

Optimizing the atom types of proteins through iterative knowledge-based potentials

Xinxiang Wang(汪心享), Sheng-You Huang(黄胜友)
Citation: Chin. Phys. B . 2018, 27(2): 020503. **doi:** 10.1088/1674-1056/27/2/020503

Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Moderate point: Balanced entropy and enthalpy contributions in soft matter

Baoji He(贺宝记), Yanting Wang(王延颋)
Chin. Phys. B . 2017, 26(3): 030506. **doi:** 10.1088/1674-1056/26/3/030506

Knowledge-based potentials in bioinformatics: From a physicist's viewpoint

Zheng Wei-Mou
Chin. Phys. B . 2015, 24(12): 0128701. **doi:** 10.1088/1674-1056/24/12/128701

Nonequilibrium thermodynamics and fluctuation relations for small systems

Cao Liang, Ke Pu, Qiao Li-Yan, Zheng Zhi-Gang
Chin. Phys. B . 2014, 23(7): 070501. **doi:** 10.1088/1674-1056/23/7/070501

Incomplete nonextensive statistics and the zeroth law of thermodynamics

Huang Zhi-Fu, Ou Cong-Jie, Chen Jin-Can
Chin. Phys. B . 2013, 22(4): 040501. **doi:** 10.1088/1674-1056/22/4/040501

Variational iteration method for solving the time-fractional diffusion equations in porous medium

Wu Guo-Cheng
Chin. Phys. B . 2012, 21(12): 0120504. **doi:** 10.1088/1674-1056/21/12/120504

SPECIAL TOPIC — Soft matter and biological physics

Optimizing the atom types of proteins through iterative knowledge-based potentials*

Xin-Xiang Wang(汪心享) and Sheng-You Huang(黄胜友)†

School of Physics, Huazhong University of Science and Technology, Wuhan 430074, *China*

(Received 28 September 2017; revised manuscript received 30 December 2017; published online 29 January 2018)

Knowledge-based scoring functions have been widely used for protein structure prediction, protein–small molecule, and protein–nucleic acid interactions, in which one critical step is to find an appropriate representation of protein structures. A key issue is to determine the minimal protein representations, which is important not only for developing of scoring functions but also for understanding the physics of protein folding. Despite significant progresses in simplifying residues into alphabets, few studies have been done to address the optimal number of atom types for proteins. Here, we have investigated the atom typing issue by classifying the 167 heavy atoms of proteins through 11 schemes with 1 to 20 atom types based on their physicochemical and functional environments. For each atom typing scheme, a statistical mechanics-based iterative method was used to extract atomic distance-dependent potentials from protein structures. The atomic distance-dependent pair potentials for different schemes were illustrated by several typical atom pairs with different physicochemical properties. The derived potentials were also evaluated on a high-resolution test set of 148 diverse proteins for native structure recognition. It was found that there was a crossover around the scheme of four atom types in terms of the success rate as a function of the number of atom types, which means that four atom types may be used when investigating the basic folding mechanism of proteins. However, it was revealed by a close examination of typical potentials that 14 atom types were needed to describe the protein interactions at atomic level. The present study will be beneficial for the development of protein related scoring functions and the understanding of folding mechanisms.

Keywords: atom types, knowledge-based potentials, statistical mechanics, iteration

PACS: 05.20.–y, 87.14.E–, 87.15.ad **DOI:** 10.1088/1674-1056/27/2/020503

1. Introduction

Appropriate representation of protein structures is important in computational structural biology. It is associated with not only the development of scoring functions but also the understanding of protein folding mechanism. Coarse-grained model is one of the methods to represent protein structures and often applied to reduce computational cost. A key factor in the model is to determine the minimal protein representations, residue-based or atom-based, without sacrificing essential precision. Simplifying twenty amino acids into fewer numbers of representative alphabets has been widely studied in protein folding,[1–5] molecular docking,[6] protein structure prediction,[7,8] protein design,[9] protein function,[10] protein classification,[11–13] and protein sequence alignment.[14,15] Baker *et al*. designed a 57-residues protein Src SH3 with a reduced alphabet using five amino acid types.[5] Motivated by their work, Wang and Wang obtained an optimal reduction with five types of residues that has the same form as the simplified palette of Baker and coworkers based on the concept of mismatch between a reduced interaction matrix and the Miyazawa and Jernigan (MJ) matrix.[1] In coarse-grained molecular simulations, lattice chain models of two residue types or more (HP, HNP, IAGEK, etc.) provided a good understanding on the folding process of natural proteins.[4,9,16–18] Although residue-based typing issues in protein modeling, folding, and design have been studied by researchers,[1,3,19–23] the optimal atom classification for proteins has received little attention. In fact, during the scoring process of protein structure prediction, protein–protein, protein–ligand, and protein–nucleic acid interactions, atom-based potentials are often required for the accurate evaluation of these interactions.[24–28] For years, various protein atom typing schemes have been developed to characterize the interactions between atoms.[29–34] However, the optimization about atom typing remains an important issue. Current approaches normally classify protein atoms based on their physical and chemical properties. These properties include atomic number, charge, polarity, hydrophobicity, hydrogen bond, local chemical and protein secondary structure environment, etc.[35–39]

Despite the successes of current atom typing schemes on some systems, their classification methods for atom types are all kind of arbitrary and depend on specific systems studied. Given the 167 heavy atoms of 20 amino acids, protein atoms could be grouped into 1 to 167 types.[36,37] If a scheme has

*Project supported by the National Natural Science Foundation of China (Grant No. 31670724), the National Key Research and Development Program of China (Grant Nos. 2016YFC1305800 and 2016YFC1305805), and the Startup Grant of Huazhong University of Science and Technology, China.

