Published in final edited form as:
J Comput Chem. 2011 June ; 32(8): 1680–1686. doi:10.1002/jcc.21747.

Protein Side Chain Modeling with Orientation Dependent Atomic Force Fields Derived by Series Expansions

Shide Liang[1,*], **Yaoqi Zhou**[2], **Nick Grishin**[3], and **Daron M. Standley**[1]

[1] Systems Immunology Lab, Immunology Frontier Research Center, Osaka University, Suita, Osaka, 565-0871, Japan

[2] Indiana University School of Informatics, Indiana University-Purdue University, Indianapolis, Indiana 46202, USA; Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA

[3] Howard Hughes Medical Institute and Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA

Abstract

We describe the development of new force fields for protein side chain modeling called OSCAR (Optimized Side Chain Atomic eneRgy). The distance-dependent energy functions (OSCAR-d) and side-chain dihedral angle potential energy functions were represented as power and Fourier series, respectively. The resulting 802 adjustable parameters were optimized by discriminating the native side chain conformations from non-native conformations, using a training set of 12000 side-chains for each residue type. In the course of optimization, for every residue, its side chain was replaced by varying rotamers, whereas conformations for all other residues were kept as they appeared in the crystal structure. Then the OSCAR-d were multiplied by an orientation dependent function to yield OSCAR-o. 1087 parameters of the orientation-dependent energy functions (OSCAR-o) were optimized by maximizing the energy gap between the native conformation and subrotamers calculated as low energy by OSCAR-d. When OSCAR-o with optimized parameters were used to model side chain conformations simultaneously for 218 recently released protein structures, the prediction accuracies were 88.8% for χ_1, 79.7% for χ_{1+2}, 1.24 Å overall RMSD (root mean square deviation), and 0.62 Å RMSD for core residues, respectively, compared with the next-best performing side-chain modeling program which achieved 86.6% for χ_1, 75.7% for χ_{1+2}, 1.40 Å overall RMSD, and 0.86 Å RMSD for core residues, respectively. The continuous energy functions obtained in this study are suitable for gradient-based optimization techniques for protein structure refinement. A program with built-in OSCAR for protein side chain prediction is available for download at http://sysimm.ifrec.osaka-u.ac.jp/OSCAR/.

Introduction

Tertiary structural information is critical for our understanding of a protein's biological function. However, experimental structure determination is far too expensive and time consuming to be applied to all proteins of interest. Computational approaches are thus expected to play a major role in determining protein structures in the future.[1] Over the last two decades, great strides have been made in exploiting distant evolutionary relationships to known structures in order to derive spatial restraints for comparative models.[2–4] One of the

[*] To whom correspondence should be addressed. shideliang@IFReC.osaka-u.ac.jp, Tel: +81-6-6879-9490, Fax: +81-6-6879-4272.

Supporting information available: Detailed definition of 16 atoms types for 20 amino acids.

remaining major challenges is in the refinement of such models to near experimental accuracy.[5] This challenge, in turn, demands the development of more accurate force fields that can be deployed in molecular mechanics simulations.

Physiochemical force fields such as CHARMM,[6] AMBER,[7] and GROMOS,[8] parameterized for use in protein simulations, are routinely applied to the refinement of comparative models. However, overall improvement in the accuracy of comparative models by such methods has not been achieved.[5] Knowledge-based potential energy functions are derived from either statistical analysis of observed protein structures[9–16] or optimization of parameters such that native structures are discriminated from non-native decoys.[17–20] They usually outperform[9,21] physiochemical force fields that lack some physical terms such as cation-π interactions and entropic effects. However, the discrete nature of statistical energy functions makes it difficult to be employed directly in energy minimization or molecular dynamics for protein-structure refinement. Moreover, most knowledge-based energy functions derived from parameter optimization are coarse-grained (i.e., at the residue-level or using simplified side chains) in order to minimize the number of adjustable parameters. Parameter optimization was considered inappropriate to derive distance-dependent energy functions of all atom types,[13] not to mention orientation dependence. Thus, it is more practical to optimize a small number of weights for mixing physiochemical terms with statistics-based potentials.[22,23] As more and more experimental protein structures become available, knowledge-based potential energy functions derived from parameter optimization may prove optimal even for all-atom force fields.

Any complicated function, including the force fields between atoms in a protein, can be decomposed as a mathematical series. For example, power-series expansions of a diatomic potential energy function are the most useful means for its analytical representation in quantum chemistry.[24] Miyazawa and Jernigan employed series expansions of spherical harmonic functions to represent the fully anisotropic distribution of the relative orientation of two residues and increased the discrimination power in fold recognition.[25] Here, we expanded atomic force fields as series. The parameters were optimized by maximizing the gap between native and non-native side-chain conformations and by minimizing the root mean square deviation (RMSD) of low-energy rotamers. A total of 5798 non-homologous proteins were used for optimizing 1889 parameters. The energy functions with optimized parameters were used to predict side chain conformations for 218 independent test proteins. The prediction accuracies of χ_1 and χ_{1+2} were improved by 2.2 and 4.0%, respectively, compared to the next best side chain modeling program. Since the expansions used here are continuous, the resulting energy functions can be used directly in gradient-based search algorithms to address the comparative model refinement problem.

