{i=1}^{n_r}\sum_{j=i+1}^{n_r} B_{ij}, \tag{4}$$

$$B_{ij}=\begin{cases}1 & \text{if residue } i \text{ and } j \text{ overlap}\\ 0 & \text{otherwise}\end{cases}. \tag{5}$$

The ADC model reveals a linear relation between the contact degree $\lambda(n_r)$ and the protein length n_r (**Figure 1** (A)). Linear fitting formula $\lambda(n_r)=kn_r+b$ reveals a spontaneous collapse character of protein structures in different sizes. If all data points are matched simultaneously, the linear relationship can be described by $\lambda(n_r)=0.6n_r-8.79$ (the lower matching line in **Figure 1** (A)) with a confidence $R^2=0.73$. R^2 is the fitting coefficient of determination. The data fitting is facilitated by the Matlab fit function [92]. The relatively low fitting confidence is the result of the deviation of some 'escaping' data points.

An interesting observation is that the ADC model has the ability to classify protein structures. The 'escaping' data points, depicted in **Figure 1** (A), constitute another group and have a distinct linear slope. Thus, $\lambda(n_r)$ is divided into two separate groups with obviously different slopes. Here we use the linear slope k as a criterion. To separate the data into two groups, we used the line that fit all the data points as a reference, where $k_{all}=0.6$ and $b_{all}=-8.79$. The slope of the data point i is calculated as $k_i=\frac{\lambda_i(n_{ri})-b_{all}}{n_{ri}}$. If $k_i>k_{all}+0.2$, the data point i is placed in another group. When all the 'escaping' data points are fitted as a separate group, the linear regression is $\lambda(n_r)=1.3n_r-8.81$ with a confidence $R^2=0.93$ (the upper matching line in **Figure 1** (A)).

Detailed evaluation suggests that proteins with steeper increasing slopes are highly compact and can be considered dense-core proteins [93]. These well-packed structures can roughly be classified into three categories: (1) nearly-perfect globular proteins with short and flexible secondary structures; (2) proteins composed of a bundle of tightly packed alpha helixes; (3) proteins composed of curly β sheets.

Figure 1 (B) shows the distribution of the overlaps in the test data set. The fact that very few proteins have overlaps in the dataset suggests that ADC is an accurate side chain model that can be used to reflect the anisotropic effect of the residue side chains. In fact, the largest number of overlaps is 4 in one protein among the entire dataset. In addition, the sparse and random data distributions suggest that systematic misinterpretation of side chain overlaps is avoided in the ADC model.

The residue contact number depends on the r_{gap}, which is included in the definition of residue contact cutoff distance r_{cnt}^{ij}

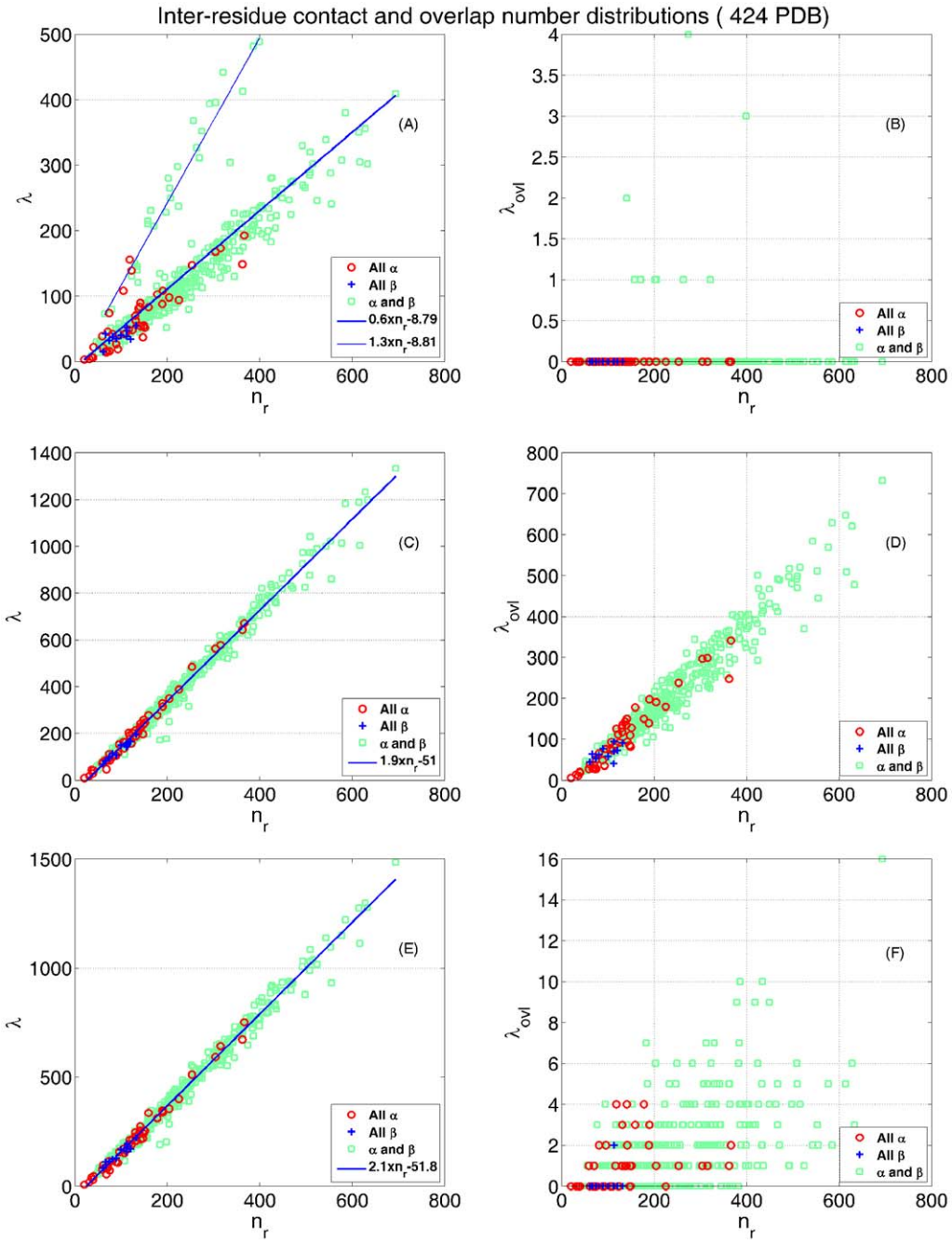

Figure 1. Residue contact distribution by ADC/ISS/AES contact model. $\lambda(n_r)$ and $\lambda_{ovl}(n_r)$ denotes the total number of contact and overlap among all residues in the protein. Surface gap distance $r_{gap}=0.0$ Å is used here. Data points for all-α helix, all-β sheet, α helix-β sheet proteins are plotted in different marker styles. (A)(B) ADC model; (C)(D) ISS model; (E)(F) AES model.
doi:10.1371/journal.pone.0019238.g001

(Equation (9)). Here r_{gap} is a gap distance between atom surfaces in ADC criteria (see Methods section). In order to investigate the influence of r_{gap} on overall residue-contact distributions, $\lambda(n_r)$ and $\lambda_{ovl}(n_r)$ are calculated for $r_{gap}=0.5$ Å, 1.0 Å, 1.5 Å and 2.0 Å, respectively. The contact cutoff distance r_{cnt}^{ij} increases as the r_{gap} increases. Residue pairs with larger distance, which were considered as separated-residues pair, are included as contact pairs. Not all these extra contact pairs reflect intrinsic residue interactions. An appropriate r_{gap} value is required to eliminate unexpected contacts. In the ADC fixed length model section, the derivation of the optimal r_{gap} is presented. Since the overlap cutoff distance is independent of the r_{gap}, the number of residue overlaps remain unchanged as the r_{gap} increases.

Residue-contact distribution for the ISS model

The ISS model uses a sphere to represent the side chain. This simple model can result in spurious side chain overlap as shown in **Figure 2**(A)–(B).