†Corresponding author. E-mail: huangsy@hust.edu.cn

© 2018 Chinese Physical Society and IOP Publishing Ltd

a small number of atom types, it will make the potentials of scoring functions simple and fast, but the resolution of the potentials will be sacrificed in characterizing the atomic interactions. On the contrary, if a scheme has a large number of atom types, the corresponding potentials will have a better resolution in describing the interactions, but the potentials may suffer from the slow speed due to more possible interacting pairs and more errors due to insufficient statistics in the derivation of scoring functions.[32–35] Therefore, an appropriate typing scheme is needed to achieve a good balance between accuracy and resolution, especially for knowledge-based scoring functions.[27]

In this work, we have used 11 schemes to categorize protein atoms based on their physical and chemical environment in proteins. The 11 schemes gave 1 to 20 atom types when different levels of details for atoms' environments were taken into account. Using a statistical mechanics-based iterative method[35] and a training set of 1225 proteins,[40] we have derived 11 sets of knowledge-based pair potentials based on the 11 protein atom typing schemes. The goodness of different atom typing schemes were assessed based on the ability of the developed scoring functions in discriminating native structures from decoys on a test set of 148 proteins with 500–1600 high resolution (HR) decoys for each protein.[40] We have also examined the derived knowledge-based potentials of some typical atom pairs to obtain a relation between the accuracy of the potentials and the dimension of the parameters. The present work provides a reference on the optimal number of atom types for proteins in the development of scoring functions and will be useful for the study of protein design and structure prediction.

2. Materials and methods

2.1. Eleven protein atom typing schemes

The 11 schemes have been used to classify the 167 heavy atoms of proteins based on their physical, chemical, and functional environment. The details of the atom typing schemes are illustrated in Fig. 1. The number of atom types for each scheme increases when more details are considered. In the present study, the simplest atom typing scheme is to group all the heavy atoms into one type. The most advanced scheme is characterized with twenty protein atom types (Table 1 and Fig. 1), which has been used in our previous study and demonstrated efficient in discriminating native structures from decoys.[35] Starting from twenty protein atom types, "similar" atom types are gradually combined together according to their similar properties, and thus the dimension of the atom typing scheme becomes smaller and smaller, until all the atom types are combined into one type.[27,35] The 11 schemes result in 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, and 20 atom types,

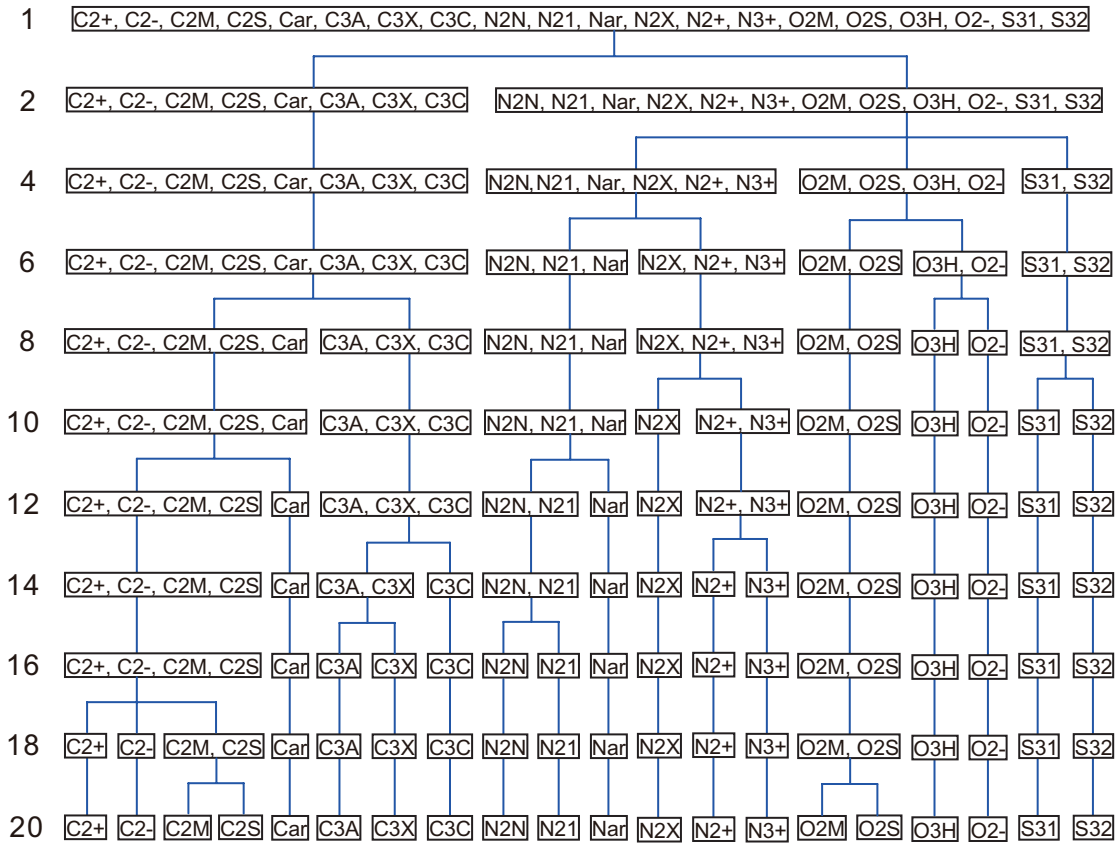


Fig. 1. (color online) Eleven typing schemes of protein heavy atoms based on their physical and chemical environments. The numbers on the left stand for the numbers of grouped atom types. The definitions of the final 20 atom types are listed in Table 1.

Table 1. List of the 20 atom types used in the iterative method for the heavy atoms in 20 standard amino acids. The symbol "*" stands for any residue.

