J. KSIAM Vol.15, No.3, 191–199, 2011

USING AN ABSTRACTION OF AMINO ACID TYPES TO IMPROVE THE QUALITY OF STATISTICAL POTENTIALS FOR PROTEIN STRUCTURE PREDICTION

JINWOO LEE

DEPARTMENT OF MATHEMATICS, KYUNGWON UNIVERSITY, SOUTH KOREA, [2]INSTITUTE OF INDUSTRIAL SCIENCE, UNIVERSITY OF TOKYO, JAPAN
E-mail address: jinwoolee@kw.ac.kr

ABSTRACT. In this paper, we adopt a position specific scoring matrix as an abstraction of amino acid type to derive two new statistical potentials for protein structure prediction, and investigated its effect on the quality of the potentials compared to that derived using residue specific amino acid identity. For stringent test of the potential quality, we carried out folding simulations of 91 residue A chain of protein 2gpi, and found unexpectedly that the abstract amino acid type improved the quality of the one-body type statistical potential, but not for the two-body type statistical potential which describes long range interactions. This observation could be effectively used when one develops more accurate potentials for structure prediction, which are usually involved in merging various one-body and many-body potentials.

1. INTRODUCTION

Ab-initio protein structure prediction is a famous unsolved problem in the computational sciences, and it is one of the biggest bottle necks in simulating biomolecules in silico [1]. the framework of every protein structure prediction procedure is based on the Afinsen's paradigm [2]; a protein's native state is the minimum free energy state. One defines a potential function, and explores the conformational space using a global optimization method.

Statistical potentials used for the protein structure prediction are called potentials of mean force, and describe favorable, or unfavorable interactions between residue specific atom types. It is well known that they are effective in selecting near native structures among many well-formed proteins like decoy configurations, and thus are widely used in the prediction community of protein structures as a component for quality assessment of generated protein models [3, 4, 5].

Most statistical potentials use residue specific atom types (see [6] and references therein.) However, even distinct amino acids could share similar structural preferences. It can be seen in that protein members in the same family are structurally well conserved even though some

Received by the editors June 1, 2011; Accepted August 18, 2011.
2000 *Mathematics Subject Classification.* 92B05.
Key words and phrases. statistical potential, protein structure prediction, position specific scoring matrix.

of their amino acids are mutated. This feature is well represented by a position specific substitution matrix (PSSM) of protein families [7], and it is an essential ingredient in every structure prediction procedure.

Since the matrix elements are more effective to describe the local environment of a residue compared to the residue atom identity, it would be feasible that they could be used instead of concrete atom types in the statistical potentials. Thus, we introduce PSSM as an abstraction of amino acid types, and tested the effect of replacing concrete atom types by PSSM.

For the test, we constructed new one-body and two-body statistical potentials named "profile contact potentials" since they describe favorable, or unfavorable interactions between abstract profile columns. The accuracy of any potential functions could be stringently judged by structure prediction test, so we investigated the effect by carrying out folding simulations of A chain of a protein 2gpi which is an alpha-beta protein with 91 residues.

Unexpectedly, the effect of the replacement was not uniform to the functional types that we used. The quality of the one-body potential which describes a local character of a given residue was improved, but not for the two-body potential which describes a rather global character of a protein shape.

2. Two New Statistical Potentials for Profile Contact

In this work, protein backbones are represented by heavy atoms, N, C_α, C, and O positions and side-chains are by side-chain centers (SC). Local backbone configurations and amino acid type determines the position of each side-chain center [8]. We included backbone heavy atoms in order to consider the excluded volume effect correctly. Side-chain center positions play an import role to define potential functions of this work.

2.1. **Abstract amino acid type.**

Most statistical potentials use residue specific atomic types such as alpha carbon of alanine, and beta carbon of lysine. However, it well known that even distinct amino acids could share similar structural preferences. For example, protein members in the same family are structurally conserved even though some of their amino acids are mutated.

A position specific substitution matrix (PSSM) [7] provides numerical values for a tendency of mutation of amino acids in structurally conserved protein families. It would be feasible to use the matrix as an abstraction of amino acid types. And thus unlike other statistical potentials, we decided to use 1 column of PSSM as an abstract representation of amino acid types a_i, and regarded as $a_i = a_j$ if $\cos^{-1} \frac{(a_i,a_j)}{\|a_i\|\|a_j\|} < \frac{\pi}{6}$, where $(\cdot,\cdot)$ is the dot-product of the column vectors, and $\|\cdot\|$ is the norm of it.

2.2. **One-body, local environment potential.**

A driving force for forming a tertiary structure (a global shape) of a protein is known as hydrophobic interaction [9] on which we lack a sufficient quantitative knowledge, and thus is deserved the term 'unsolved'. By that interaction, aquaphobia residues are buried inside the protein, and hydrophilic residues are exposed to solvent.

In [10], Hamelryck proposed a new solvent exposure measure called half-sphere exposure (HSE) which counts the number of C_α atoms within a half sphere which contains a vector from C_α to C_β and centered at a given C_α atom. Since proteins have two sides, buried and exposed, their idea is quite reasonable. A new one body potential in this paper is a direct conversion of their idea in the form of a statistical potential with a slight modification.

Let $S(x)$ be the area surrounded by the sphere of radius 11Å centered at x and the upper part of the plane normal to the vector from x to its side-chain center. Let $N(x)$ be the number of side-chain centers within $S(x)$.

We define the one-body score $S^1_{(\cdot)}(X|A)$ for a given protein model X with the amino acid sequence A as follows:

$$S^1_{(\cdot)}(X|A) = -\sum_{i=1}^{m} P_1(N(x_i)|a_i) \tag{2.1}$$

where m is the number of residues of X, x_i is the C_α position of i-th residue, and $P_1(\cdot|a_i)$ is the pre-calculated conditional probability distribution over $N(\cdot)$.