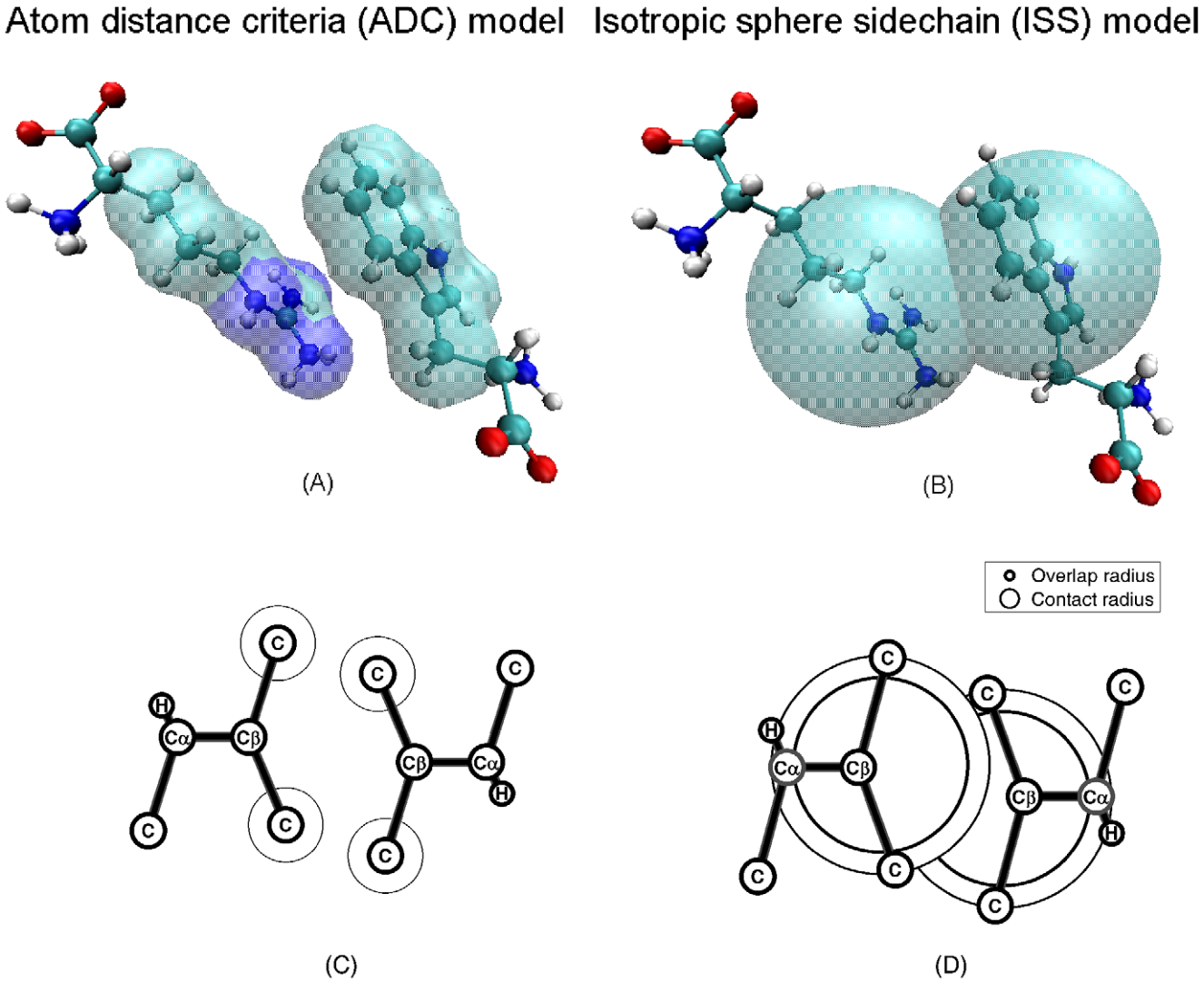

Figure 2. Residue side chain contact model. (A) All-atom sidechain model; (B) Simple isotropic sphere side chain model will cause spurious overlaps; (C) Effective overlap radius and contact radius of residue side chain atoms in the ADC contact model. (D) Effective overlap radius and contact radius in the ISS model.
doi:10.1371/journal.pone.0019238.g002

In this study, we used the geometry center $\mathbf{x}_0$ of heavy-atoms as the center of side chain.

$$\mathbf{x}_0 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i. \quad (6)$$

Here $\mathbf{x}_i$ is the coordinate of atom i. n is the number of heavy atoms.

In the ISS method, the error caused by neglecting side chain anisotropy can also be observed in the distributions of $\lambda_{ovl}(n_r)$, which increases linearly with respect to protein size (**Figure 1**(D)).

It is not surprising that the bulky side chains, such as Trp and Arg, lead to more significantly spurious overlaps. For example, phenyl rings (**Figure 2**(A)) prefer to form parallel or vertical orientations, and the spherical representation can overestimate the size of it. We also observed that the two lines fitted in the contact degree (Figure 1(A)) appear as one line (Figure 1(C)) in a linear regression of $\lambda(n_r) = 1.9n_r - 51$ with a confidence $R^2 = 0.99$. The sensitivity to anisotropy and ability to discriminate among different structure packings are lost in the ISS model. The $\lambda(n_r)$ behavior of the ISS model with a surface-gap distance $r_{gap} = 0.0$ Å appears to be equivalent to that in the ADC model with an atom surface-gap distance of $r_{gap} = 1.0$ Å or $r_{gap} = 1.5$ Å.

We also calculated the distributions of $\lambda(n_r)$ and $\lambda_{ovl}(n_r)$ with $r_{gap} =$ 0.5, 1.0, 1.5 and 2.0 Å respectively for the ISS model (data not shown). The increase in r_{gap} leads to a simultaneous increase in the residue-contact cutoff distance r_{cnt}^{ij}. As a consequence, more residue side chain pairs are considered to be within the contact range. However, $\lambda_{ovl}(n_r)$ distributions do not change with different r_{gap} for the reason that the overlap cutoff distance r_{ovl}^{ij} is independent of r_{gap} (see Equation (15)). For all types of r_{gap}, ISS model has significantly more overlaps than ADC model.

Residue-contact distribution for the AES model

In AES model, the residue side chain is represented as an ellipsoid with anisotropic radii in three principal dimensions. An ellipsoid collision-detection algorithm [94,95] was used to determine the side chain contact and overlap. With the anisotropic

radii in three principal dimensions, the AES model is much more accurate than the ISS model. Although the AES still has false positive determination of overlaps, the number of misjudgements is less than 5% of that in the ISS model.

Figure 1(E)–(F) show the distribution of $\lambda(n_r)$ and $\lambda_{ovl}(n_r)$ calculated by the AES model. The $\lambda_{ovl}-n_r$ ratio for the AES (**Figure 1**(F)) is much less than that in the ISS model (**Figure 1**(D)). With the ellipsoid overlap criterion, more than 95% of the false residue overlaps in the ISS model have been avoided, and most of the λ_{ovl} are less than 10. This number appear to be less than that in previous works [96].

Although the side chain anisotropy is taken into consideration to some extent, the 20 types of side chain conformations are still encompassed in a quadratic surface. The bulky side chain volume will lead to an underestimation of residue side chain distances. As a result, many closely packed residue pairs are mistakenly judged as overlaps. The anisotropic ellipsoidal radii help to improve accuracy in discriminating contact-residue pairs from separate-residue pairs in the AES model. However, the AES model still encounters difficulties in assessing the difference between overlap and contact. Part of residue contacts are taken as overlaps.

We analyzed contact determination algorithms with the three models (see method section). The ISS model is the least computationally intensive, followed by ADC and then AES. The accuracy rank is ADC, AES, ISS from high to low respectively. Algorithm accuracy/computing-cost ratio suggests that the ADC model is a cost effective model with the best accuracy. AES model is the most computational method among the three because the detection of ellipsoid collision algorithm is the most intensive step and needs to be improved in the future.

An analysis of the number of overlaps determined using the three models show that less than 50% of the total pairs of ADC contact are correctly predicted by ISS model. Whereas most of the 424 proteins have more than 95% ADC contacts shared by AES model. **Figure 3** shows the overlap number distribution for the 20×20 pairs of residues using the AES model. It appears that AES model is successful in determination of contact for most of the pairs. However, AES fails in most of the pairs involving Arg-Glu, Arg-Asp and Arg-Tyr. Arg-Glu pair is one of the most frequently seen false positive overlaps due to their large side chains. **Figure 4** (A) illustrated the ellipsoids calculated for an Glu-Arg pair. It appears that the overlapping volume is not much in this case. In another example of Asp-Arg (**Figure 4**(C)), the false positive overlap involves quite a lot of overlapping volume. **Figure 4**(B)(D) show the all-atom side chain positions of residue pair Glu260-Arg286 in 1IO0 (PDB ID) and residue pair Asp49-Arg51 in 1C7K (PDB ID). It is possible to develop an improved AES model that involves pair-specific and relative orientation dependent distance criteria for more accurate representation of the side chains.

Pair-specific contact cutoff distance

A popular contact cutoff is 5 Å between two NHSA atoms. We investigate if this threshold is a good estimation for all pairs of residues in this section. In theory, the cutoff distance in the ADC model should depend on the specific radii of atoms that are in contact and the atom surface gap distance r_{gap}. This is because the two residues can interact through different pairs of atoms. For example, the minimal distance of Val–Phe may occur either between atom pairs CG1–CE1 or CG2–CZ. The contact/overlap distances are distributions, rather than a single value (such as 5 Å).

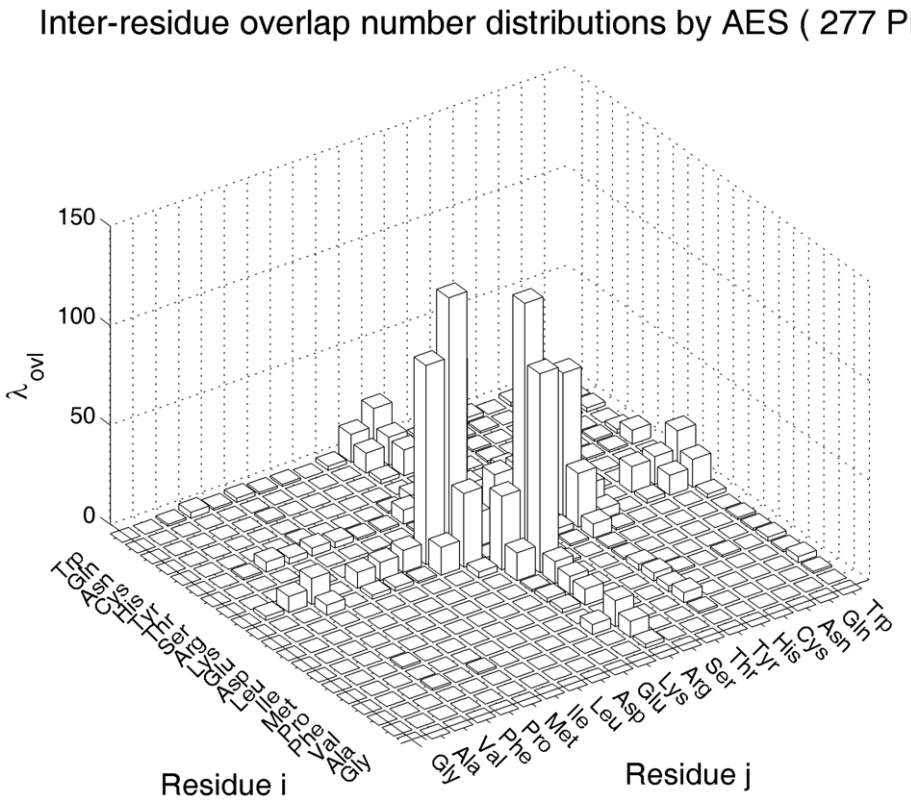

Figure 3. Overlap distributions of 20×20 residue pairs for AES model. AES-determined overlaps emerge in 277 out of 424 PDBs.
doi:10.1371/journal.pone.0019238.g003