No.	Symbol	Description	Atom name
1	C2+	carbons bonded to a positively charged nitrogen	ARG_CZ
2	C2-	carbons bonded to a negatively charged nitrogen	*_C (on terminal residues), ASP_CG, GLU_CD
3	C2M	carbonyl carbons on the mainchain	*_C (not on terminal residues)
4	C2S	carbonyl carbons on sidechains	ASN_CG, GLN_CD
5	Car	aromatic carbons	HIS_CD2, HIS_CE1, HIS_CG, PHE_CD1, PHE_CD2, PHE_CE1, PHE_CE2, PHE_CG, PHE_CZ, TRP_CD1, TRP_CD2, TRP_CE2, TRP_CE3, TRP_CG, TRP_CH2, TRP_CZ2, TRP_CZ3, TYR_CD1, TYR_CD2, TYR_CE1, TYR_CE2, TYR_CG, TYR_CZ
6	C3C	aliphatic carbons bonded to carbons or hydrogens only	ALA_CB, ARG_CB, ARG_CG, ASN_CB, ASP_CB, GLN_CB, GLN_CG, GLU_CB, GLU_CG, HIS_CB, ILE_CB, ILE_CD1, ILE_CG1, ILE_CG2, LEU_CB, LEU_CD1, LEU_CD2, LEU_CG, LYS_CB, LYS_CD, LYS_CG, MET_CB, PHE_CB, PRO_CB, PRO_CG, SER_CB, THR_CG2, TRP_CB, TYR_CB, VAL_CB, VAL_CG1, VAL_CG2
7	C3A	C_α atoms	*_CA
8	C3X	aliphatic carbons on sidechains bonded to a polar atom	ARG_CD, CYS_CB, LYS_CE, MET_CE, MET_CG, PRO_CD, THR_CB
9	N2N	amide nitrogens with one hydrogen	*_N (not on terminal residues)
10	N2X	amide nitrogens with two hydrogens	*_N (on terminal residues), ASN_ND2, GLN_NE2
11	Nar	aromatic nitrogens	HIS_ND1, HIS_NE2, TRP_NE1
12	N2+	guanidine nitrogens with two hydrogens	ARG_NH1, ARG_NH2
13	N21	guanidine nitrogen with one hydrogen	ARG_NE
14	N3+	nitrogen with three hydrogens	LYS_NZ
15	O2M	oxygens in carbonyl groups of the mainchain	*_O (not on terminal residues)
16	O2S	oxygens in carbonyl groups of sidechains	ASN_OD1, GLN_OE1
17	O3H	oxgens in hydroxyl groups	SER_OG, THR_OG1, TYR_OH
18	O2-	oxgens in carboxyl groups	*_O or *_OXT (on terminal residues), ASP_OD1, ASP_OD2, GLU_OE1, GLU_OE2
19	S31	sulfur with one hydrogen	CYS_SG
20	S32	sulfur without hydrogens	MET_SD

respectively. During the classification of the protein atoms, one important aspect is to consider the effects of physical and chemical environments like van der Waals interactions, electrostatics, hydrophobicity, entropy effect, and hydrogen bonding, but the ultimate goal of atom typing is to develop an accurate scoring function based on the atom typing scheme.[25,26,35,41–43] Therefore, we have derived the corresponding scoring functions of pair potentials based on different atom typing schemes through a statistical mechanics-based iterative method.[35]

2.2. The iterative method to extract effective potentials

During our computations, a large training set of experimentally determined protein structures and computationally generated decoys was used to derive the scoring function of pair potentials.[40] Through a statistical mechanics-based iterative method, all the native structures are expected to have the lowest energy scores compared with their respective decoys. The basic idea of the iteration process is described by the following expressions:[35]

$$u_{ij}^{k+1}(r) = u_{ij}^{k}(r) + \Delta u_{ij}^{k}(r), \tag{1}$$

$$\Delta u_{ij}^{k}(r) = \lambda k_{\mathrm{B}} T [\ln g_{ij}^{k}(r) - \ln g_{ij}^{\mathrm{obs}}(r)], \tag{2}$$

where i and j represent the types of a protein atom pair, $u_{ij}^{k}(r)$ is the statistical interaction potential of atom pair ij at a distance of r in the k-th iterative step, $u_{ij}^{k+1}(r)$ is the improved potential through iteration in the $(k+1)$-th step, λ represents a convergence parameter with $0 < \lambda \leqslant 1$, k_{B} is the Boltzmann constant, and T is the absolute temperature. Without loss of generality, $k_{\mathrm{B}}T$ is set to unit one in the present study. $g_{ij}^{\mathrm{obs}}(r)$ is the pair distribution function for the experimentally observed (i.e. native) structures. $g_{ij}^{k}(r)$ is the weighted average of the predicted pair distribution functions for the k-th step based on a training set of experimentally determined native protein structures and computationally generated decoys

according to the Boltzmann's probability distribution. To reduce the local correlation effects of the covalent bonds and obtain an unbiased scoring function, only the interatomic interactions between non-neighboring residues were considered in the calculation of pair distribution functions and interaction potentials. During the iteration, the standard potentials of mean force (PMF)[35,44] were used as the initial potentials,

$$u_{ij}^{0}(r) = -k_{\mathrm{B}}T \ln g_{ij}^{\mathrm{obs}}(r). \tag{3}$$

The statistical energy score of each native structure or non-native/decoy structure at the k-th iterative step can be calculated as follows:

$$E^{k}(S_{pq}) = \sum_{i=1}^{N}\sum_{j=i}^{N}\int_{0}^{r_{\mathrm{cut}}} n_{pq}^{ij}(r)u_{ij}^{k}(r)\,\mathrm{d}r, \quad p \in [1,P], \quad q \in [0,D_p], \tag{4}$$

where P stands for the number of proteins in the training set, D_p stands for the number of decoys for the p-th protein, and S_{pq} stands for the q-th decoy of the p-th protein, with $p \in [1,P]$, $q \in [0,D_p]$, and the native structures correspond to $q = 0$. $n_{pq}^{ij}(r)$ is the number density of atom pair ij at distance r for the structure of S_{pq}. As the size of each protein structure is finite, there exists a distance cutoff r_{cut} to consider the pairs between atoms, which was set to be 12 Å in this work. N is the number of protein atom types for a typing scheme (Fig. 1). The iteration is repeated until a maximum number of native structures are discriminated from decoys

$$E^{k}(S_{p0}) < E^{k}(S_{pq}), \quad \forall p \in [1,P], \quad \forall q \in [1,D_p]. \tag{5}$$

