2011

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

Weitao Sun

Jing He
Old Dominion University

Follow this and additional works at: http://digitalcommons.odu.edu/computerscience_fac_pubs

Part of the Computer Sciences Commons, and the Mathematics Commons

Repository Citation
Sun, Weitao and He, Jing, 'From Isotropic to Anisotropic Side Chain Representations: Comparison of Three Models for Residue Contact Estimation' (2011). Computer Science Faculty Publications. 52.
http://digitalcommons.odu.edu/computerscience_fac_pubs/52

Original Publication Citation

This Article is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Faculty Publications by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.
From Isotropic to Anisotropic Side Chain Representations: Comparison of Three Models for Residue Contact Estimation

Weitao Sun¹, Jing He²

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 \( C\alpha - C\alpha \) 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.

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 utilized 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-depending 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 \( i \) and \( j \) 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 \( C\alpha \) atom [62,63], \( C\beta \) 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 \( C\alpha \) 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 \( C_i \) atom pair or \( C_g \) 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 (pENM) 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 anisotropic 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. We derive a dataset and illustrate the pairs that were wrongly estimated using the AES model for future study to improve the AES model.

### 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 \( n_{cnt}(i) \).

\[
n_{cnt}(i) = \sum_{j=1}^{n_r} A_{ij}
\]

\( A_{ij} = \begin{cases} 1 & \text{if residue } i \text{ and } j \text{ are in contact} \\ 0 & \text{otherwise} \end{cases} \)

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 \( n_{cnt}(i) = n_{ovl}(i) + 1 \) 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.

\[
\lambda(n_r) = \sum_{i=1}^{n_r} \sum_{j=1}^{n_r} A_{ij}.
\]

\( \lambda(n_r) \) 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.

\[
\lambda_{ovl}(n_r) = \sum_{i=1}^{n_r} \sum_{j=1}^{n_r} B_{ij},
\]

\( B_{ij} = \begin{cases} 1 & \text{if residue } i \text{ and } j \text{ overlap} \\ 0 & \text{otherwise} \end{cases} \)

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 \( \lambda(n_r) = k n_r + b \) 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 \( \lambda(n_r) = 0.5 n_r - 8.79 \) (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 \( k_i = \frac{\lambda(n_r) - b_{all}}{n_r} \). 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 \( \lambda(n_r) = 1.3 n_r - 8.81 \) 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 helices; (3) proteins composed of curvy 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 \( r_{cut} \).
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, $\lambda(n_c)$ and $\lambda_{out}(n_c)$ are calculated for $r_{gap} = 0.5\,\text{Å}, 1.0\,\text{Å}, 1.5\,\text{Å}$ and $2.0\,\text{Å}$, respectively. The contact cutoff distance $r_{cut}$ 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 1. Residue contact distribution by ADC/ISS/AES contact model. $\lambda(n_c)$ and $\lambda_{out}(n_c)$ denotes the total number of contact and overlap among all residues in the protein. Surface gap distance $r_{gap} = 0.0\,\text{Å}$ is used here. Data points for all-$\alpha$ helix, all-$\beta$ sheet, $\alpha$ helix-$\beta$ 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
In this study, we used the geometry center $x_0$ of heavy-atoms as the center of side chain.

$$x_0 = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

(6)

Here $x_i$ 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 $\lambda_{ovl}(n_t)$, 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 $\lambda(n_t) = 1.9n_t - 51$ 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 $\lambda(n_t)$ behavior of the ISS model with a surface-gap distance $r_{gap} = 0.0 \text{ Å}$ appears to be equivalent to that in the ADC model with an atom surface-gap distance of $r_{gap} = 1.0 \text{ Å}$ or $r_{gap} = 1.5 \text{ Å}$.