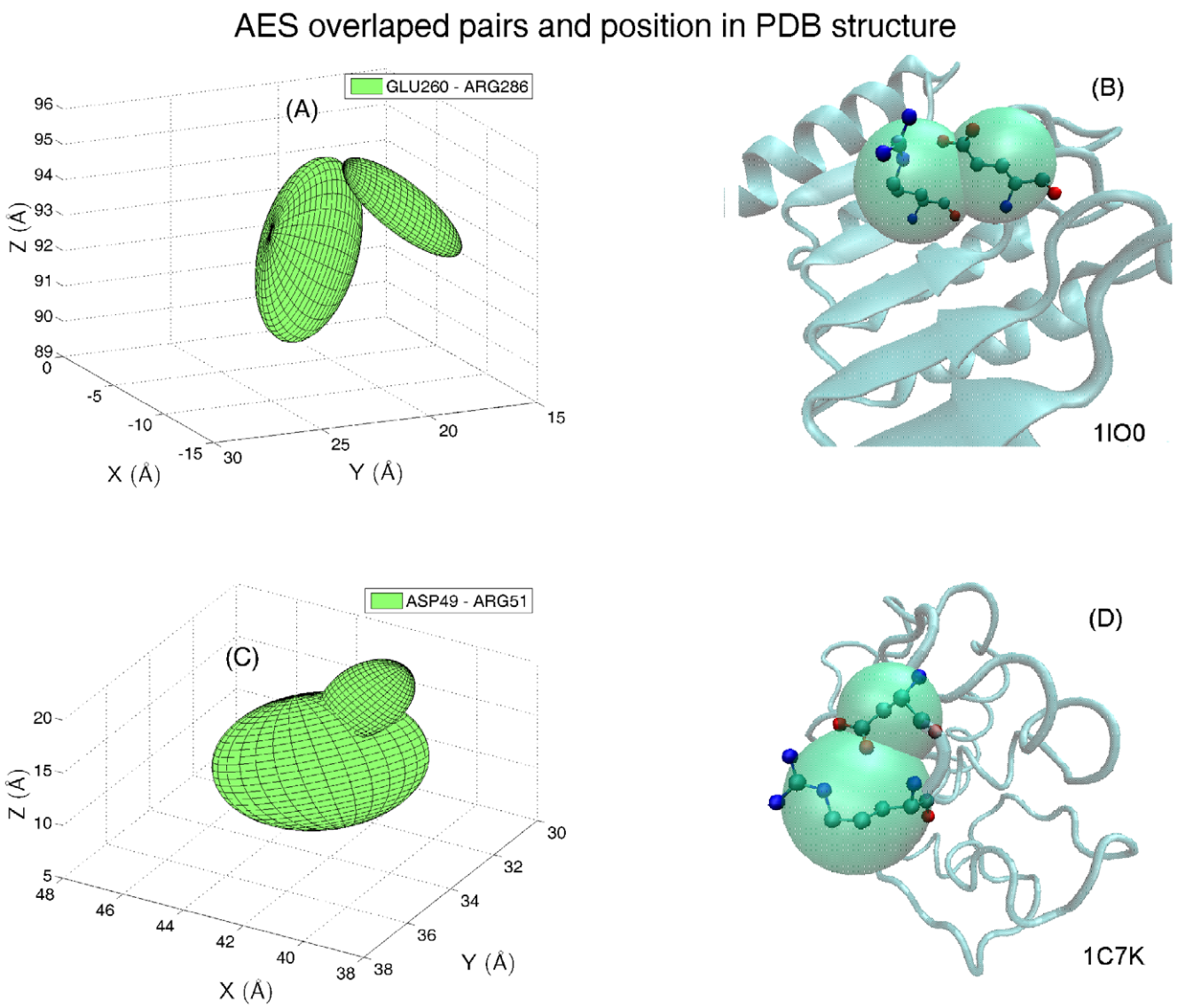

Figure 4. Examples of the ellipsoid overlap and side chain positions in PDB structures. (A) Glu-Arg overlap in 1IO0; (B) Glu260-Arg286 side chain postions in 1IO0 ; (C) Asp-Arg overlap in 1C7K; (D) Asp49-Arg51 side chain postions in 1C7K.
doi:10.1371/journal.pone.0019238.g004

The relation between r_{cnt}^{ij} distributions and the 5 Å model is an interesting topic. In addition, the method of how to estimate an appropriate r_{gap} value for the cutoff distance is discussed in this section.

The interactions between two residues have preferred distances and orientations, rather than a random packing. For any two residues, the most frequently occurring residue distance is considered its major contact distance. As the two residues depart from, or come close to, each other, the occurrence probability decreases gradually and the interaction energy becomes relatively unstable till the contact distance increases to the upper limit, the cutoff distance.

Figure 5 depicts the contact-distance distribution of Val-Phe pair. The histogram of packing distances between Val and Phe is a Gaussian-shaped function with a peak close to 3.9 Å. The peak position, i.e. the preferred contact distance, is almost independent of the cutoff distance. Only when the gap distance r_{gap} is beyond 3 Å will a second peak arise gradually. From the linearly increasing manner and peak-position shift, we ascertained that this peak ('non-local contact') is the result of the increase of cutoff distance, rather than an intrinsic interaction between Val and Phe. We further investigated all the pairs involving Val. The peak distributions of all residue pairs containing Val confirm the steady behavior of residue contact (**Figure 6**). As the gap distance r_{gap} increases, the peak positions corresponding to the preferred Val–XXX contact distance remain constant. While the positions of 'non-local contact' peak increase linearly with respect to r_{gap}.

The peak distributions allow us to set the cutoff distance for residue contacts. Between the two contact peaks, there is a low occurrence valley close to 5 Å (**Figure 5**). The valley position provides a rough estimation of the cutoff distance. We determined the cutoff distance for all the 210 pairs of residues using the position of the valleys (Table 1). The popular cutoff distance of $r_{cnt}^{ij} = 5$ Å appears to be effective in most of the pairs. However, some residue pairs such as Gly-XXX have complicated distributions with multiple valleys. In such cases, a larger cutoff distance will be chosen as the optimal value such that all the preferred contact distances (the stable peaks) can be included.

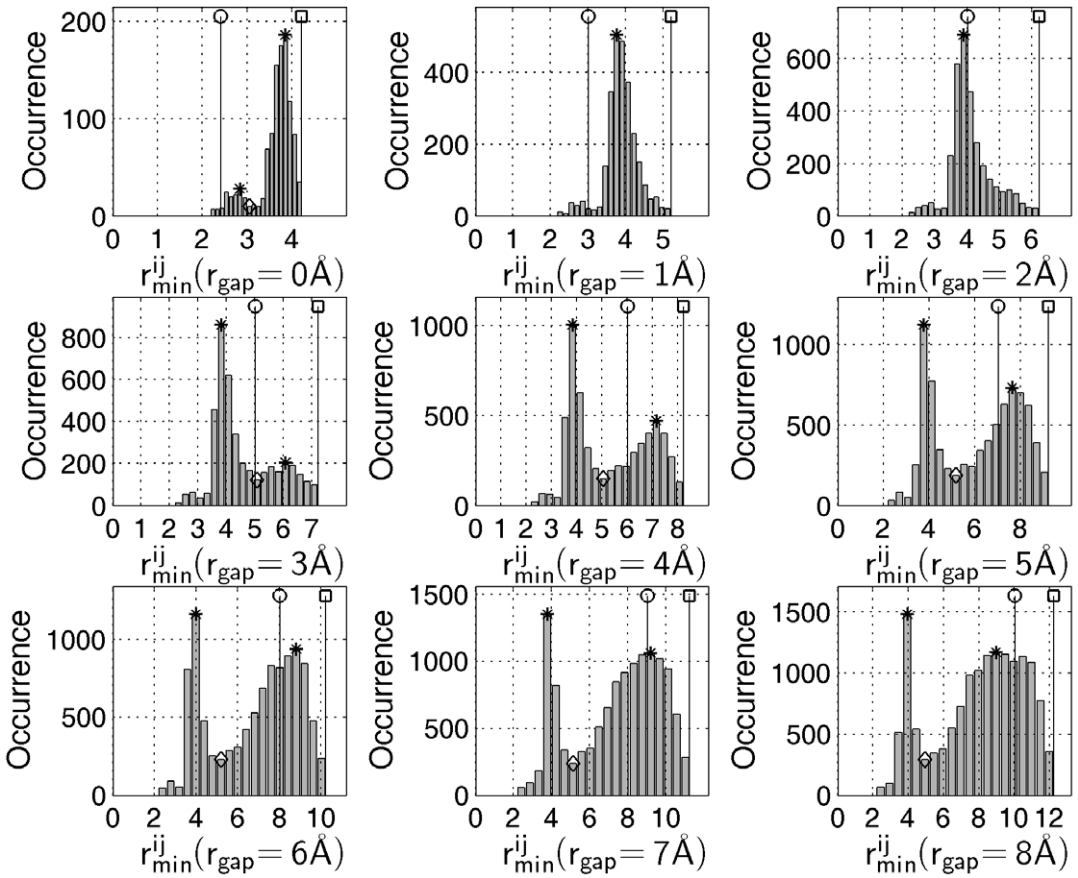

Figure 5. The contact distance distribution for residue pair Val–Phe at difference gap distances. The stem with circle indicates the minimal cutoff distance r^{ij}_{cnt}. The stem with a square indicates the maximal cutoff distance. The stars indicate high occurrence peaks. The diamonds indicate occurrence valley position. The r^{ij}_{min} is the minimal atom-to-atom distance between Val and Phe side chains. As the gap distance r^{ij}_{gap} increases from 0 to 8 Å, the cutoff distance r^{ij}_{cnt} increases simultaneously.
doi:10.1371/journal.pone.0019238.g005