Thus, at the end of the iteration, we can obtain a set of knowledge-based pair potentials $u_{ij}(r) = u_{ij}^{k}(r)$ for each atom typing scheme. The details on the pair distribution functions and the statistical mechanics theory of the iteration method have been described in our previous study.[35]

The iterative method is robust and has been widely used in studying protein–protein interactions, protein–ligand interactions, protein–RNA/DNA interactions, and protein structure prediction.[32–35,42,45] The iterative method circumvents the long-standing reference state problem in the development of knowledge-based scoring functions by improving the pair potentials iteratively through comparison of the physics-based pair distribution functions. The scoring function ITScorePP, that was derived by the similar iterative method, has proven to be effective in past critical assessment of prediction of interactions (CAPRI) experiments.[46,47]

2.3. Training and test data sets

In the present work, we have used the training set of 1225 proteins and the test set of 148 proteins that were generated by Rajgaria *et al.*[40] All the proteins, that were selected by Zhang and Skolnick, are nonredundant single domain proteins with a maximum pairwise sequence similarity of 35%.[48] The length of these proteins ranges from 41 to 200 amino acids. This set also has a uniformly distribution of α, β, and α/β protein structures.

Out of the 1225 proteins, we have randomly chosen 500 proteins as the training set for our iterative computation. The similar iteration was run 10 times, and the final potentials were the average values over all the resulting iterative potentials of 10 runs. For the sake of computational efficiency, we have selected 400 decoys for each protein in the iteration. As our main purpose is to obtain the relationship between the effectiveness of the statistical interaction potentials and the dimension of the atom typing schemes, we have focused on the comparison between the effectiveness of different atom typing schemes in protein structure prediction so as to obtain an optimal atom typing scheme. Therefore, as long as the statistics for atom pairs are sufficient, the quantity of proteins in the training set and the number of decoys for each protein are relatively less relevant factors in our calculations. Given the interaction pair potentials for different atom typing schemes, we have closely examined the potential energy curves of several typical atom pairs with different physicochemical properties. All the derived scoring functions of potentials were also evaluated in terms of native structure recognition and five other parameters on the test set of 148 proteins with 500–1600 HR decoys for each protein.

Considering the set of randomly selected 500 nonhomologous proteins, our large training database results in significant statistics of frequencies > 1000 for 209 of the 210 possible pairs in the case of 20 atom types. For example, the frequency is 1454923 for C3C–C3C pair. It is expected that the average frequencies will be higher when fewer atom types are adopted. The high frequencies of most atom pairs occurring in the training set warrant sufficient statistics to derive the distance-dependent pair potentials. Therefore, the ensemble is large enough in terms of atom pair frequencies for statistical study.

3. Results and discussion

3.1. The potentials for different atom typing schemes

Through a statistical mechanics-based iterative method, we have developed 11 sets of atomic distance-dependent pair potentials corresponding to 11 atom typing schemes. To illustrate how the atom typing schemes impact the derived statistical interaction potentials and obtain an overall physical picture of the atomic interactions, we have plot the potential energy curves of several typical atom pairs with different physicochemical properties when the protein atoms were categorized into one to 20 atom types, respectively. Similar to our previous study,[35] we have chosen several representative atom pairs

for demonstration in the present work. They stand for electrostatic interactions [O2-]–[N2+] (e.g., ASP_OD2–ARG_NH1) and [O2-]–[N3+] (e.g., ASP_OD2–LYS_NZ) (Figs. 2 and 3), hydrogen bonding interactions [O2M]–[N2N] (e.g., *_O–*_N) and [O3H]–[O3H] (e.g., SER_OG–TYR_OH) (Figs. 4 and 5), and hydrophobic interactions [Car]–[Car] (e.g., PHE_CZ–TRP_CH2) and [C3C]–[C3C] (e.g., ILE_CD1–LEU_CD2) (Figs. 6 and 7), respectively.

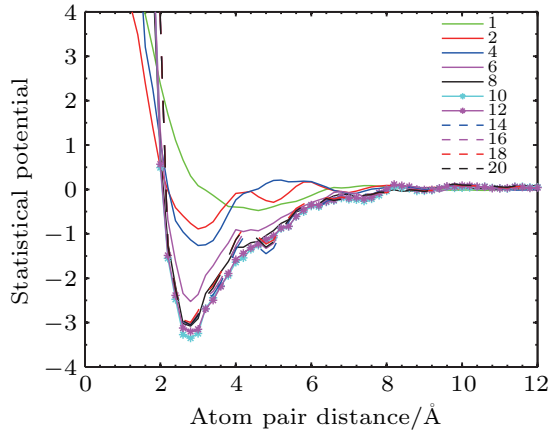


Fig. 2. (color online) Comparison of the derived pair potentials of [O2–]–[N2+] for different atom typing schemes with 1 to 20 atom types.