We also calculated the distributions of $\lambda(n_t)$ and $\lambda_{ovl}(n_t)$ with $r_{gap} = 0.5, 1.0, 1.5$ and $2.0 \text{ Å}$ respectively for the ISS model (data not shown). The increase in $r_{gap}$ leads to a simultaneous increase in the residue-contact cutoff distance $r_{ij}^{cnt}$. As a consequence, more residue side chain pairs are considered to be within the contact range. However, $\lambda_{ovl}(n_t)$ distributions do not change with different $r_{gap}$ for the reason that the overlap cutoff distance $r_{ij}^{ovl}$ 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 model (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 $\lambda_{ovl}$ are less than 10. This number appears 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 20x20 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).
The relation between $r_{ij}$ 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 $r_{ij} = 5$ 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 4. Examples of the ellipsoid overlap and side chain positions in PDB structures. (A) Glu-Arg overlap in 1IO0; (B) Glu260-Arg286 side chain positions in 1IO0; (C) Asp-Arg overlap in 1C7K; (D) Asp49-Arg51 side chain positions in 1C7K.
doi:10.1371/journal.pone.0019238.g004
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 contact distance \( r_{cnt} \). First, \( r_{cnt} \) 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 \( r_{cnt} \). Then, from among all gap distances, there should be one critical value at which all contact-pair distances are less than the optimal \( r_{cnt} \). 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 \( r_{cnt} = 5 \) 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 \( r_{cnt}^{ij} \), 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
closely packed and loosely packed groups with the help of two different linear fit of $n_i$ 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 model. 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, which reflects the anisotropy in side chain orientations. With the current definition, the cutoff distance can be different values for different residues.

$$\text{rij}^\text{cnt} = \text{R}^\text{vdw}_i + \delta_H d_H.$$  (7)

Where $R^\text{vdw}_i$ 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; $d_H = 0.4R^\text{vdw}_H$ was used in the current study; $R^\text{vdw}_H$ denotes the van der Waals radius of hydrogen atom; and $\delta_H$ is a constant value, which is defined as:

$$\delta_H = \begin{cases} 0 & \text{Has attached hydrogen atom} \\ 1 & \text{No attached hydrogen atom} \end{cases}.$$  (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 $r_{ij} < r^\text{cnt}_{ij}$, the atoms are considered to be in contact. Other than a predetermined fixed value, the atom-contact cutoff distance $r^\text{cnt}_{ij}$ is calculated based on side chain atom-surface distance, which reflects the anisotropy in side chain orientations.

$$r^\text{cnt}_{ij} = r^\text{cnt}_i + r^\text{cnt}_j + r_{\text{gap}}.$$  (9)

$R^\text{cnt}_i, R^\text{cnt}_j$ are contact radii of atoms $i$ and $j$. The $r_{\text{gap}}$ is the gap distance representing the decay of atomic interaction. With the current definition, the cutoff distance $r^\text{cnt}_{ij}$ can be different values...
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., \( r_g < r_{ol}^{ij} \).

\[
\frac{\rho_{ol}^{ij}}{R_{ol}} = R^{ol} + R^{ovl}. \tag{10}
\]

Here \( R^{ovl} \) is the ‘overlap radius’ for residue side chain atom \( i \).

\[
R^{ol} = R^{cov} + \delta_H R^{H}_{cov}. \tag{11}
\]

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

### Isotropic sphere 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:

\[
R_{cnt} = r_{\text{max}} + R^{\text{max}}_{\text{ovl}} + \delta_H d_H. \tag{12}
\]

Here \( r_{\text{max}} \) 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 \( I_{\text{max}} \). The \( R^{\text{max}}_{\text{ovl}} \) is the van der Waals radius of atom \( I_{\text{max}} \); \( d_H \) and \( \delta_H \) have the same definitions as in the ADC model.

The effective ‘overlap radius’ is defined as:

\[
R^{ol}_{\text{cnt}} = r_{\text{max}} + R^{\text{max}}_{\text{ovl}} + \delta_H R^{H}_{\text{cov}}. \tag{13}
\]

Here \( R^{\text{max}}_{\text{ovl}} \) is the covalent radius of atom \( I_{\text{max}} \). The \( R^{H}_{\text{cov}} \) 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:

\[
r^{ij}_{\text{cnt}} = r^{\text{cnt}}_{ij} + R^{\text{cnt}}_{ol} + r_{\text{gap}}. \tag{14}
\]

\[
r^{ij}_{\text{ol}} = R^{ij}_{ol} + R^{\text{ol}}_{ol}, \tag{15}
\]

\[
\begin{cases}
  r_{ij} \leq r^{ij}_{\text{ol}} & \text{overlap} \\
  r^{ij}_{\text{ol}} < r_{ij} \leq r^{ij}_{\text{cnt}} & \text{contact} \\
  r^{ij}_{\text{cnt}} < r_{ij} & \text{separated}
\end{cases} \tag{16}
\]

Where \( r_{ij} \) is the distance between the center of the sphere of...
residue $i$ and $j$; $r_{\text{gap}}$ 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 $M$.

\[
M = \begin{bmatrix}
m_{11} & m_{12} & m_{13} \\
m_{21} & m_{22} & m_{23} \\
m_{31} & m_{32} & m_{33}
\end{bmatrix}.
\]