A proper estimation of the gap distance r_{gap} is crucial in the ADC model. The optimal r_{gap} is expected to cover the most intrinsic contact distance between two residues, especially the major preferred contact distance. Meanwhile, the optimal r_{gap} must be small enough to exclude the fake "preferred contact distance" (the linearly increased peak position in **Figure 6**). The appropriate r_{gap} values are estimated statistically from the optimal cutoff distance r^{ij}_{cnt}. First, r^{ij}_{cnt} are selected from the valley positions under different gap-distance conditions (**Figure 5**). The valley position does not shift drastically when the gap distance r_{gap} increases. This stable behavior aids us to identify a statistically optimal r^{ij}_{cnt}. Then, from among all gap distances, there should be one critical value at which all contact-pair distances are less than the optimal r^{ij}_{cnt}. The critical value is the appropriate gap distance r_{gap}.

Take Val–Phe for example, **Figure 5** shows the valley position near 5 Å, i.e. the optimal cutoff distance $r^{ij}_{cnt} = 5$ Å. When the r_{gap} varies from 0 to 8 Å, the positions of occurrence bin extend from 4 to 12 Å along the horizontal axis. At the critical gap-distance value $r_{gap} \approx 1$ Å, all bin positions are lower than 5 Å. In such cases, the optimal gap distance is $r_{gap} \approx 1$ Å. Although small fluctuations do occur, we note that the optimal gap distance for all 210 residue pair types is around 1 Å.

Instead of the fixed value of 5 Å, the optimal cutoff distances in **Table 1** provide pair-specific cutoff distances. The ADC model uses the cutoff distance r^{ij}_{cnt}, which depends on the specific atom pairs between two side chains (see Equation (9). For the same type of residue pair, such as Val–Phe, the minimal side chain distance may occur between different pairs of atoms and hence the cutoff contact distances are usually different. The maximal and minimal cutoff-distance variations can be seen in **Figure 5**. No matter what cutoff distances are used, either the popular 5-Å criterion or the optimal ones in **Table 1**, single-value cutoff distances can hardly deal with various atom-to-atom contact cases. The 5 Å may be a good choice for two atoms surrounded by hydrogen atoms. However, it may be too large for two atoms that have no hydrogen atoms attached to them. A fixed, large cutoff-distance value is more convenient for most residue side chain contact, but over-estimation can happen when the cutoff distance is adopted for heavy atom pairs in more compactly packed side chains.

Conclusion

The influence of residue side chain anisotropy has been studied for three side chain contact models. The atom distance criteria (ADC) contact model shows high accuracy in the determination of residue contact and overlaps. Protein structures can be classified as

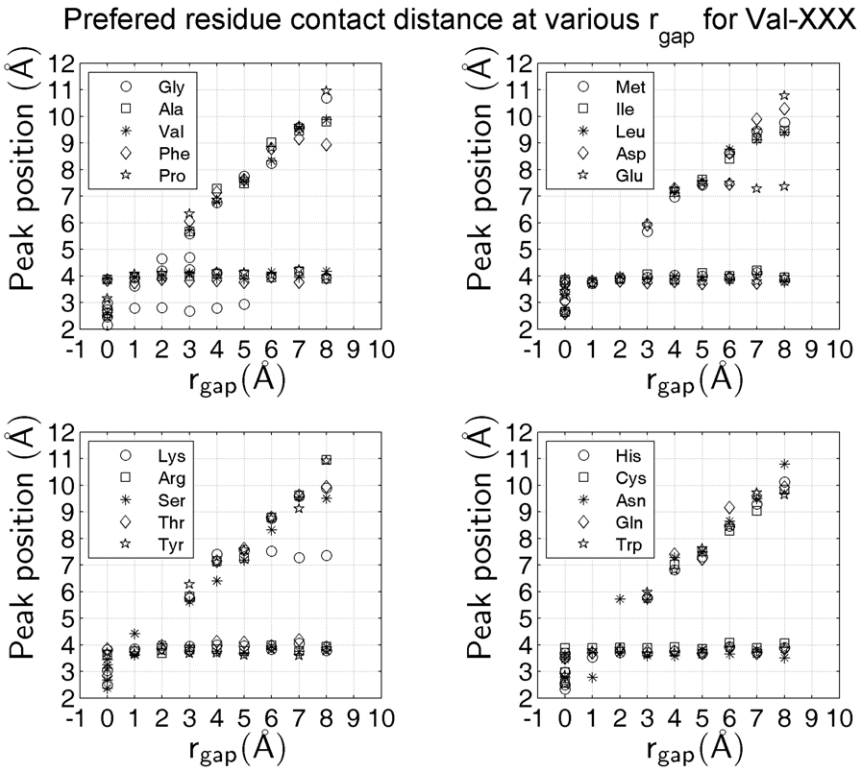

Figure 6. The preferred contact distance for Val-involved residue pairs at different gap distances. The positions of occurrence peaks in Val–XXX contact-distance distribution show two different behaviors. One is the stable, high occurrence-peak positions (the preferred contact distances) close to 4 Å, which are independent of the gap distance r_{gap}. The other is the linearly increasing peak positions (contact distances caused by increase in r_{gap}).
doi:10.1371/journal.pone.0019238.g006

closely packed and loosely packed groups with the help of two different linear fit of $\lambda(n_r)$ by ADC method. The isotropic sphere side chain (ISS) model has systematically misjudgements in determining of both residue contact and overlaps. The residue surface distances are underestimated and more side chain overlaps emerged. With the different radii in three principle directions, anisotropic ellipsoid side chain (AES) model is more accurate than ISS in determining residue contact. The number of misjudgement is less than 5% of that in ISS method. However, AES need much more computations than ISS model. Based on the algorithm accuracy and complexity analysis, ADC model is recommended as the best all-atom side chain contact determination method. And AES is the most promising coarse-grained method.

Methods

Atom distance criteria (ADC) for residue side chain contact and overlap

Two atoms can be considered in 'contact' when they are in close interaction. Van der Waals interaction, a commonly employed close interaction, decreases rapidly with the distance between atoms. Residue contact can be defined when any two non-hydrogen side chain atoms (NHSA) from two residues are in the range of van der Waals interaction. This interaction-based contact definition are usually converted to distance-based contact definition by a cutoff distance of van der Waals interaction [56,57,58]. A popular cutoff distance between atoms from different residues is 5 Å [97]. We discuss the atom distance criteria in details in this section.

X-ray crystallography can barely resolve hydrogen atoms in most protein crystals. As a consequence, hydrogen atoms are absent in most PDB files. Thus the influence of hydrogen atoms that are attached at the NHSA has to be approximately included in determining residue-contact relations. we define the contact radius R_{cnt} as in the following (**Figure 2** (C)).

$$R_{cnt} = R_{vdw} + \delta_H d_H. \tag{7}$$

Where R_{vdw} is the van der Waal's radius of the side chain atom; d_H denotes the additional volume thickness if this atom has an attached hydrogen atom; $d_H = 0.4R^H_{vdw}$ was used in the current study; R^H_{vdw} denotes the van der Waals radius of hydrogen atom; and δ_H is a constant value , which is defined as:

$$\delta_H = \begin{cases} 0 & \text{Has attached hydrogen atom} \\ 1 & \text{No attached hydrogen atom} \end{cases}. \tag{8}$$

Atom interactions are confined to a limited range, such as the contact radius of an atom. If the distance r_{ij} between atom i and j satisfies the criterion $r_{ij} < r^{ij}_{cnt}$, the atoms are considered to be in contact. Other than a predetermined fixed value, the atom-contact cutoff distance r^{ij}_{cnt} is calculated based on side chain atom-surface distance, which reflects the anisotropy in side chain orientations.

$$r^{ij}_{cnt} = R^i_{cnt} + R^j_{cnt} + r_{gap}. \tag{9}$$

R^i_{cnt}, R^j_{cnt} are contact radii of atoms i and j. The r_{gap} is the gap distance representing the decay of atomic interaction. With the current definition, the cutoff distance r^{ij}_{cnt} can be different values

Table 1. The optimal atom–atom cutoff distance for all types of residue side chain contact pairs (Unit: Angstrom).