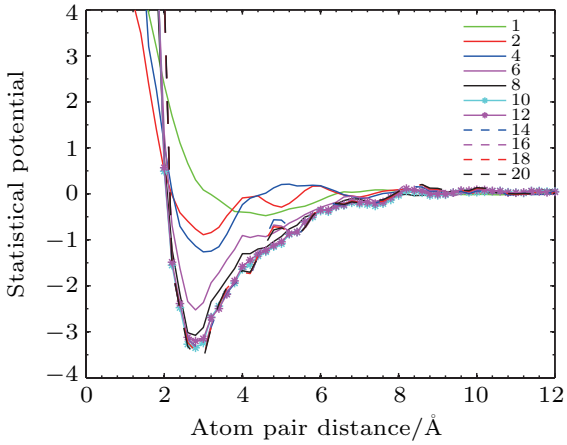


Fig. 3. (color online) Comparison of the derived pair potentials of [O2–]–[N3+] for different atom typing schemes with 1 to 20 atom types.

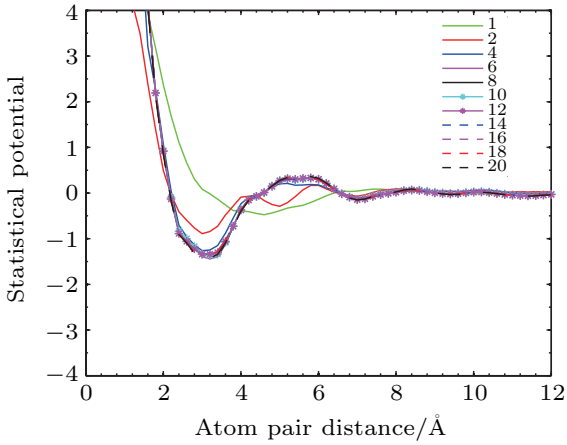


Fig. 4. (color online) Comparison of the derived pair potentials of [O2M]–[N2N] for different atom typing schemes with 1 to 20 atom types.

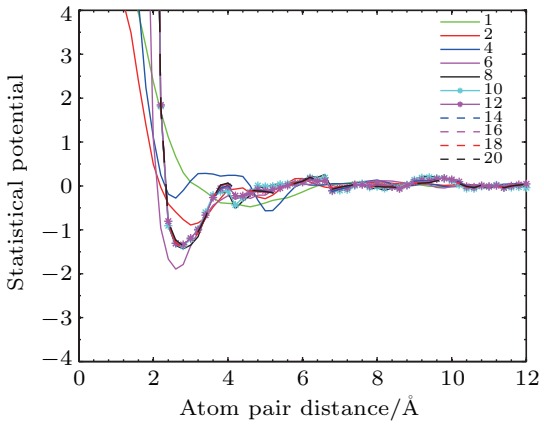


Fig. 5. (color online) The comparison of the derived pair potentials of [O3H]–[O3H] for different atom typing schemes with 1 to 20 atom types.

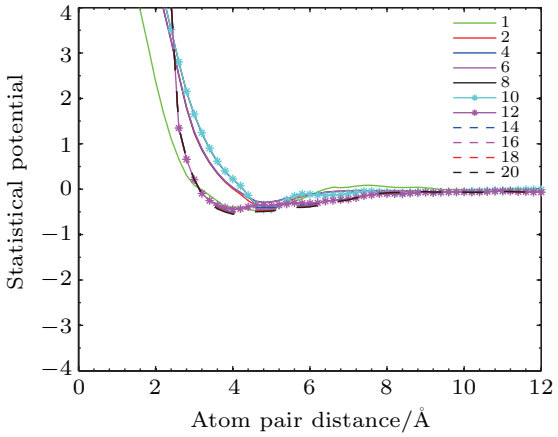


Fig. 6. (color online) The comparison of the derived pair potentials of [Car]–[Car] for different atom typing schemes with 1 to 20 atom types.

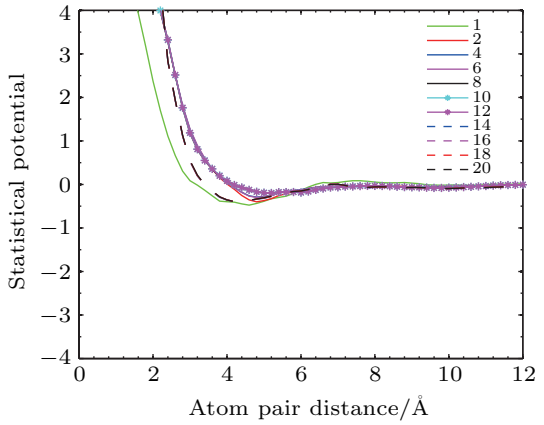


Fig. 7. (color online) The comparison of the derived pair potentials of [C3C]–[C3C] for different atom typing schemes with 1 to 20 atom types.

Several common trends can be observed from the pairs of knowledge-based potentials under different atom typing schemes. First, with the increasing number of classified atom types, the equilibrium positions of the potentials tend to move left. Second, the depths of the potential wells become deeper with more atom types classified. For clarity, the equilibrium positions and depths of the potential wells of all typical atom pairs are listed in Table 2, and illustrated in Figs. 8 and 9.

It can be seen from the figures that the equilibrium positions and well depths become relatively stable at the alphabet of 14 atom types. Third, the pair potentials gradually converge into a stable curve, though the converging speed depends on the specific atom pairs. For electrostatic interactions, the derived potentials become stable when the protein atoms are grouped into eight or more atom types (Figs. 2 and 3). For hydrogen bond interactions, the pair potentials converge at the scheme of four atom types, and become consistent for the schemes of eight or more atom types (Figs. 4 and 5). For hydrophobic interactions, the potential curves start to merge together when 14 or more atom types are used (Figs. 6 and 7). Therefore, when considering all the atom pair potentials of different atom typing schemes, categorizing protein heavy atoms into 14 atom types seems to be the optimal choice in terms of atomic pair potentials.