Methods

Training and test sets

30 non-homologous proteins are used as the first test set, as described previously.[23] The training proteins were chosen according to the following criteria: the sequence identity between any two pairs was less than 30%, the resolution was less than 2.5 Å, and the R-factor was less than 1.0. A total of 6254 chains that met the above criteria were downloaded from the Dunbrack Lab website http://dunbrack.fccc.edu/PISCES.php in Jun, 2008. A protein was discarded if more than 5% of its residues had incomplete side chain atomic coordinates or the sequence identity with any of the 30 test proteins was more than 50% following local alignment. As a result, the training set contains 5798 proteins. We also compiled a second test set. A total of 5279 chains with sequence identity less than 30%, resolution less than 2.0 Å, and R-factor less than 0.25 were downloaded from the Dunbrack Lab website in Jun 2009. Those proteins were discarded if they met any of the following

conditions: the same PDB ID as the 6254 protein chains downloaded in Jun 2008, more than 5% residues with incomplete side chain atomic coordinates, less than 100 residues(excluding Gly, Ala, and incomplete side chains) for the prediction accuracy assessment, or a sequence identity of more than 50% with any of the other training proteins following local alignment. This leads to the second test set of 218 proteins. Hydrogen atoms were added with the REDUCE program for all protein structures.[26]

Rotamer library

The rotamer library is from Dunbrack and Cohen.[27] We generate sub-rotamers by giving a perturbation to each dihedral angle of the rotamer. $(f_1+f_2+f_3+f_4+f_5)\times\sigma$ is added to the original dihedral angle. Here f_i is a generated random number in the range of (−1,1) and σ is the standard deviation of the dihedral angle included in the library. Bond lengths and angles from Engh and Huber[28] are used to build the rotamer library. Polar hydrogen atoms are added since they are absent in the Dunbrack library and considered explicitly in this study. Each χ_2 for Ser and Thr and χ_3 for Tyr(θ) are assigned three possible values: −60°, 60°, and 180°. The dihedral angle varies from $\theta - 30°$ to $\theta + 30°$ with even distribution for the subrotamers.

Rotamer internal energy

$$E_{\text{torsion}}=t_1\times\cos\alpha+t_2\times\sin\alpha+t_3\times\cos2\alpha+t_4\times\sin2\alpha+t_5\times\cos3\alpha+t_6\times\sin3\alpha \quad (1)$$

where α is a dihedral angle of the side chain rotamer and t_{1-6} are optimized parameters. 258 parameters are used for the 43 dihedral angles of the 20 amino acids. The rotamer internal energy is summarized over all dihedral angles of the modeled side chain. The interactions between bonded atoms are not calculated. Atomic interaction energy beyond 1,4 interactions are calculated as for typical non-bonded atoms.

Distance dependent energy function

The distance-dependent optimized side-chain atomic energy (OSCAR-d) is calculated by

$$E_{(\text{d})}=a_1\times d^{-2}+a_2\times d^{-4}+a_3\times d^{-6}+a_4\times d^{-8} \quad (2)$$

where d is the distance between two atoms and a_{1-4} are optimized parameters. We define 16 atom types for 20 amino acids and employ a total 544 parameters. Atoms with a similar charge and radius according to CHARMM are defined as the same type. The distance cutoff is set to 10 Å for any two interacting atoms. The definition of the atom types can be found in the supporting information.

Orientation dependent energy function

The orientation-dependent optimized side-chain atomic energy (OSCAR-o) is calculated by

$$E=(E_{(\theta,\phi,\psi)}+C)\times E_{(\text{d})} \quad (3)$$

where $E_{(\theta,\phi,\Psi)}$ is an orientation dependent function and C is a constant. θ, ϕ, and Ψ are Euler angles of two interacting dipoles(Fig 1). The dipole points to the interacting atom from the center of its base atoms. $E_{(\theta,\phi,\Psi)}$ is given by

$$E_{(\theta,\phi,\psi)}=b_1\times\cos^2\theta+b_2\times\cos^2\phi+b_3\times\cos^2\psi+b_4\times\cos\theta\cos\phi+b_5\times\cos\theta\cos\psi+b_6\times\cos\phi\cos\psi+b_7\times\cos\theta+b_8\times\cos\phi+b_9\times\cos\psi \quad (4)$$

There are 1087 parameters including b_{1-9} in eq (4) and the constant C in eq (3) optimized for 16 atom types(for sp2 hybridized atoms, the dipole is perpendicular to the hybridization plane and the parameters for the related one order terms are set to 0 so that the calculated energy is not affected by inversion of the dipole direction). Here, we assume the interaction energy is comprised of a distance dependent term and an orientation dependent term. In extreme case when $E_{(\theta,\phi,\Psi)}$ equals to 0 and C equals to 1, Eq (3) becomes a distance dependent energy function only. We optimized the parameters(b_{1-9} and C) simultaneously so that the interaction energy could be correctly calculated even if the distance dependent term and the orientation dependent terms overlap somewhat.