ID	Gly	Ala	Val	Phe	Pro	Met	Ile	Leu	Asp	Glu	Lys	Arg	Ser	Thr	Tyr	His	Cys	Asn	Gln	Trp
Gly	3.9	5.8	4.4	3.7	5.4	4.8	6.2	4.2	4.6	5.5	4.6	5.0	4.1	5.8	5.5	3.9	5.6	4.1	4.5	3.5
Ala	5.8	5.1	5.0	5.1	4.7	6.0	5.1	4.9	5.0	4.9	5.0	4.5	4.9	5.0	4.8	4.8	5.0	4.8	5.2	5.0
Val	4.4	5.0	5.2	5.1	5.2	5.0	5.1	5.2	4.7	5.0	5.0	4.8	5.0	5.1	5.2	5.1	5.1	4.9	5.0	4.9
Phe	3.7	5.1	5.1	4.9	4.8	5.1	5.2	5.2	4.9	4.6	4.8	5.1	4.9	4.9	4.9	4.6	4.7	5.0	4.9	5.4
Pro	5.4	4.7	5.2	4.8	4.8	5.0	5.0	5.1	4.6	5.1	4.8	4.8	4.7	5.1	4.8	5.2	4.9	4.5	4.7	5.1
Met	4.8	6.0	5.0	5.1	5.0	5.7	5.4	5.2	4.5	4.5	5.5	4.7	4.9	4.9	5.4	5.7	5.0	4.7	5.1	4.9
Ile	6.2	5.1	5.1	5.2	5.0	5.4	5.2	5.5	4.7	4.8	5.0	5.1	4.8	5.1	5.2	4.7	5.4	4.9	5.3	4.9
Leu	4.2	4.9	5.2	5.2	5.1	5.2	5.5	5.5	4.5	5.0	5.0	5.0	4.8	4.9	5.2	5.1	4.9	5.0	4.8	5.4
Asp	4.6	5.0	4.7	4.9	4.6	4.5	4.7	4.5	4.9	4.5	4.4	4.5	5.0	4.5	4.8	4.5	4.4	4.5	4.9	4.7
Glu	5.5	4.9	5.0	4.6	5.1	4.5	4.8	5.0	4.5	5.0	5.0	4.9	5.1	4.8	4.7	5.0	4.4	4.3	4.6	5.1
Lys	4.6	5.0	5.0	4.8	4.8	5.5	5.0	5.0	4.4	5.0	4.8	4.7	4.4	4.9	4.8	4.8	5.0	4.4	4.5	5.0
Arg	5.0	4.5	4.8	5.1	4.8	4.7	5.1	5.0	4.5	4.9	4.7	5.0	4.5	4.7	4.7	4.5	5.0	4.9	4.9	4.6
Ser	4.1	4.9	5.0	4.9	4.7	4.9	4.8	4.8	5.0	5.1	4.4	4.5	4.6	4.5	4.7	4.6	5.0	4.4	5.1	5.1
Thr	5.8	5.0	5.1	4.9	5.1	4.9	5.1	4.9	4.5	4.8	4.9	4.7	4.5	5.0	4.7	4.6	5.0	4.7	4.7	4.7
Tyr	5.5	4.8	5.2	4.9	4.8	5.4	5.2	5.2	4.8	4.7	4.8	4.7	4.7	4.7	5.2	4.7	5.4	4.6	4.7	5.3
His	3.9	4.8	5.1	4.6	5.2	5.7	4.7	5.1	4.5	5.0	4.8	4.5	4.6	4.6	4.7	4.6	5.1	5.2	4.8	4.9
Cys	5.6	5.0	5.1	4.7	4.9	5.0	5.4	4.9	4.4	4.4	5.0	5.0	5.0	5.0	5.4	5.1	4.9	4.7	5.0	4.8
Asn	4.1	4.8	4.9	5.0	4.5	4.7	4.9	5.0	4.5	4.3	4.4	4.9	4.4	4.7	4.6	5.2	4.7	4.3	4.6	4.7
Gln	4.5	5.2	5.0	4.9	4.7	5.1	5.3	4.8	4.9	4.6	4.5	4.9	5.1	4.7	4.7	4.8	5.0	4.6	4.7	5.2
Trp	3.5	5.0	4.9	5.4	5.1	4.9	4.9	5.4	4.7	5.1	5.0	4.6	5.1	4.7	5.3	4.9	4.8	4.7	5.2	5.0

doi:10.1371/journal.pone.0019238.t001

for different pairs of side chains, or for the same pair of side chain with different torsion angles.

Generally speaking, the atom distance between different residues cannot be less than the sum of covalent radii (the disulfide bond is about 2.05 Å in length, which is almost equal to the sum of covalent radii of the S atom). Overlaps happen when one atom is within the range of the covalent volume of other atoms, i.e., $r_{ij} < r^{ij}_{ovl}$.

$$r^{ij}_{ovl} = R^{i}_{ovl} + R^{j}_{ovl}. \tag{10}$$

Here R^{i}_{ovl} is the 'overlap radius' for residue side chain atom i.

$$R_{ovl} = R_{\mathrm{cov}} + \delta_H R^{H}_{\mathrm{cov}}. \tag{11}$$

R_{cov} is the covalent radius of the atom; δ_H is a constant value as defined in the contact radius; and R^{H}_{cov} is the covalent radius of the hydrogen atom.

Isotropic sphere side chain (ISS) contact model

In many coarse-grained protein structure models, the isotropic sphere is used as a simplification of residue side chains [2,3,4,14,25]. The sphere model depends on two parameters: center position and radius. The geometry or mass center of heavy-atom collections is usually chosen as the sphere-shaped side chain center (**Figure 2** (D)). Although the radius of gyration, R_g, is commonly used in describing the size of the residue side chain, some atoms will be located outside the range of R_g. In the present study, the side chain radius is scaled to envelop all atoms. In order to determine the contact and overlap relationships between residues, effective 'contact radius' and 'overlap radius' have been proposed for sphere-shaped side chains.

The effective 'contact radius' R_{cnt} for isotropic sphere side chain is defined as:

$$R_{cnt} = r_{I_{\max}} + R^{I_{\max}}_{vdw} + \delta_H d_H. \tag{12}$$

Here $r_{I_{\max}}$ is the maximal radius of all r_i. The r_i denotes the distance between the side chain atom i and the center of the sphere. The atom index corresponding to the maximal radius is the $I_{\max}$. The $R^{I_{\max}}_{vdw}$ is the van der Waals radius of atom $I_{\max}$; d_H and δ_H have the same definitions as in the ADC model.

The effective 'overlap radius' is defined as:

$$R_{ovl} = r_{I_{\max}} + R^{I_{\max}}_{\mathrm{cov}} + \delta_H R^{H}_{\mathrm{cov}}. \tag{13}$$

Here $R^{I_{\max}}_{\mathrm{cov}}$ is the covalent radius of atom $I_{\max}$. The R^{H}_{cov} is the covalent radius of the hydrogen atom. Based on the contact and overlap radii of isotropic sphere side chain model, two cutoff distances are proposed:

$$r^{ij}_{cnt} = R^{i}_{cnt} + R^{j}_{cnt} + r_{gap}, \tag{14}$$

$$r^{ij}_{ovl} = R^{i}_{ovl} + R^{j}_{ovl}, \tag{15}$$

$$\begin{cases} r_{ij} \le r^{ij}_{ovl} & \text{overlap} \\ r^{ij}_{ovl} < r_{ij} \le r^{ij}_{cnt} & \text{contact} \\ r^{ij}_{cnt} < r_{ij} & \text{separated} \end{cases}. \tag{16}$$

Where r_{ij} is the distance between the center of the sphere of

residue i and j; r_{gap} is a gap distance between spherical surfaces, which can be adjusted to provide some flexibility to residue-attraction interactions.

Anisotropic ellipsoid side chain (AES) contact model

Although the ISS model works well for equi-axial, spheroidal atom systems, the radius-of-gyration techniques do not retain three-dimensional anisotropic properties with regard to side chain orientations. A more general ellipsoid side chain model is proposed in this work. The residue side chain is simulated as ellipsoids with three principal axes for arbitrarily shaped atom clusters. The orientations of resulting ellipsoids are then used to study relative positions of the residue side chain. All residue NHSAs are used to calculate three ellipsoidal radii.

The principal radii of the best-fit ellipsoid are along the transformed Cartesian coordinates axes. Principal axes can be obtained from the diagonalization of matrix $\mathbf{M}$.

$$\mathbf{M}=\begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix}. \tag{17}$$

Here, $\mathbf{M}$ represents the moment of inertia. The elements m_{ij} are calculated from the atom positions $\mathbf{x}=(x_1,x_2,x_3)$ relative to the side chain center $\mathbf{x}^0=(x_1^0,x_2^0,x_3^0)$, averaged over side chain atom number m [98,99]. The subscripts 1, 2 and 3 are coordinate indices.

$$\begin{aligned} m_{11} &= \frac{1}{m}\sum_{i=1}^{m}\left[(x_2(i)-x_2^0)^2+(x_3(i)-x_3^0)^2\right] \\ m_{22} &= \frac{1}{m}\sum_{i=1}^{m}\left[(x_1(i)-x_1^0)^2+(x_3(i)-x_3^0)^2\right], \\ m_{33} &= \frac{1}{m}\sum_{i=1}^{m}\left[(x_1(i)-x_1^0)^2+(x_2(i)-x_2^0)^2\right] \end{aligned} \tag{18}$$

$$\begin{aligned} m_{12}=m_{21} &= -\frac{1}{m}\sum_{i=1}^{m}\left[(x_1(i)-x_1^0)(x_2(i)-x_2^0)\right] \\ m_{13}=m_{31} &= -\frac{1}{m}\sum_{i=1}^{m}\left[(x_1(i)-x_1^0)(x_3(i)-x_3^0)\right]. \\ m_{23}=m_{32} &= -\frac{1}{m}\sum_{i=1}^{m}\left[(x_2(i)-x_2^0)(x_3(i)-x_3^0)\right] \end{aligned} \tag{19}$$