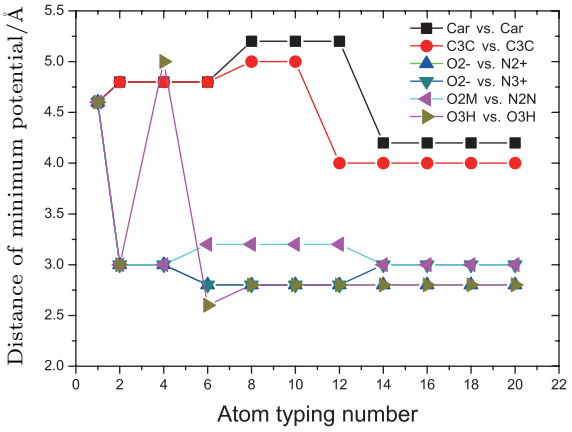


Fig. 8. (color online) The equilibrium positions of six typical pair potentials under different atom typing schemes.

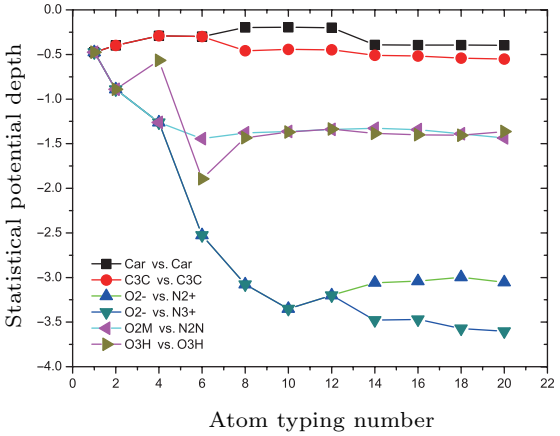


Fig. 9. (color online) The well depths of six typical pair potentials under different atom typing schemes.

Table 2. The equilibrium positions and well depths of several typical pair potentials for 1–20 atom tying schemes.

Scheme	[Car]–[Car]	[C3C]–[C3C]	[O2–]–[N2+]	[O2–]–[N3+]	[O2M]–[N2N]	[O3H]–[O3H]
1	(4.6, –0.475)	(4.6, –0.475)	(4.6, –0.475)	(4.6, –0.475)	(4.6, –0.475)	(4.6, –0.475)
2	(4.8, –0.399)	(4.8, –0.399)	(3.0, –0.891)	(3.0, –0.891)	(3.0, –0.891)	(3.0, –0.891)
4	(4.8, –0.290)	(4.8, –0.290)	(3.0, –1.263)	(3.0, –1.263)	(3.0, –1.263)	(5.0, –0.566)
6	(4.8, –0.299)	(4.8, –0.299)	(2.8, –2.526)	(2.8, –2.526)	(3.2, –1.444)	(2.6, –1.895)
8	(5.2, –0.197)	(5.0, –0.457)	(2.8, –3.078)	(2.8, –3.078)	(3.2, –1.379)	(2.8, –1.433)
10	(5.2, –0.195)	(5.0, –0.442)	(2.8, –3.350)	(2.8, –3.350)	(3.2, –1.363)	(2.8, –1.371)
12	(5.2, –0.200)	(4.0, –0.447)	(2.8, –3.200)	(2.8, –3.200)	(3.2, –1.340)	(2.8, -1.337)
14	(4.2, –0.393)	(4.0, –0.508)	(2.8, –3.058)	(3.0, –3.477)	(3.0, –1.327)	(2.8, –1.385)
16	(4.2, –0.394)	(4.0, –0.518)	(2.8, –3.041)	(3.0, -3.470)	(3.0, –1.343)	(2.8, –1.399)
18	(4.2, –0.394)	(4.0, –0.542)	(2.8, –2.996)	(3.0, –3.571)	(3.0, –1.389)	(2.8, –1.406)
20	(4.2, –0.398)	(4.0, –0.552)	(2.8, –3.054)	(3.0, –3.603)	(3.0, –1.436)	(2.8, –1.365)

3.2. Performances in discriminating native structures

Atom typing is a critical aspect in the development of knowledge-based scoring functions. Therefore, an important criterion for the goodness of an atom typing scheme is the ability of the scoring function with the atom typing scheme in discriminating native structures from decoys. Therefore, we have evaluated the performances of the 11 scoring functions derived from the 11 atom typing schemes through a statistical mechanics-based iteration method on the high-resolution (HR) decoy set of 148 proteins generated by Rajgaria *et al.*[40] Table 3 lists the six assessment parameters for different atom typing schemes, in which the success rate of identifying native structures, the correlations between the energy scores and the RMSDs of decoys, and the *Z*-score of native structures are show in Figs. 10–12, respectively. Here, the success rate is defined as the number of the proteins with native structures as rank #1 divided by the total number of the proteins in the test

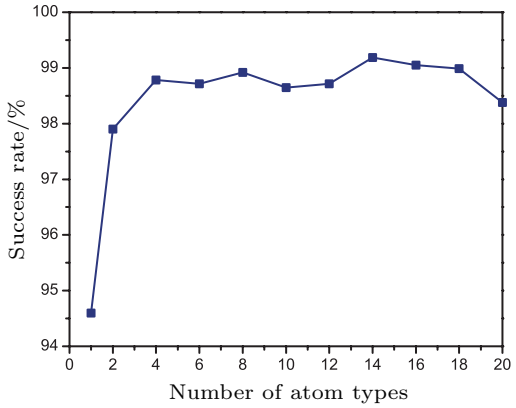


Fig. 10. (color online) Success rates of the proteins with native structures as rank 1 for 11 atom typing schemes with 1 to 20 atom types.