Optimizing parameters for the distance dependent energy functions and dihedral angle potential functions

The parameters are initialized with random values. The sum of Eq (1) and Eq (2) is used to calculate energies for the native side chain conformation and rotamers at a specific position. The side chain conformation of other residues is fixed at observed atomic coordinates. Energies for 12000 residues from the training proteins are calculated for each of the 18 residue types (excluding Gly and Ala). Residues from high-resolution proteins are used with a priority. Similar to our previous study,[29] Monte Carlo simulation annealing is used to optimize the parameters by minimizing the following objective function:

$$\sum_{k=1}^{M}\frac{\sum_{i=1}^{N}e^{-E(i)}}{M\times(N\times e^{-E(r)}+\sum_{i=1}^{N}e^{-E(i)})} \quad (5)$$

where N is the number of rotamers, $E(r)$ is the energy of the native conformation r, $E(i)$ is the energy of rotamer i, and M is the total number of calculated residues (18×12000=216000).

Optimizing parameters for orientation-dependent energy functions

Firstly, for each of the N backbone dependent rotamers at the modeled position, we generate 60 sub-rotamers and select the one with the lowest energy by the distance-dependent and rotamer internal energy functions. The parameters of the orientation-dependent functions are optimized so that the native conformation has a lower energy than the selected N sub-rotamers by minimizing eq (5). The optimized parameters of the distance-dependent functions are fixed during this procedure. The parameters of the rotamer internal energy functions are initialized to previously optimized values and then re-optimized. For the parameters of the orientation-dependent functions, b_{1-9} in eq (4) are initialized to 0 and C in eq (3) is initialized to 1.

In the next step, we increase the number of residues used in training up to 40000 for each residue type. For rare residue types such as Cys, Met, Trp, and His, less than 40000 residues are used. Instead of eq (5), which is continuous and easy to minimize, the rmsd value of the sub-rotamer with the lowest energy is averaged over all training residues and adopted as the objective function to minimize. For each modeled position, the lowest-energy sub-rotamer calculated by the distance-dependent functions and the rotamer internal energy functions are selected for 4 rotamers. Those with a relatively high energy are not used. Similarly, we select 4 lowest-energy sub-rotamers calculated by the orientation-dependent functions with parameters optimized in the first step. A total of 8 sub-rotamers are considered at each position. The parameters are initialized to the same value as optimized in the first step. After optimization, we employed the new parameters to select an additional 4 lowest-energy sub-rotamers and optimize the parameters with 12 sub-rotamers at each modeled position. This procedure is repeated 3 more times based on the observation that the results improve slightly with each iterative optimization.

Predicting side chain conformation of a single residue

We do not use any information of the native side chain conformation. To predict the side chain conformation, we generate 60 sub-rotamers for each rotamer and the sub-rotamer with the lowest energy among 60*N* ones constitutes the prediction.

Side chain modeling of the whole protein

We predict side chain conformations of entire proteins by combing a genetic algorithm with Monte Carlo(MC) simulation as follows: 1) generate a pool of 20 structures with the same native backbone structure but with randomly initialized side chain conformations; 2) exchange side chain conformations among those with lower energy values; 3) optimize side chain conformations for all of the 20 protein structures by the Monte Carlo method; and 4) repeat steps 2 and 3 for 30 cycles during which the MC simulation temperature decreases after every cycle. The energy values of final 20 structures are compared and the structure with the lowest energy is the predicted structure.

Evaluation

The methods for accuracy evaluation are similar to those described previously.[23] Residues with <20% solvent accessibility are considered as core residues. The χ_1 angle of a residue is correctly predicted if it is within 40° of the experimental value. The χ_{1+2} angle is correctly predicted when both χ_1 and χ_2 are within 40° of their experimental values. For residues with multiple side chain conformations in the observed structure, we compare withthe first conformation in the PDB file only; other conformations are not considered. Residues with incomplete side-chain atomic coordinates are modeled but not evaluated.

Results

Performance of distance-dependent energy functions and rotamer-dependent internal energy functions

We decomposed atomic distance-dependent energy functions and dihedral angle potentials as power and Fourier series, respectively. The parameters of the series were initially assigned as random values. The MC-optimized objective function (Eq. 5) converged to similar values (0.088~0.093) with different starting parameter values. As an example, Figure 2 shows the resulting optimized, distance-dependent component of the CH3-CH3 and H-H interaction energy functions. CH3 (the terminal methyl carbon) and H (polar hydrogen from non-charged residues) are 2 of the 16 defined atom types. The H-H interaction energy is unfavorable in the range of 0~10 Å while the CH3-CH3 interaction has an attractive well around 4 Å, similar to the Van der Waals energy. Table I tabulates the performance of the optimized energy functions for the training and 30-protein test sets. The accuracies of χ_1 and χ_{1+2} are 90.3 and 81.3%, respectively, for the 30 test proteins. The accuracy is much higher for core residues (96.4% for χ_1 and 92.5% for χ_{1+2}). For different residue types, the prediction accuracy varies significantly from the lowest (80.1% for Glu χ_1 and 68.1% for Glu χ_{1+2}) to the highest (100% for Phe χ_1 and 98.9% for Phe χ_{1+2}). For core residues, the prediction accuracy of the training proteins is slightly (approx 1% for χ_1 or χ_{1+2}) better than for the 30 test proteins. However, the accuracies are nearly identical for the two sets of proteins when both surface and core residues are included in the evaluation, implying that the training set is sufficiently large, in spite of the large number of adjustable parameters.