Where $(x_1(i),x_2(i),x_3(i))$ are the coordinates of atom i.The major and minor radii (known as the principal radius [99]) are determined directly from the Eigen values $(\lambda_1,\lambda_2,\lambda_3)$ of $\mathbf{M}$.

$$D(\mathbf{M})=\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}. \tag{20}$$

Here $D(\mathbf{M})$ denotes the diagonalization of $\mathbf{M}$ based on the cyclic Jacobi method [100]; $\lambda_1\geq\lambda_2\geq\lambda_3$ are three eigen values; and $r_3\geq r_2\geq r_1$ are the major and two minor semi-axes of the best-fit ellipsoid, respectively. If the atomic mass is assumed to be uniform and the side chain to have a unit mass, the eigen values are the principal moments of inertia for the ellipsoid side chain models (I_1,I_2,I_3).

$$\begin{aligned} \lambda_1 &= I_1=\frac{1}{5}(r_2^2+r_3^2) \\ \lambda_2 &= I_2=\frac{1}{5}(r_1^2+r_3^2). \\ \lambda_3 &= I_3=\frac{1}{5}(r_1^2+r_2^2) \end{aligned} \tag{21}$$

The principal radii of ellipsoid side chain are [99]:

$$\begin{aligned} r_1 &= \sqrt{\frac{5}{2}(\lambda_2+\lambda_3-\lambda_1)} \\ r_2 &= \sqrt{\frac{5}{2}(\lambda_3+\lambda_1-\lambda_2)}. \\ r_3 &= \sqrt{\frac{5}{2}(\lambda_1+\lambda_2-\lambda_3)} \end{aligned} \tag{22}$$

The ellipsoid orientation vector $\mathbf{v}$, with respect to the reference coordinate can also be obtained from the eigen vectors [98]. Using the ellipsoid side chain model, many spurious side chain overlaps can be avoided (see Results).

Some studies have been reported with regard to the detection of ellipsoid overlap [94,95]. In this study, we apply the algorithm to residue side chain contact determinations. For two ellipsoids A: $\mathbf{X}^T\mathbf{A}\mathbf{X}=0$ and B: $\mathbf{X}^T\mathbf{B}\mathbf{X}=0$, the solution of characteristic equation $\det(\lambda\mathbf{A}+\mathbf{B})=0$ is used as a simple algebraic condition for the separation of the ellipsoids. Here $\mathbf{X}=(x_1,x_2,x_3,w)^T$, where w is the 4^{th} dimension that represents the constant term in the ellipsoid formula; $\mathbf{A}$ and $\mathbf{B}$ are 4×4 real, symmetric matrices. The interiors of two ellipsoids are represented by $\mathbf{X}^T\mathbf{A}\mathbf{X}<0$ and $\mathbf{X}^T\mathbf{B}\mathbf{X}<0$. Then,

(1) A and B are separated if and only if $\det(\lambda\mathbf{A}+\mathbf{B})=0$ has two distinct positive roots.
(2) A and B touch externally if and only if $\det(\lambda\mathbf{A}+\mathbf{B})=0$ has a positive double root.
(3) A and B overlap if their characteristic equation has no positive root.

Matrix $\mathbf{A}$ and $\mathbf{B}$ can be constructed with the ellipsoid principal direction vectors $\mathbf{v}_i$ ($\mathbf{v}_i'$ for $\mathbf{B}$) and principal axis radii r_i (r_i' for $\mathbf{B}$) [94].

$$\mathbf{A}=\mathbf{P}_A\mathbf{D}\mathbf{P}_A{}^T=[\mathbf{v}_1\mathbf{v}_2\mathbf{v}_3\mathbf{v}_w]\begin{bmatrix} \frac{1}{r_1^2} & & & \\ & \frac{1}{r_2^2} & & \\ & & \frac{1}{r_3^2} & \\ & & & -1 \end{bmatrix}\begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \mathbf{v}_3^T \\ \mathbf{v}_w^T \end{bmatrix}, \tag{23}$$

$$\mathbf{B}=\mathbf{P}_B\mathbf{D}_B{\mathbf{P}_B}^T=[\mathbf{v'}_1\mathbf{v'}_2\mathbf{v'}_3\mathbf{v'}_w]$$
$$\begin{bmatrix} \frac{1}{r'_1{}^2} & & & -\frac{x_c}{r'_1{}^2} \\ & \frac{1}{r'_2{}^2} & & -\frac{y_c}{r'_2{}^2} \\ & & \frac{1}{r'_3{}^2} & -\frac{z_c}{r'_3{}^2} \\ -\frac{x_c}{r'_1{}^2} & -\frac{y_c}{r'_2{}^2} & -\frac{z_c}{r'_3{}^2} & -1+\frac{x_c^2}{r'_1{}^2}+\frac{y_c^2}{r'_2{}^2}+\frac{z_c^2}{r'_3{}^2} \end{bmatrix} \begin{bmatrix} \mathbf{v'}_1{}^T \\ \mathbf{v'}_2{}^T \\ \mathbf{v'}_3{}^T \\ \mathbf{v'}_w{}^T \end{bmatrix}. \quad (24)$$

Here (x_c,y_c,z_c) denotes the central position of ellipsoid B relative to the center of ellipsoid A. The fourth dimension of $\mathbf{v}_i$ corresponds to the constant term as in vector $\mathbf{X}$.

To avoid unnecessary calculations, two screening conditions are introduced prior to collision detection. The ellipsoid overlap and contact will be evaluated only if the side chain distances r_{ij} fall within a suitable range.

$$\begin{cases} r_{ij}<r_1^i+r_1^j & \text{overlap} \quad \text{(screening condition 1)} \\ r_{ij}>r_3^i+r_3^j+r_{gap} & \text{separated} \quad \text{(screening condition 2).} \\ r_1^i+r_1^j\le r_{ij}\le r_3^i+r_3^j+r_{gap} & \text{collision detection} \end{cases} \quad (25)$$

Here $r_3\ge r_2\ge r_1$ are the major and two minor semi-axes of the best-fit ellipsoid; r_{gap} is a gap distance between ellipsoid surfaces, which represents the decay zone of attraction interaction.

If the residue side chain distances clear the first two screening conditions, there still are three possibilities: overlap, contact or separated. According to the ellipsoid collision conditions, overlap can easily be sorted out. The problem is with regard to how contact from separated cases can be discriminated. In a similar manner as in the ADC and ISS models, 'contact radius' are proposed for the ellipsoid side chain model. With this contact radius, the ellipsoids can be scaled to include all atoms represented by van der Waals radius. Then, the ellipsoid collision conditions are checked for the enlarged ellipsoids. If the scaled ellipsoids are still separated, the two residue side chains are separated. Otherwise, the residue pair is said to be in contact.

Analysis of algorithm complexity

For a protein chain with N amino acids, the number of NHSA of residue k is $m(k)$. The ADC side chain contact algorithm needs to calculate the distances between two heavy atoms i and j. Let t_d be the complexity of distance operation $\|\mathbf{x}_i-\mathbf{x}_j\|$. The total complexity of ADC for a whole protein chain is:

$$\begin{aligned} T_{ADC} &= \left\{ \frac{\sum_{k=1}^{N} m(k)\left[\sum_{k=1}^{N} m(k)-1\right]}{2} - \sum_{k=1}^{N}\frac{m(k)[m(k)-1]}{2} \right\} t_d \\ &= \frac{N\cdot\bar{\boldsymbol{m}}(N\cdot\bar{\boldsymbol{m}}-1)-N\cdot\bar{\boldsymbol{m}}(\bar{\boldsymbol{m}}-1)}{2}t_d \\ &= \frac{1}{2}N^2\cdot\bar{\boldsymbol{m}}^2\cdot t_d\left(1-\frac{1}{N}\right) \end{aligned} \quad (26)$$

Here $\sum_{k=1}^{N} m(k)$ is the total number of NHSA. Distance calculation is not essential for atoms within the same residue. Thus, there is a deduction $\sum_{k=1}^{N}\frac{m(k)[m(k)-1]}{2}$; $\bar{\boldsymbol{m}}$ is the average NHSA number in side chain k with respect to all the N residues. As protein size N increases to a large value, T_{ADC} asymptotically approaches $\frac{N^2\bar{\boldsymbol{m}}^2}{2}t_d$.

In the case of the ISS model, there are three main steps. The bulky spherical centers have to be estimated first. Then the sphere radii are determined. Finally, the distance between two side chains is calculated and checked. If the geometrical center is considered the side chain center $\mathbf{x}_0$, the coordinate-averaging operation $\mathbf{x}_0=\frac{1}{m(k)}\sum_{i=1}^{m(k)}\mathbf{x}_i$ will be involved in calculations for residue k. Here $\mathbf{x}_i$ is the location of the non-hydrogen atom i. The largest atom-to-center distance $r_i^{\max}$ in the side chain k is utilized as isotropic sphere radius. In order to determine the $r_i^{\max}$, distance operation $\|\mathbf{x}_i-\mathbf{x}_0\|$ is carried out for all $m(k)$ atoms. Finally, the distances between any two side chain centers are calculated and checked.