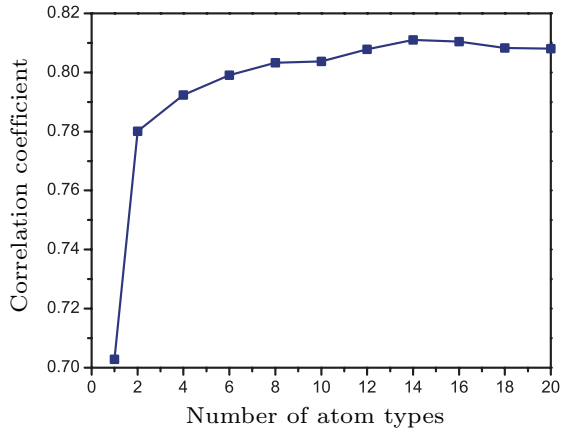


Fig. 11. (color online) The average Pearson correlation coefficients (CC) between the energy scores and the RMSD of decoys for 11 atom typing schemes with 1 to 20 atom types.

set. The Z-score is defined as

$$Z\text{-score} = (E_{\text{average}} - E_{\text{native}})/\sigma, \tag{6}$$

where E_{average} is the average energy score of all the decoys for a protein, E_{native} is the energy score of the native structure, and σ is the energy deviation of all the decoys. The Z-score measures the relative energy separation between the native structure of a protein and its decoys.

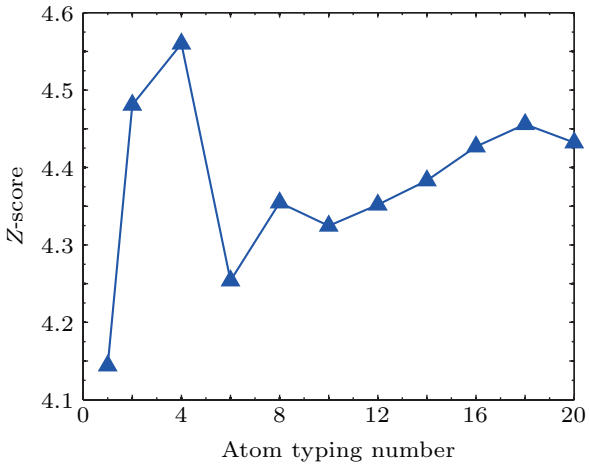


Fig. 12. (color online) The average Z-score for native structures for 11 atom typing schemes with 1 to 20 atom types.

Table 3. Testing results for 11 atom typing schemes on HR decoy sets.[40]

Scheme	Avg rank[a]	No of firsts[b]	Avg Z-score[c]	Avg CC[d]	Avg Cα-RMSD[e]	Avg Cα-RMSD[f]
1	1.644	94.60	4.144	0.703	0.269	2.365
2	1.090	97.90	4.481	0.780	0.058	1.926
4	1.017	98.78	4.560	0.792	0.033	1.729
6	1.017	98.71	4.254	0.799	0.034	1.743
8	1.023	98.92	4.355	0.803	0.027	1.682
10	1.028	98.65	4.325	0.804	0.033	1.658
12	1.039	98.72	4.352	0.808	0.030	1.644
14	1.024	99.19	4.383	0.811	0.015	1.628
16	1.028	99.05	4.427	0.810	0.018	1.618
18	1.045	98.98	4.456	0.808	0.019	1.630
20	1.072	98.38	4.432	0.808	0.033	1.625

[a]The average (Avg) rank of the native structures. The best rank is 1.
[b]The number of the proteins with native structures as rank #1 in terms of the calculated energy scores.
[c]The average Z-score of natives, measuring the relative energetic separation of the native structure of a protein with respect to its decoys.
[d]The average Pearson correlation coefficients (CC) between the energy scores and the RMSD of decoys.
[e]The average RMSD of the best predicted structures with the lowest energy scores (native structures included).
[f]The average RMSD of the best predicted structures with the lowest energy scores with native structures excluded.

One common feature can be found from five assessment parameters (except Z-score) as a function of the number of atom types. Namely, the values of the five parameters change fast at the beginning and then become relatively stable with the increasing number of atom types (Table 3). There is a crossover between fast and slow changes at the scheme of four atom types (Figs. 10 and 11). For example, there is a maximum at the scheme of four atom types for the Z-score of native structures, before which the Z-score increases fast and after which the Z-score drops again (Fig. 12). Therefore, the scheme of four atom types seems to give the best balance between the accuracy and resolution in terms of the success rate of a scoring function in discriminating native structures. It can also be seen from Fig. 10 that the scheme of 14 atom types gives a slightly higher success rate than the other schemes. A similar trend was also found in the work by Mintseris *et al.*[28] Namely, the highest peak of effective mutual information (MI) occurs at alphabet of size four, corresponding to hydrophobic, polar, positive, and negative types, and the effectiveness of all pair-wise potentials flattens out at alphabet of size 12.[28]

From the above assessments over different atom typing schemes, we have found that the scheme of four atom types did the best in low dimensions while the scheme of 14 atom types obtained the best performance in high dimensions. The four atom types correspond to the natural atom typing scheme of C, N, O, and S. The scheme can give excellent average Z-score and rank, and the other parameters are not far from the maximum values in the high dimensions. Therefore, we may

use the scheme of four atom types as rough screening (low precision) for protein structure prediction, as the simple scheme can greatly cut down the computational cost. When accurate prediction is needed, we can turn to the scheme of 14 atom types to evaluate structures in high resolution.