Performance of orientation-dependent energy functions

The distance dependent energy functions are multiplied by an orientation-dependent factor. The parameters of the orientation-dependent functions are optimized as described in Methods and the performance is shown in Table I. For the 30 test proteins, the prediction accuracies of χ_1 and χ_{1+2} are further improved by 1.7 and 3.0%, respectively, compared to the distance dependent energy functions. The improvement is mostly due to hydrophilic residues that are responsible for specific interactions. For example, the prediction accuracies of χ_1 and χ_{1+2} for Asp increased significantly from 88.7 and 76.6% to 94.5 and 88.4%, respectively. The χ_1 accuracy of Cys also improved from 96.2 to 98.1%, due to the orientation-dependent nature of the disulfide bridge. The prediction accuracy in χ_1 and χ_{1+2} of the training set is slightly higher than that of the test set (Table I), indicating some degree of over optimization

We compared our energy functions with the two most popular force fields, AMBER and CHARMM, in order to predict side chain conformations of a single residue. This procedure was already described by Wilson and co-workers[30] and Petrella and co-workers[31] to test AMBER and CHARMM, respectively. We employed the same protein as Wilson and co-workers (PDB identifier 2alp) and the 10 proteins of Petrella and co-workers, for comparison. As shown in Table II, our results on single side-chain prediction are significantly more accurate than those from the physics-based force fields despite the fact that Petrella and co-workers have used native bond lengths and angles in their predictions. Rather than using rotamers, Petrella and co-workers rotated χ_1 and χ_2 in native side-chains at intervals of 5° or 10°, which made the prediction easier.

Side-chain modeling for whole proteins

We compared side chain modeling programs with built-in distance-dependent force fields(OSCAR-d) and orientation-dependent force fields(OSCAR-o), respectively, to our previous side chain modeling program LGA[23] in Table III based on 30 test proteins. The χ_1 accuracy of OSCAR-d is only slightly higher (0.5%) than LGA while the χ_{1+2} accuracy is 3.2% higher. The OSCAR-o has the highest accuracy (89.4% for χ_1 and 80.8% for χ_{1+2}) among the three methods. The rmsd values of core residues in the current predictions are much smaller than those of LGA, in part due to the use of the sub-rotamer model.

Comparison to other algorithms

To provide an additional test for our methods, we collected 218 recently released non-homologous proteins (see Methods). Three side-chain modeling programs, NCN,[32] OPUS_Rota,[9] and LGA,[23] with top prediction accuracy ranked by Lu and co-workers[9] and the recently updated SCWRL4[33] were compared. The prediction accuracy of OSCAR-d is similar to other side chain modeling programs while significant improvement is achieved by OSCAR-o. The accuracies of χ_1 and χ_{1+2} are improved by 2.2 and 4.0%, respectively(Table IV), compared to the next-best side-chain modeling program, OPUS_Rota. The improvement for χ_1 is remarkable considering the moderate improvement (2.9%) that resulted from combining a knowledge-based term and multiple physics-based terms instead of a simple Van der Waals energy.[32] For the prediction of individual residue types, OSCAR-o has a higher accuracy than OPUS_Rota for χ_1 and χ_{1+2} in all cases except Pro (Fig 3). We also modeled side-chain conformations for additional 595 proteins. These proteins were selected according to the same criteria as the 218 test proteins, but with the maximum sequence identity to the training proteins in the range of 50.1% and 100.0%. The prediction accuracy (88.8% for χ_1 and 79.5% for χ_{1+2}), is almost the same as that for the 218 proteins, indicating that the results are not biased to proteins with a high sequence identity to the training proteins.

Discussion

We have expanded energy functions as mathematical series and improved prediction accuracy for side chain modeling. The energy functions were first expanded as distance dependent power-series and further enhanced by an orientation dependent factor. The large number of optimized parameters and training proteins is the main reason for its performance. In previous study,[23] we combined contact surface, volume overlap, backbone dependency, electrostatic interactions, and desolvation energy; the weights of energy terms were optimized in a manner similar to the current study with 15 training proteins. The prediction accuracy was not improved when more proteins were used for training. Here, by series expansion, more parameters are available for optimization, which makes it possible to improve the prediction accuracy by using a huge number of training proteins. The lowest-energy subrotamer model also plays a key role in optimizing the parameters for OSCAR-o. We used OSCAR-d and the rotamer internal energy functions with optimized parameters to select the subrotamer with the lowest energy out of 60 possible states for each rotamer at the modeled position. The parameters of OSCAR-o were optimized so that the native conformation had a lower energy than the selected lowest-energy subrotamers. If only one sub-rotamer was generated and employed, i.e. OSCAR-d were not used to select the lowest energy subrotamer, there was little improvement by OSCAR-o over OSCAR-d.