If t_a is the approximate complexity of each add operation for $\frac{\mathbf{x}_i}{m(k)}$, then $m(k)\cdot t_a$ is the complexity of the coordinate-averaging operation $\mathbf{x}_0=\frac{1}{m(k)}\sum_{i=1}^{m(k)}\mathbf{x}_i$. Let t_d be the complexity of the distance operation $\|\mathbf{x}_i-\mathbf{x}_j\|$, and the total complexity of the ISS model will be:

$$\begin{aligned} T_{ISS} &= \sum_{k=1}^{N}[m(k)\cdot t_a+m(k)\cdot t_d]+\frac{N(N-1)}{2}t_d \\ &= N\cdot\bar{\boldsymbol{m}}(t_a+t_d)+\frac{N(N-1)}{2}t_d. \end{aligned} \quad (27)$$

Although residues have different rotamers, the side chain radius will not change too much for such conformational isomers. To simplify the process, the same type of residues is assumed to have the same radius. As a consequence, the calculation of radii is only necessary for 20 types of amino acid, rather than for all the N residues. The complexity can be re-written as:

$$\begin{aligned} T_{ISS} &= \sum_{k=1}^{20} m(k)\cdot t_d+\sum_{k=1}^{N} m(k)\cdot t_a+\frac{N(N-1)}{2}t_d \\ &= 20\bar{\boldsymbol{m}}_{20}t_d+N^2\left(\frac{\bar{\boldsymbol{m}}}{N}\right)t_a+N^2\left(\frac{1}{2}-\frac{1}{2N}\right)\cdot t_d. \end{aligned} \quad (28)$$

Here $\bar{\boldsymbol{m}}_{20}=\frac{1}{20}\sum_{i=1}^{20}m(i)$ is the average number of NHSA for 20 amino acids. There is little difference between $\bar{\boldsymbol{m}}_{20}$ and the average number $\bar{\boldsymbol{m}}=\frac{1}{N}\sum_{k=1}^{N}m(k)$ along the chain. When protein chain length N increases to a large value ($N\gg\bar{\boldsymbol{m}}$), T_{ISS} asymptotically approximates to $\left(20\bar{\boldsymbol{m}}_{20}+\frac{N^2}{2}\right)\cdot t_d$.

The complexity of the AES algorithm is comprised of several aspects: the creation of moment of inertia matrix $\mathbf{M}$, the

diagonalization of **M** and calculation of principal semi-radii, and the determ-
ination of residue contact according to ellipsoid collision conditions.

The elements of **M** are calculated from the relative positions of side chain atom to side chain center. In the same way as in the ISS algorithm, side chain-center calculation complexity is derived by $\sum_{k=1}^{N} m(k)\cdot t_a$. Here t_a is the approximate complexity related to coordinate-averaging operations. Considering the symmetry of the 3×3 matrix **M**, relative position estimations require $m(k)$ partial distance operations (only two coordinate axes are used) for each matrix element. Let t_d be the complexity of distance operation. Matrix creation has a complexity as:

$$\begin{aligned} T_{M1} &= \sum_{k=1}^{N} m(k)\cdot t_a + \sum_{k=1}^{N} 6m(k)\cdot t_d \\ &= N\cdot\bar{m}\cdot t_a + 6N\cdot\bar{m}\cdot t_d. \end{aligned} \tag{29}$$

The direct diagonalization of matrix **M** results in an algorithm complexity t_{diag}, which covers the computing cost of principal radii vectors. The total complexity for the whole protein chain is $T_{M2} = N\cdot t_{diag}$.

The ellipsoid-collision conditions are based on the solution of the characteristic equation $\det(\lambda A + B) = 0$. The constructions of A and B need products of three 4×4 matrices and the complexity is $2\cdot 2\cdot t_A$. Here t_A is the matrices multiplication complexity. The number of solutions can be obtained by solving the characteristic equation, which has a complexity independent of protein size and total atom number. This complexity is represented as $t_{\det}$. The ellipsoid collision complexity for N residues is $T_{M3} = 4N\cdot t_A + \frac{N(N-1)}{2} t_{\det}$.

From all the above analysis, the total complexity of the AES algorithm for an entire protein is:

$$\begin{aligned} T_{AES} &= N\cdot\bar{m}\cdot t_a + 6N\cdot\bar{m}\cdot t_d + N\cdot t_{diag} + 4N\cdot t_A + \frac{N(N-1)}{2} t_{\det} \\ &= N^2\left[\left(\frac{1}{2} - \frac{1}{2N}\right) t_{\det} + \frac{4}{N} t_A + \frac{1}{N} t_{diag} + \frac{6\bar{m}}{N} t_d + \frac{\bar{m}}{N} t_a\right]. \end{aligned} \tag{30}$$

When the protein size N increases to a very large value ($N \gg \bar{m}$), T_{AES} has an asymptotic approximation as $\frac{N^2}{2} t_{\det}$.

For large proteins ($N \gg \bar{m}$), the asymptotic approximation complexity of the ADC, ISS and AES algorithms are $\frac{N^2\bar{m}^2}{2} t_d$, $\left(\frac{N^2}{2} + 20\bar{m}_{20}\right)\cdot t_d$ and $\frac{N^2}{2} t_{\det}$, respectively; $\bar{m}$ is number of NHSA with respect to all the N residues in a protein; and $\bar{m}_{20}$ is the average number of NHSA with respect to 20 types of amino acids. The difference between $\bar{m}$ and $\bar{m}_{20}$ is trivial. Thus, an obvious fact is that the ADC model needs a significantly larger number of computations than the ISS model. The AES appears less complex than the ADC model. However, the $t_{\det}$ is much larger than t_d. If $t_{\det}$ can be written as $t_{\det} = c\cdot t_d$, the AES complexity will be $\frac{N^2}{2} c t_d$. The average number of NHSA usually satisfies $\bar{m} \approx 5$. As a consequence, ADC complexity is around $\frac{N^2}{2}\cdot 25 t_d$. When $t_{\det} \geq 25 t_d$, AES complexity exceeds that of the ADC model. In current algorithms, the complexity $t_{\det}$ for solving a fourth-order equation $\det(\lambda A + B) = 0$ and determining the number of different solutions is significantly greater than $25 t_d$. Stated briefly, the AES is currently the most computationally intensive algorithm model.

Author Contributions

Conceived and designed the experiments: WS JH. Performed the experiments: WS. Analyzed the data: WS JH. Contributed reagents/materials/analysis tools: WS JH. Wrote the paper: WS JH.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
2. Tanaka S, Scheraga HA (1975) Model of Protein Folding - Inclusion of Short-Range, Medium-Range, and Long-Range Interactions. Proceedings of the National Academy of Sciences of the United States of America 72: 3802–3806.
3. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. . Macromolecules 18: 534–552.
4. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 256: 623–644.
5. Bethe HA (1935) Statistical Theory of Superlattices. Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 150: 552–575.
6. Chang TS (1939) Statistical Theory of the Adsorption of Double Molecules. Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 169: 512–531.
7. Rushbrooke GS (1938) A Note on Guggenheim's Theory of Strictly Regular Binary Liquid Mixtures. Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 166: 296–315.
8. Miyazawa S (1983) Cooperative ligand binding on multidimensional lattices: Bethe approximation. Biopolymers 22: 2253–2271.
9. Guggenheim EA (1932) On the Statistical Mechanics of Dilute and of Perfect Solutions. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character 135: 181–192.
10. Guggenheim EA (1935) The Statistical Mechanics of Regular Solutions. Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 148: 304–312.
11. Guggenheim EA (1938) The Statistical Mechanics of Co-operative Assemblies. Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 169: 134–148.
12. Guggenheim EA (1944) Statistical Thermodynamics of Mixtures with Non-Zero Energies of Mixing. Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 183: 213–227.
13. Guggenheim EA (1944) Statistical Thermodynamics of Mixtures with Zero Energies of Mixing. Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 183: 203–212.
14. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213: 859–883.
15. Lu H, Skolnick J (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins-Structure Function and Genetics 44: 223–232.
16. Samudrala R, Moult J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology 275: 895–916.
17. Buchete NV, Straub JE, Thirumalai D (2004) Continuous anisotropic representation of coarse-grained potentials for proteins by spherical harmonics synthesis. Journal of Molecular Graphics & Modelling 22: 441–450.
18. Miyazawa S, Jernigan RL (2005) How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? J Chem Phys 122: 024901.
19. Lin W, Sun F, Rao ZH (2002) Tri-residue contact potential: a new knowledge-based energetic method. Progress in Natural Science 12: 826–840.
20. Godzik A, Kolinski A, Skolnick J (1992) Topology fingerprint approach to the inverse protein folding problem. J Mol Biol 227: 227–238.
21. Godzik A, Skolnick J (1992) Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. Proc Natl Acad Sci U S A 89: 12098–12102.
22. Singh RK, Tropsha A, Vaisman II (1996) Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol 3: 213–221.