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

\[
m_{11} = \frac{1}{m} \sum_{i=1}^{m} \left( (x_2(i)-x_2^0)^2 + (x_3(i)-x_3^0)^2 \right)
\]

\[
m_{22} = \frac{1}{m} \sum_{i=1}^{m} \left( (x_1(i)-x_1^0)^2 + (x_3(i)-x_3^0)^2 \right),
\]

\[
m_{33} = \frac{1}{m} \sum_{i=1}^{m} \left( (x_1(i)-x_1^0)^2 + (x_2(i)-x_2^0)^2 \right)
\]

\[
m_{12} = m_{21} = -\frac{1}{m} \sum_{i=1}^{m} \left( (x_1(i)-x_1^0)(x_2(i)-x_2^0) \right)
\]

\[
m_{13} = m_{31} = -\frac{1}{m} \sum_{i=1}^{m} \left( (x_1(i)-x_1^0)(x_3(i)-x_3^0) \right).
\]

\[
m_{23} = m_{32} = -\frac{1}{m} \sum_{i=1}^{m} \left( (x_2(i)-x_2^0)(x_3(i)-x_3^0) \right)
\]

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 $M$.

\[
D(M) = \begin{bmatrix}
\lambda_1 & 0 & 0 \\
0 & \lambda_2 & 0 \\
0 & 0 & \lambda_3
\end{bmatrix}.
\]

Here $D(M)$ denotes the diagonalization of $M$ based on the cyclic Jacobi method [100]; $\lambda_1 \geq \lambda_2 \geq \lambda_3$ are three eigen values; and $r_1 \geq r_2 \geq r_3$ 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 $(I_1, I_2, I_3)$:

\[
\lambda_1 = I_1 = \frac{1}{5} (r_1^2 + r_2^2)
\]

\[
\lambda_2 = I_2 = \frac{1}{5} (r_2^2 + r_3^2),
\]

\[
\lambda_3 = I_3 = \frac{1}{5} (r_1^2 + r_3^2).
\]

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

\[
r_1 = \sqrt{\frac{5}{2} (\lambda_2 + \lambda_3 - \lambda_1)}
\]

\[
r_2 = \sqrt{\frac{5}{2} (\lambda_1 + \lambda_3 - \lambda_2)},
\]

\[
r_3 = \sqrt{\frac{5}{2} (\lambda_1 + \lambda_2 - \lambda_3)}
\]

The ellipsoid orientation vector $v$, 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 $A$: $X^TAX = 0$ and $B$: $X^TBX = 0$, the solution of characteristic equation $\det (\lambda A + B) = 0$ is used as a simple algebraic condition for the separation of the ellipsoids. Here $X = (x_1, x_2, x_3, w)^T$, where $w$ is the 4th dimension that represents the constant term in the ellipsoid formula; $A$ and $B$ are $4 \times 4$ real, symmetric matrices. The interiors of two ellipsoids are represented by $X^TAX < 0$ and $X^TBX < 0$. Then,

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

Matrix $A$ and $B$ can be constructed with the ellipsoid principal direction vectors $v_i$ ($v_i$ for $B$) and principal axis radii $r_i$ ($r_i$ for $B$) [94].

\[
A = P_4DP_4^T = [v_1 v_2 v_3 v_w]
\]

\[
\begin{bmatrix}
\frac{1}{r_1^2} & 0 & 0 & 0 \\
0 & \frac{1}{r_2^2} & 0 & 0 \\
0 & 0 & \frac{1}{r_3^2} & 0 \\
0 & 0 & 0 & -1
\end{bmatrix}
\]

(23)
\[ B = P_{\beta} D_{\beta} P_{\beta}^T = \left[ v_1^T v_2^T v_3^T v_4^T \right] \]

Here \((x_1, y_1, z_1)\) denotes the central position of ellipsoid \(B\) relative to the center of ellipsoid \(A\). The fourth dimension of \(v_j\) corresponds to the constant term as in vector \(X\).

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.

\[
\begin{align*}
\text{Overlap:} & \quad r_{ij} < r_1 + r_1' & \quad (\text{screening condition 1}) \\
\text{Separation:} & \quad r_{ij} > r_1 + r_1' + r_{gap} & \quad (\text{screening condition 2}) \\
\end{align*}
\]

Here \(r_1 \geq r_2 \geq r_1'\) 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 \(\|x_i - x_j\|\). The total complexity of ADC for a whole protein chain is:

\[
T_{ADC} = \left\{ \sum_{k=1}^{N} \frac{m(k)}{2} \right\} \cdot T_d
\]

The largest \(N\) atoms of protein chain length \(N\) increases to a large value \((N \gg \bar{m})\), \(T_{ADC}\) asymptotically approaches \(N^2 \bar{m}^2 \cdot T_d\).