It takes approximately 15 CPU hours for OSCAR-o or NCN[32] to model the side chain conformations for the 30 test proteins. By comparison, OPUS_Rota[9] or SCWRL4[33] takes 5 CPU minutes only. Both OPUS_Rota and SCWRL4 use rigid rotamers. As a result, all rotamer-rotamer or rotamer-backbone interaction energies can be pre-calculated and employed directly in modeling the whole protein. Using flexible rotamer model prohibits OSCAR-o for employing the pre-calculated values. Nevertheless, our energy functions are continuous and fast to calculate. In our future studies, we will explore more efficient gradient-based search algorithms for modeling side chain conformations on a flexible backbone.

We also expanded the series with different orders for the distance dependent energy functions, $\Sigma a_i \cdot d^{-2i}$ (i=1,2,3…n). The accuracies of χ_1 in predicting single residues were 85.0%(n=2), 89.6%(n=3), 90.3%(n=4), 90.1%(n=5), and 90.3%(n=6), respectively, for the 30 test proteins. The prediction accuracy was not improved by a higher order expansion(n>4) and even lower(<85%) by expansions in a different manner(i=3,4 or i=1,4). The Fourier series to calculate rotamer internal energy is expanded up to the third order because the rotatable bonds are connected to at least one sp3 hybridized atom. The prediction accuracy of χ_1 decreased to 88.8% by one order expansion($t_1 \times \cos\alpha + t_2 \times \sin\alpha$). For the orientation dependent functions, we tried a simpler formula($b_1 \times \cos\theta + b_2 \times \cos\boldsymbol{\phi} + b_3 \times \cos\Psi + C$). The accuracy was the same as the eq (4) in predicting side chain conformations for single residues. However, when the simple formula was used to model side chain conformations for the whole protein, the accuracy was slightly decreased for the 218 test proteins(88.5% for χ_1 and 79.1% for χ_{1+2}). We preferred to using eq (4) in side chain conformation search because the running time was reduced only 10% by using the simpler formula. Nevertheless, it may be appropriate to use this simpler formula if the energy functions are used with gradient-based optimization methods in future.

Employing native conformations in generated sub-rotamers improves the prediction accuracy. For 30-test proteins, the prediction accuracies of χ_1 and χ_{1+2} for single-residue conformations by OSCAR-o increase by 0.9 and 1.7%, respectively. On the other hand, the prediction accuracy does not improve with more sampling (for example, generating 200 sub-rotamers instead of 60 for each rotamer and excluding the native conformation). This indicates that it is the energy function that limits the final accuracy.

In addition to energy functions and sampling techniques, other factors can affect the prediction accuracy. The accuracy of OPUS_Rota in table IV is lower than the reported values (89.0 for χ_1 and 79.1% for χ_{1+2}).[9] This is mainly due to different evaluation methods. For residues with multiple conformations, we only compared the predicted one with the first conformation in the PDB file. In Lu and co-worker's report, the predicted one was considered to be correct if it satisfied any of the alternative positions. Using the same evaluation method as this study, the prediction accuracy of OPUS_Roda decreased to 88.0% for χ_1 and 77.8% for χ_{1+2}, respectively, for their 65 test proteins. In addition, the prediction accuracy is protein dependent. For the same 65 test proteins, which also included our 30 test proteins, OSCAR-o achieved a relatively high accuracy (90.0% for χ_1 and 81.1% for χ_{1+2}). For the 43 small proteins, which were selected according to the same criteria as the 218 test proteins, but with less than 100 evaluated residues, the prediction accuracy of OSCAR-o is only 85.0% for χ_1 and 73.4% for χ_{1+2}. Nevertheless, the prediction accuracy for core residues is still high (96.0% for χ_1 and 92.0% for χ_{1+2}). The decreased overall prediction accuracy is mostly due to the relatively low percentage of core residues for these small proteins.

Conclusion

Protein tertiary structure prediction is now more important than ever. But the prediction accuracy is limited by the availability of high quality energy functions. A complicated function, such as that describing the forces between atoms in a protein, can be decomposed as a mathematical series even if we do not know the analytical form in detail. We derived orientation dependent knowledge-based atomic force fields (OSCAR-o) by series expansions. The solvation energy and entropic effect, which are of the most difficult items to calculate, are not considered explicitly because the energy functions, in principle, include all known or unknown interactions. The parameters were optimized by discriminating the native side chain conformations from non-native conformations. When OSCAR-o were used to predict side chain conformations for single residues, the prediction accuracy in χ_1 and χ_{1+2} was >5% higher than AMBER or CHARMM force fields. We also used OSCAR-o to model side chain conformations of entire proteins. For 218 independent test proteins, the prediction accuracy was significantly higher (2% for χ_1 and 4% for χ_{1+2}) than the next-best performing side chain modeling program. Since OSCAR-o are continuous, accurate, and fast to calculate, we expect a wide-range of applications in protein structure prediction and protein design.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

S.L. is grateful for the financial support from Institute for Protein Research, Osaka University and illuminating discussions with Prof. Haruki Nakamura(Osaka University). This project was partly supported by Natural Science

Foundation of China(Grant Number: 30400096). Y.Z. was supported by the National Institutes of Health (grants R01 GM 085003 and R01 GM 067168).