23. Munson PJ, Singh RK (1997) Multi-body interactions within the graph of protein structure. Proc Int Conf Intell Syst Mol Biol 5: 198–201.
24. Munson PJ, Singh RK (1997) Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Sci 6: 1467–1481.
25. Mayewski S (2005) A multibody, whole-residue potential for protein structures, with testing by Monte Carlo simulated annealing. Proteins 59: 152–169.
26. Carter Jr. CW, LeFebvre BC, Cammer SA, Tropsha A, Edgell MH (2001) Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. Journal of Molecular Biology 311: 625–638.
27. Gan HH, Tropsha A, Schlick T (2001) Lattice protein folding with two and four-body statistical potentials. Proteins-Structure Function and Genetics 43: 161–174.
28. Deutsch C, Krishnamoorthy B (2007) Four-Body Scoring Function for Mutagenesis. Bioinformatics 23: 3009–3015.
29. Feng YP, Kloczkowski A, Jernigan RL (2007) Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. Proteins-Structure Function and Bioinformatics 68: 57–66.
30. Krishnamoorthy B, Tropsha A (2003) Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. Bioinformatics 19: 1540–1548.
31. Zhang C, Kim SH (2000) Environment-dependent residue contact energies for proteins. Proceedings of the National Academy of Sciences of the United States of America 97: 2550–2555.
32. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17: 355–362.
33. DeBolt SE, Skolnick J (1996) Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. Protein Eng 9: 637–655.
34. Gatchell DW, Dennis S, Vajda S (2000) Discrimination of near-native protein structures from misfolded models by empirical free energy functions. Proteins 41: 518–534.
35. Topf M, Sali A (2005) Combining electron microscopy and comparative protein structure modeling. Curr Opin Struct Biol 15: 578–585.
36. Topf M, Baker ML, Marti-Renom MA, Chiu W, Sali A (2006) Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. J Mol Biol 357: 1655–1668.
37. Maiorov VN, Crippen GM (1992) Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 227: 876–888.
38. Sippl MJ, Weitckus S (1992) Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 13: 258–271.
39. Bryant SH, Lawrence CE (1993) An empirical energy function for threading protein sequence through the folding motif. Proteins 16: 92–112.
40. Jones DT, Thornton JM (1996) Potential energy functions for threading. Curr Opin Struct Biol 6: 210–216.
41. Miyazawa S, Jernigan RL (1999) An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins 36: 357–369.
42. Miyazawa S, Jernigan RL (2000) Identifying sequence-structure pairs undetected by sequence alignments. Protein Eng 13: 459–475.
43. Skolnick J, Kolinski A, Ortiz A (2000) Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 38: 3–16.
44. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, et al. (1990) Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol 216: 167–180.
45. Casari G, Sippl MJ (1992) Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. J Mol Biol 224: 725–732.
46. Bauer A, Beyer A (1994) An improved pair potential to recognize native protein folds. Proteins 18: 254–261.
47. Samudrala R, Moult J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 275: 895–916.
48. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, et al. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 34: 82–95.
49. Sun WT, He J (2009) Native secondary structure topology has near minimum contact energy among all possible geometrically constrained topologies. Proteins-Structure Function and Bioinformatics 77: 159–173.
50. Sun WT, He J (2009) Reduction of the secondary structure topological space through direct estimation of the contact energy formed by the secondary structures. BMC Bioinformatics 10(suppl 1): S40.
51. Al Nasr K, Sun WT, He J (2010) Structure prediction for the helical skeletons detected from the low resolution protein density map. BMC Bioinformatics 11(suppl 1): S44.
52. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253: 164–170.
53. Sun S (1993) Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. Protein Sci 2: 762–785.
54. Tobi D, Elber R (2000) Distance-dependent, pair potential for protein folding: results from linear optimization. Proteins 41: 40–46.
55. Tobi D, Shafran G, Linial N, Elber R (2000) On the design and analysis of protein folding potentials. Proteins 40: 71–85.
56. Li J, Wang J, Wang W (2008) Identifying folding nucleus based on residue contact networks of proteins. Proteins 71: 1899–1907.
57. Kannan N, Vishveshwara S (1999) Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol 292: 441–464.
58. Greene LH, Higman VA (2003) Uncovering network systems within protein structures. J Mol Biol 334: 781–791.
59. Cohen M, Potapov V, Schreiber G (2009) Four Distances between Pairs of Amino Acids Provide a Precise Description of their Interaction. Plos Computational Biology 5: e1000470.
60. Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Sci 11: 430–448.
61. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714–2726.
62. Levitt M, Warshel A (1975) Computer simulation of protein folding. Nature 253: 694–698.
63. Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104: 59–107.
64. Lee J, Liwo A, Scheraga HA (1999) Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10–55 fragment of staphylococcal protein A and to apo calbindin D9K. Proc Natl Acad Sci U S A 96: 2025–2030.
65. Lee J, Ripoll DR, Czaplewski C, Pillardy J, Wedemeyer WJ, et al. (2001) Optimization of parameters in macromolecular potential energy functions by conformational space annealing. Journal of Physical Chemistry B 105: 7291–7298.
66. Liwo A, Arlukowicz P, Czaplewski C, Oldziej S, Pillardy J, et al. (2002) A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field. Proceedings of the National Academy of Sciences of the United States of America 99: 1937–1942.
67. Manavalan P, Ponnuswamy PK (1977) Study of Preferred Environment of Amino-Acid Residues in Globular Proteins. Archives of Biochemistry and Biophysics 184: 476–487.
68. Manavalan P, Ponnuswamy PK (1978) Hydrophobic Character of Amino-Acid Residues in Globular Proteins. Nature 275: 673–674.
69. Gromiha MM, Selvaraj S (1999) Amino acid clustering pattern and medium and long-range interactions in (alpha/beta)(8) barrel proteins. Periodicum Biologorum 101: 333–338.
70. Selvaraj S, Gromiha MM (2000) Inter-residue interactions in protein structures. Current Science 78: 129–131.
71. Debe DA, Goddard WA (1999) First principles prediction of protein folding rates. Journal of Molecular Biology 294: 619–625.
72. Gromiha MM, Selvaraj S (1999) Importance of long-range interactions in protein folding. Biophysical Chemistry 77: 49–68.
73. Gromiha MM (2001) Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. Biophysical Chemistry 91: 71–77.
74. Atilgan AR, Akan P, Baysal C (2004) Small-world communication of residues and significance for protein dynamics. Biophysical Journal 86: 85–91.
75. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297: 1551–1555.
76. Tudos E, Fiser A, Simon I (1994) Different Sequence Environments of Amino-Acid-Residues Involved and Not Involved in Long-Range Interactions in Proteins. International Journal of Peptide and Protein Research 43: 205–208.
77. Barah P, Sinha S (2008) Analysis of protein folds using protein contact networks. Pramana-Journal of Physics 71: 369–378.
78. Yang L, Song G, Jernigan RL (2009) Protein elastic network models and the ranges of cooperativity. Proc Natl Acad Sci U S A 106: 12347–12352.
79. Rao F, Caflisch A (2004) The protein folding network. Journal of Molecular Biology 342: 299–306.
80. Bode C, Kovacs IA, Szalay MS, Palotai R, Korcsmaros T, et al. (2007) Network analysis of protein dynamics. FEBS Letters 581: 2776–2782.
81. Krishnan A, Zbilut JP, Tomita M, Giuliani A (2008) Proteins as networks: Usefulness of graph theory in protein science. Current Protein & Peptide Science 9: 28–38.
82. Chandrasekaran R, Ramachandran GN (1970) Studies on the conformation of amino acids. XI. Analysis of the observed side group conformation in proteins. Int J Protein Res 2: 223–233.
83. Sasisekharan V, Ponnuswamy PK (1970) Backbone and side-chain conformations of amino acids and amino acid residues in peptides. Biopolymers 9: 1249–1256.
84. von Schnakenburg K (1971) [Light and electron microscopy studies on brain tissue changes in acute experimental oxygen intoxication]. Virchows Arch B Cell Pathol 8: 230–242.
85. Janin J, Wodak S (1978) Conformation of amino acid side-chains in proteins. J Mol Biol 125: 357–386.
86. Bhat TN, Sasisekharan V, Vijayan M (1979) An analysis of side-chain conformation in proteins. Int J Pept Protein Res 13: 170–184.
87. Benedetti E, Morelli G, Nemethy G, Scheraga HA (1983) Statistical and energetic analysis of side-chain conformations in oligopeptides. Int J Pept Protein Res 22: 1–15.

88. James MN, Sielecki AR (1983) Structure and refinement of penicillopepsin at 1.8 A resolution. J Mol Biol 163: 299–361.
89. Ponder JW, Richards FM (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 193: 775–791.
90. Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. Proteins-Structure Function and Genetics 40: 389–408.
91. Richardson DC, Chen VB, Arendall WB, Headd JJ, Keedy DA, et al. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D-Biological Crystallography 66: 12–21.
92. Gilat A (2004) MATLAB: An Introduction with Applications 2nd Edition: John Wiley & Sons.
93. Sun W, He J (2010) Understanding on the Residue Contact Network Using the Log-Normal Cluster Model and the Multilevel Wheel Diagram. Biopolymers 93: 904–916.
94. Wang W, Wang J, Kim M-S (2001) An algebraic condition for the separation of two ellipsoids. Comput Aided Geom Des 18: 531–539.
95. Wang W, Choi Y-K, Chan B, Kim M-S, Wang J (2004) Efficient collision detection for moving ellipsoids using separating planes. Computing 72: 235–246.
96. Sun W, He J (2009) Effect of sidechain anisotropy on residue contact determination. In: Chen J, Reddy CK, Chen X, Ruan J, Ely J, et al., eds. 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshops; 2009 Nov; Washington D.C, USA. pp 181–188.
97. Lu M, Dousis AD, Ma J (2008) OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol 376: 288–301.
98. Karnesky RA, Sudbrack CK, Seidman DN (2007) Best-fit ellipsoids of atom-probe tomographic data to study coalescence of gamma' (L1(2)) precipitates in Ni-Al-Cr. Scripta Materialia 57: 353–356.
99. Nye JF (1985) Physical Properties of Crystals: Their Representation by Tensors and Matrices: Oxford University Press.
100. Forsythe GE, Henrici P (1960) The Cyclic Jacobi Method for Computing the Principal Values of a Complex Matrix. Transactions of the American Mathematical Society 94: 1–23.