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 \(\sum_{k=1}^{N} \frac{m(k)(m(k) - 1)}{2}\), \(\bar{m}\) 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 \(N^2 \bar{m}^2 \cdot T_d\).

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 \(x_0\), the coordinate-averaging operation \(x_0 = \frac{1}{m(k)} \sum_{i=1}^{m(k)} x_i\) will be involved in calculations for residue \(k\).

Here \(x_i\) is the location of the non-hydrogen atom \(i\). The largest atom-to-center distance \(r_{\text{max}}\) in the side chain \(k\) is utilized as isotropic sphere radius. In order to determine the \(r_{\text{max}}\), distance operation \(\|x_i - x_0\|\) 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 \(\frac{m(k)}{m(k)} \), then \(m(k)\cdot T_a\) is the complexity of the coordinate-averaging operation \(x_0 = \frac{1}{m(k)} \sum_{i=1}^{m(k)} x_i\). Let \(T_d\) be the complexity of the distance operation \(\|x_i - x_j\|\), and the total complexity of the ISS model will be:

\[
T_{ISS} = \sum_{k=1}^{N} \left[ m(k) \cdot T_a + m(k) \cdot T_d \right] + \frac{N(N - 1)}{2} T_d
\]

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:

\[
T_{ISS} = \sum_{k=1}^{20} \left[ m(k) \cdot T_a + m(k) \cdot T_d \right] + \frac{N(N - 1)}{2} T_d
\]

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

The complexity of the AES algorithm is comprised of several aspects: the creation of moment of inertia matrix \(M\), the
diagonalization of $\mathbf{M}$ and calculation of principal semi-radii, and the determination of residue contact according to ellipsoid collision conditions.

The elements of $\mathbf{M}$ 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 $N\sum_{k=1}^{n} m(k)\cdot t_a$. Here $t_a$ is the approximate complexity related to coordinate-averaging operations. Considering the symmetry of the $3 \times 3$ matrix $\mathbf{M}$, 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:

$$T_{M1} = N\sum_{k=1}^{n} m(k)\cdot t_a + \sum_{k=1}^{n} 6m(k)\cdot t_d$$

$$= N\hat{m}\cdot t_a + 6N\hat{m}\cdot t_d. \quad (29)$$

The direct diagonalization of matrix $\mathbf{M}$ 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 $\det(\mathbf{A}+\mathbf{B})=0$. The constructions of $A$ and $B$ need products of three $4 \times 4$ matrices and the complexity is $2\cdot 2\cdot t_d$. Here $t_d$ 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 $T_{M3} = 4N\cdot t_d + \frac{N(N-1)}{2} t_{det}$.

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

$$T_{AES} = N\hat{m}\cdot t_a + 6N\hat{m}\cdot t_d + 4N\cdot t_{det} + \frac{N(N-1)}{2} t_{det}$$

$$= N^2 \left[ \frac{1}{2} - \frac{1}{2N} \right] t_{det} + 4 \frac{N\hat{m}}{N} t_a + \frac{1}{N} 6m t_d + \frac{m}{N} t_d \right] \cdot (30)$$

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

For large proteins ($N \gg m$), the asymptotic approximation complexity of the ADC, ISS and AES algorithms are $\frac{N^2\hat{m}^2}{2} t_d$, $\frac{N^2\hat{m}^2}{2} t_d$ and $\frac{N^2}{2} t_{det}$, respectively; $\hat{m}$ is number of NHSA with respect to all the $N$ residues in a protein; and $\hat{m}_p$ is the average number of NHSA with respect to 20 types of amino acids. The difference between $\hat{m}$ and $\hat{m}_p$ 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 $\frac{N^2}{2} ct_d$. The average number of NHSA usually satisfies $\hat{m} \approx 5$. As a consequence, ADC complexity is around $\frac{N^2}{2} \cdot 25t_d$. When $t_{det} \approx 25t_d$, AES complexity exceeds that of the ADC model. In current algorithms, the complexity $t_{det}$ for solving a fourth-order equation $\det(\mathbf{A}+\mathbf{B})=0$ and determining the number of different solutions is significantly greater than $25t_d$. 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**