References

1. Baker D, Sali A. Science. 2001; 294(5540):93–96. [PubMed: 11588250]
2. Zhang Y. Curr Opin Struct Biol. 2008; 18(3):342–348. [PubMed: 18436442]
3. Dunbrack RL Jr. Curr Opin Struct Biol. 2006; 16(3):374–384. [PubMed: 16713709]
4. Qian B, Ortiz AR, Baker D. Proc Natl Acad Sci U S A. 2004; 101(43):15346–15351. [PubMed: 15492216]
5. MacCallum JL, Hua L, Schnieders MJ, Pande VS, Jacobson MP, Dill KA. Proteins. 2009; 77(Suppl 9):66–80. [PubMed: 19714776]
6. MacKerell AD, Bashford D, Bellott, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. The Journal ofPhysical Chemistry B. 1998; 102(18):3586–3616.
7. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. Journal of the American Chemical Society. 1996; 118(9):2309–2309.
8. Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, van Gunsteren WF. The Journal of Physical Chemistry A. 1999; 103(19):3596–3607.
9. Lu M, Dousis AD, Ma J. J Mol Biol. 2008; 376(1):288–301. [PubMed: 18177896]
10. Majek P, Elber R. Proteins. 2009; 76(4):822–836. [PubMed: 19291741]
11. Miyazawa S, Jernigan RL. Macromolecules. 1985; 18(3):534–552.
12. Rata IA, Li Y, Jakobsson E. J Phys Chem B. 114(5):1859–1869. [PubMed: 20070091]
13. Shen MY, Sali A. Protein Sci. 2006; 15(11):2507–2524. [PubMed: 17075131]
14. Zhou H, Zhou Y. Protein Sci. 2002; 11(11):2714–2726. [PubMed: 12381853]
15. Yang YD, Zhou YQ. Proteins. 2008; 72(2):793–803. [PubMed: 18260109]
16. Huang SY, Zou X. Proteins. 2008; 72(2):557–579. [PubMed: 18247354]
17. Chiu TL, Goldstein RA. Fold Des. 1998; 3(3):223–228. [PubMed: 9669880]
18. Thomas PD, Dill KA. Proc Natl Acad Sci U S A. 1996; 93(21):11628–11633. [PubMed: 8876187]
19. Bastolla U, Vendruscolo M, Knapp EW. Proc Natl Acad Sci U S A. 2000; 97(8):3977–3981. [PubMed: 10760269]
20. Winther O, Krogh A. Phys Rev E Stat Nonlin Soft Matter Phys. 2004; 70(3 Pt 1):030903. [PubMed: 15524499]
21. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Proteins. 2008; 70(3):834–843. [PubMed: 17729286]
22. Kuhlman B, Baker D. Proc Natl Acad Sci U S A. 2000; 97(19):10383–10388. [PubMed: 10984534]
23. Liang S, Grishin NV. Protein Sci. 2002; 11(2):322–331. [PubMed: 11790842]
24. Löwdin, P-O. Academic Press. New York; London: 1964.
25. Miyazawa S, Jernigan RL. Journal of Chemical Physics. 2005; 122(2)
26. Word JM, Lovell SC, Richardson JS, Richardson DC. Journal of Molecular Biology. 1999; 285(4): 1735–1747. [PubMed: 9917408]
27. Dunbrack RL Jr, Cohen FE. Protein Sci. 1997; 6(8):1661–1681. [PubMed: 9260279]
28. Engh RA, Huber R. Acta Crystallographica Section A. 1991; 47:392–400.
29. Liang SD, Grishin NV. Proteins. 2004; 54(2):271–281. [PubMed: 14696189]
30. Wilson C, Gregoret LM, Agard DA. Journal of Molecular Biology. 1993; 229(4):996–1006. [PubMed: 8445659]
31. Petrella RJ, Lazaridis T, Karplus M. Folding & Design. 1998; 3(6):588–588.
32. Peterson RW, Dutton PL, Wand AJ. Protein Science. 2004; 13(3):735–751. [PubMed: 14978310]

33. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Proteins. 2009; 77(4):778–795. [PubMed: 19603484]

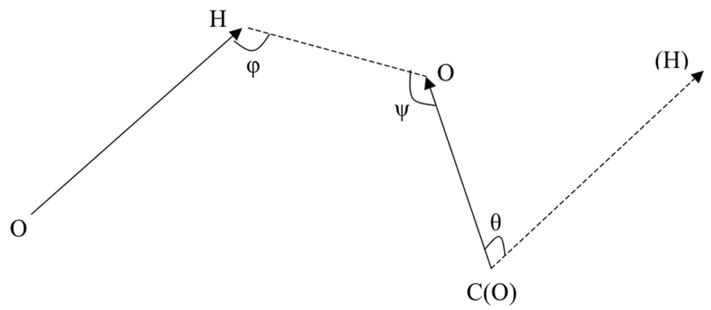


Fig 1.