4. Conclusion

In this work, we have addressed the protein atom typing problem by categorizing protein heavy atoms into 11 atom typing schemes. The knowledge-based pair potentials for the 11 atom typing schemes were derived using a statistical mechanics-based iterative method. The performances of different atom typing schemes were evaluated by the comparison of the derived knowledge-based pair potentials and the ability of the corresponding scoring functions in discriminating native structures from decoys. It was found that the derived pair potentials started to converge when 14 or more atom types were used, while four types were enough to obtain a satisfactory success rate in native structure recognition. The results suggested that the number of atom types could range from 4 to 14 in practical applications depending on the studied systems. The scheme of four atom types (i.e., C, N, O, and S) could be used for assessing the overall quality of protein structures, while an accurate description of interactions for protein structures at atomic level may require a finer scheme of 14 atom types. The present study provides a basic guidance for the classification of protein atoms, and is expected to benefit the development of scoring functions and the understanding of interaction mechanisms in proteins.

References

[1] Wang J and Wang W 1999 *Nat. Struct. Mol. Biol.* **6** 1033
[2] Wang J and Wang W 2000 *Phys. Rev. E* **61** 6981
[3] Fan K and Wang W 2003 *J. Mol. Biol.* **328** 921
[4] Dill K A 1985 *Biochemistry* **24** 1501
[5] Riddle D S, Santiago J V, Bray-Hall S T, Doshi N, Grantcharova V P, Yi Q and Baker D 1997 *Nat. Struct. Biol.* **4** 805
[6] Launay G, Mendez R, Wodak S and Simonson T 2007 *BMC Bioinformatics* **8** 270
[7] Luthra A, Jha A N, Ananthasuresh G K and Vishveswara S 2007 *J. Biosci.* **32** 883
[8] Li T, Fan K, Wang J and Wang W 2003 *Protein Eng.* **16** 323
[9] Walter K U, Vamvaca K and Hilvert D 2005 *J. Biol. Chem.* **280** 37742
[10] Akanuma S, Kigawa T and Yokoyama S 2002 *Proc. Natl. Acad. Sci. USA* **99** 13549
[11] Peterson E L, Kondev J, Theriot J A and Phillips R 2009 *Bioinformatics* **25** 1356
[12] Albayrak A, Out H H and Sezerman U O 2010 *BMC Bioinformatics* **11** 428
[13] Etchebest C, Benros C, Bornot A, Camproux A C and de Brevern A G 2007 *Eur. Biophys. J.* **36** 1059
[14] Melo F and Marti-Renom M A 2006 *Proteins* **63** 986
[15] Cannata N, Toppo S, Romualdi C and Valle G 2002 *Bioinformatics* **18** 1102
[16] Shi Y Z, Wu Y Y, Wang F H and Tan Z J 2015 *Chin. Phys. B* **24** 116802
[17] Zhang W, Sun Z B and Zou X W 2005 *Chin. Phys. Lett.* **22** 2133
[18] Li W F, Zhang J, Wang J and Wang W 2015 *Acta Phys. Sin.* **64** 098701 (in Chinese)
[19] Solis A D 2015 *Proteins* **83** 2198
[20] Huang J T, Wang T, Huang S R and Li X 2015 *Proteins* **83** 631
[21] Levitt M and Warshel A 1975 *Nature* **253** 694
[22] Wang J and Wang W 2002 *Phys. Rev. E* **65** 041911
[23] Chen H, Zhou X and Ou-Yang Z C 2002 *Phys. Rev. E* **65** 061907
[24] Liu Y, Zeng J and Gong H 2014 *Proteins* **82** 2383
[25] Yang Y and Zhou Y 2008 *Protein Sci.* **17** 1212
[26] Yang Y and Zhou Y 2008 *Proteins* **72** 793
[27] Mintseris J and Weng Z 2004 *Genome Informatics* **15** 160
[28] Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R and Weng Z 2007 *Proteins* **69** 511
[29] Zhao Y, Cheng T and Wang R 2007 *J. Chem. Inf. Model.* **47** 1379
[30] Jiang L, Gao Y, Mao F, Liu Z and Lai L 2002 *Proteins* **46** 190
[31] Zhang M, Chen C, He Y and Xiao Y 2005 *Phys. Rev. E* **72** 051919
[32] Huang S Y and Zou X 2014 *Nucleic Acids Res.* **42** e55
[33] Huang S Y and Zou X 2014 *Proteins* **72** 557
[34] Huang S Y and Zou X 2006 *J. Comput. Chem.* **27** 1866
[35] Huang S Y and Zou X 2011 *Proteins* **79** 2648
[36] Zhou H and Zhou Y 2002 *Protein Sci.* **11** 2714
[37] Zhou H and Skolnick J 2011 *Biophys J.* **101** 2043
[38] Mitchell J, Alex A and Snarey M 1999 *J. Chem. Inf. Comput. Sci.* **39** 751
[39] Labute P 2005 *J. Chem. Inf. Model.* **45** 215
[40] Rajgaria R, McAllister S R and Floudas C A 2008 *Proteins* **70** 950
[41] Lu M, Dousis A D and Ma J 2008 *J. Mol. Biol.* **376** 288
[42] Huang S Y and Zou X 2010 *J. Chem. Inf. Model.* **50** 263
[43] Kadukova M and Grudinin S 2016 *J. Chem. Inf. Model.* **56** 1410
[44] Xu W X, Li Y and Zhang J Z 2012 *Chin. Phys. Lett.* **29** 068702
[45] Yan Y M, Zhang D, Zhou P, Li B T and Huang S Y 2017 *Nucleic Acids Res.* **45** w365
[46] Huang S Y 2014 *Drug Discov. Today* **19** 1081
[47] Lensink M F, Velankar S, Kryshtafovych A, Huang S Y, *et al.* 2016 *Proteins* **84** 323
[48] Zhang Y and Skolnick J 2004 *Proc. Natl. Acad. Sci. USA* **101** 7594