Euler angles to define orientation dependent interaction demonstrated with O-H and C=O interacting dipoles

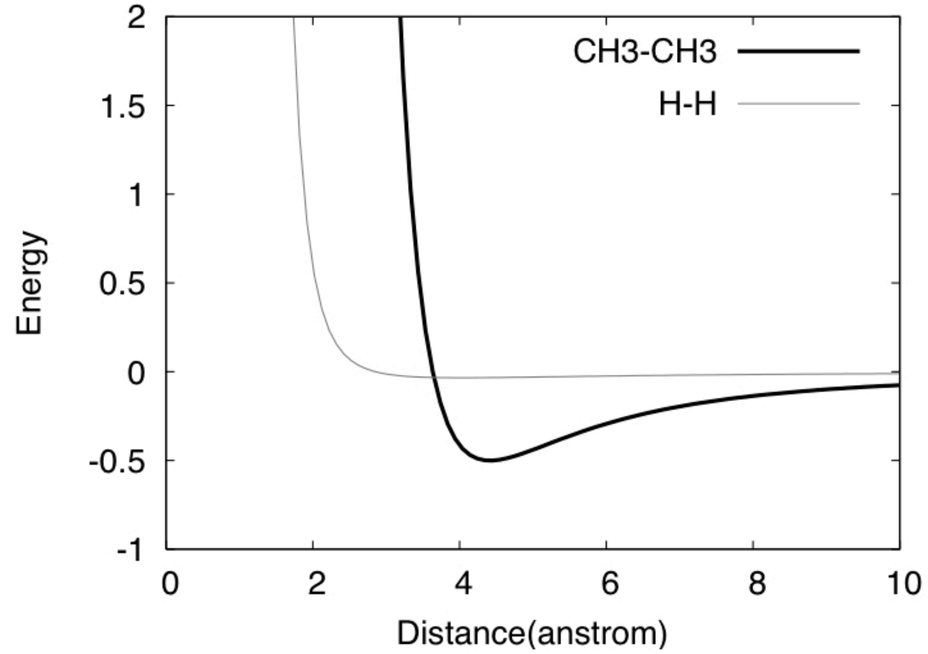
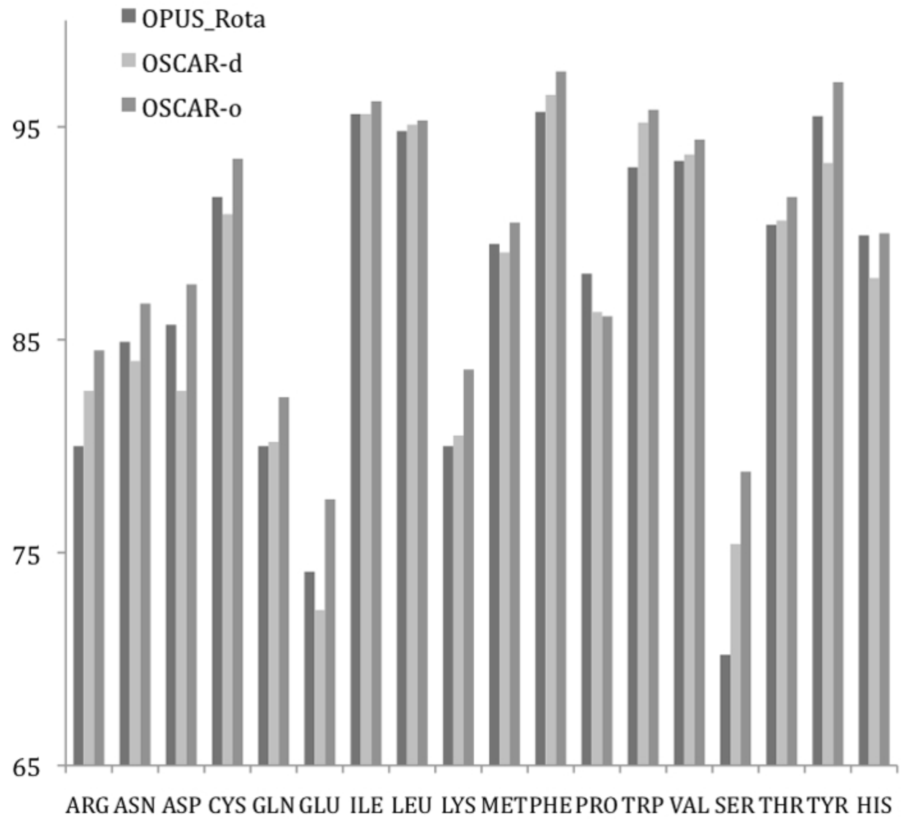


Fig 2.

The distance dependent interaction energies between CH3 and CH3 and between H and H.


OPUS_Rota
OSCAR-d
OSCAR-o
95
85
75
65
ARG
ASN
ASP
CYS
GLN
GLU
ILE
LEU
LYS
MET
PHE
PRO
TRP
VAL
SER
THR
TYR
HIS

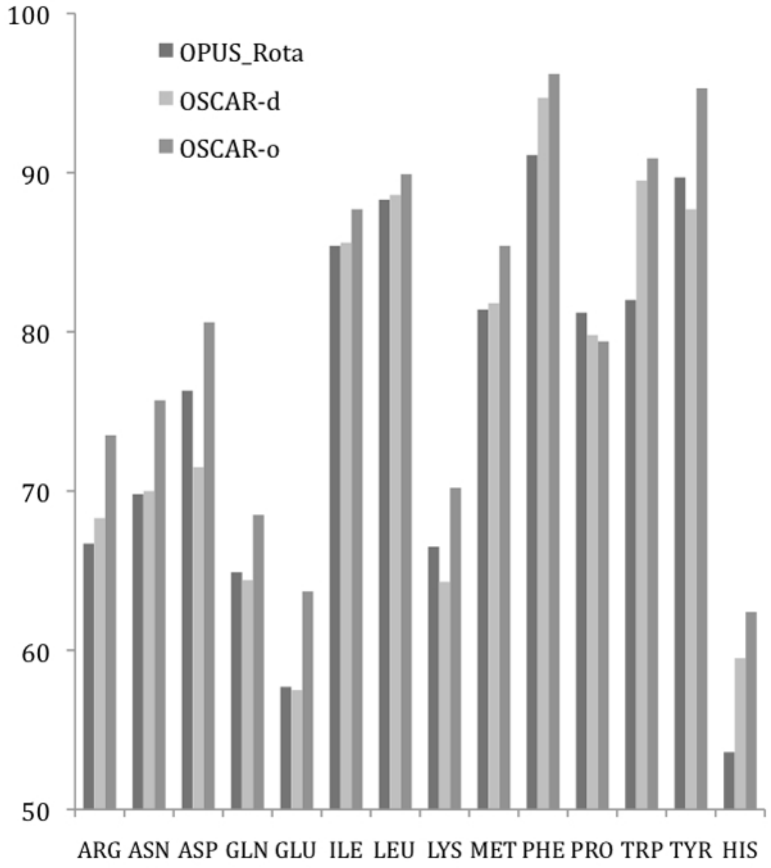


Fig 3.

Prediction accuracy of 218 test proteins for different residue types.

Table I

Prediction of side chain conformations for single residues.

	Training set[a]						30-protein test set[b]					
	All[c]	All[d]	All[e]	Core[f]	Core[f]	Core[f]	All[c]	All[d]	All[e]	Core[f]	Core[f]	Core[f]
	χ_1	χ_{1+2}	rmsd	χ_1	χ_{1+2}	rmsd	χ_1	χ_{1+2}	rmsd	χ_1	χ_{1+2}	rmsd
OSCAR-d	90.5	81.3	0.67	97.3	94.0	0.36	90.3	81.3	0.68	96.4	92.5	0.40
OSCAR-o	92.6	85.2	0.56	97.8	94.9	0.33	92.0	84.3	0.59	96.4	93.4	0.37

[a] 12000 residues of each residue type(excluding Gly and Ala) were calculated for the training proteins.

[b] The prediction accuracy of each residue type was calculated and averaged over 18 amino acids for the 30 test proteins.

[c] Fraction of correctly predicted $\chi 1$ for all residues.

[d] Fraction of residues with both $\chi 1$ and $\chi 2$ correctly predicted.

[e] The root mean square deviation between predicted and native side chain conformations.

[f] Prediction for core residues.

Table II

Single side-chain prediction by AMBER, CHARM and OSCAR-o.

			$\%\chi_1$ correct		$\%\chi_1 \times \chi_2$ correct	
	Average rmsd (A)	**Overall rmsd (A)**	**All**	**Core**	**All**	**Core**
AMBER	0.68	1.21	82	—	—	—
OSCAR-o	0.46	0.88	94	—	—	—
CHARMM	—	—	86.8	94.9	77.4	89.5
OSCAR-o	—	—	91.6	96.8	83.4	92.8

$\chi 1 \times \chi 2$, the number of residues with both dihedral angles correct, or single angle correct in the cases of valine, threonine, serine, and cysteine. Results for the AMBER and CHARMM force fields were obtained from Wilson and co-workers[30] and Petrella and co-workers,[31] respectively.

Table III

The accuracy of side-chain modeling for whole proteins by various energy functions.

	All χ_1	All χ_{1+2}	All rmsd[a]	Core χ_1	Core χ_{1+2}	Core rmsd[a]
LGA	87.5	73.3	1.34	93.7	84.5	0.85
OSCAR-d	88.0	76.5	1.32	95.1	89.5	0.65
OSCAR-o	89.4	80.8	1.20	95.5	92.1	0.60

[a] The overall rmsd of each of the 30 test proteins was calculated and averaged. The rows correspond to the three energy functions optimized in a similar way: our previous method LGA resulted from combining a knowledge-based term and multiple physics-based terms,[23] the distance-dependent optimized side chain atomic energy derived by series expansions (OSCAR-d), and the orientation-dependent version (OSCAR-o).

Table IV

Performance of different side chain modeling programs on recently released PDB entries.

	All χ_1	All χ_{1+2}	All rmsd	Core χ_1	Core χ_{1+2}	Core rmsd
SCWRL4	85.1	74.0	1.48	93.0	86.9	0.96
NCN	86.3	74.3	1.48	93.8	87.9	0.87
OPUS_Rota	86.6	75.7	1.40	94.3	87.6	0.86
LGA	86.1	72.3	1.42	93.9	85.9	0.91
OSCAR-d	86.6	75.3	1.41	95.5	90.4	0.70
OSCAR-o	88.8	79.7	1.24	95.9	91.9	0.62

Data are shown for three third-party methods (SCWRL4,[33] NCN,[32] and OPUS_Rota[9]), our previous method LGA,[23] and current OSCAR methods.