To estimate $P_1(\cdot|a_i)$, we used 5717 representative protein structures with less than 25% sequence identity from PISCES server [11], and the following formula:

$$P^1(n|a_i) = \frac{\sum_{k=1}^{5717}\sum_{i=1}^{m_k}\delta([N(x_i^k)/2],[n/2])\delta(a_i^k,a_i)}{\sum_{k=1}^{5717}\sum_{i=1}^{m_k}\delta(a_i^k,a_i)} \tag{2.2}$$

where x_i^k and a_i^k are the location and the (abstract) amino acid type of the i-th residue of the k-th protein in the database, respectively, and m_k is the number of residues of the k-th protein. $\delta(p,q) = 1$ if $p = q$, otherwise it is 0. $[x]$ is the greatest integer which is less than or equal to x. Let us denote (2.1) plus an excluded volume term taken from [12] with a sufficient large coefficient as $S^1_{atom}(X|A)$ if it is defined using residue specific atom identity. If a profile column is used as an abstract amino acid type, let us denote by $S^1_{profile}(X|A)$.

In short, (2.2) is just the probability for a specific residue type a_i to have n neighboring side-chain center atoms within the half-sphere in the experimental protein structures. For example, hydrophobic residues which are usually buried may have high n values with high probability compared to hydrophilic residues. (2.1) assigns a lower value to a model structure in which the distribution of aquaphobia and hydrophilic residues consistently follows the tendency of the known structures.

2.3. **Two-body potential.** We define two body profile potential $S^2_{(\cdot)}(X|A)$ which is distance dependent for a given protein model X with the amino acid sequence A as follows:

$$S^2_{(\cdot)}(X|A) = -\sum_{i=1}^{m-3}\sum_{j=i+3}^{m} P_2(r_{ij}|a_i,a_j), \tag{2.3}$$

where m is the number of residues of X, and r_{ij} is the distance between corresponding side-chain centers of (abstract) amino acid pairs a_i, a_j. $P_2(\cdot|a_i,a_j)$ is the pre-calculated conditional probability distribution over pair-wise distances for side-chain centers of amino acid pairs a_i, a_j.

To estimate $P_2(\cdot|a_i, a_j)$, we used the same database as the one-body case, and the following formula:

$$P_2(r_{ij}|a_i, a_j) = \frac{\sum_{k=1}^{5717} \sum_{i,j=1(i+1<j)}^{m_k} \delta([r_{ij}^k], [r_{ij}])\delta(a_i^k, a_i)\delta(a_j^k, a_j)}{\sum_{k=1}^{5717} \sum_{i,j=1(i+1<j)}^{m_k} \delta(a_i^k, a_i)\delta(a_j^k, a_j)} \tag{2.4}$$

where a_i^k, a_j^k and r_{ij}^k are the (abstract) amino acid pairs of the i-th and j-th residues of the k-th protein in the database, and the distance between their corresponding side-chain center atoms, respectively. m_k is the number of residues of the k-th protein in the database. $\delta(p, q) = 1$ if $p = q$, otherwise it is 0. $[x]$ is the greatest integer which is less than or equal to x. Let us denote (2.3) plus excluded volume term taken from [12] with a sufficient large coefficient as $S^2_{atom}(X|A)$ if it is defined using residue specific atom identity. If a profile column is used as an abstract amino acid type, let us denote by $S^2_{profile}(X|A)$.

In short, (2.4) is just the probability for the corresponding side-chain centers of a specific residue pairs a_i, a_j to be in the distance r_{ij} in known protein structures. For example, hydrophobic residue pairs could have low r_{ij} values with high probability since if they are both buried, they would be in contact with each other. (2.3) assigns a lower value to a model structure in which the distribution of the distances between amino acid pairs consistently follows the tendency of the known structures.

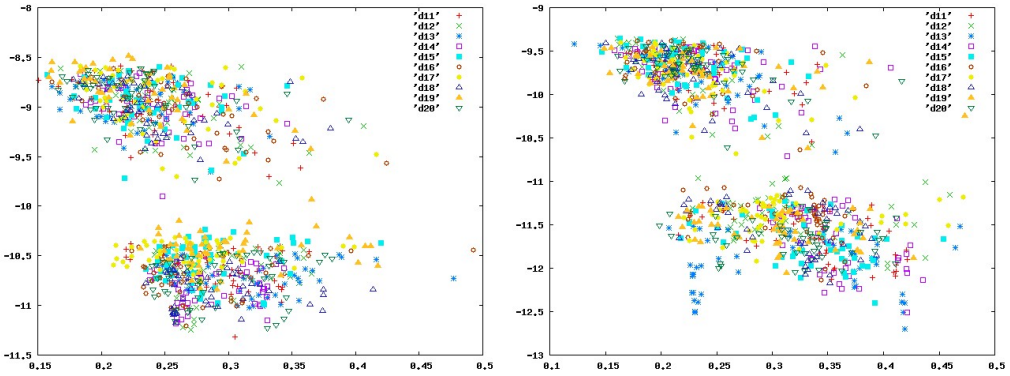
3. NUMERICAL SIMULATION

Local structures were sampled from the fragment library generated by Rosetta [12], and the assembly of them is guided by the proposed statistical potentials in Section 2. The scores contain the excluded volume term used in Rosetta with a high weight coefficient so that any overlap between heavy atoms is prohibited.

Most successful methods for protein structure prediction such as I-tasser, and Rosetta take the form of fragment assembly [8, 12]. By using well formed fragment from a protein database for local structures, the degrees of freedom to be searched is significantly reduced. Moreover, interactions that act to form the local structures need not to be considered explicitly. And thus we can concentrate solely on global interactions that are responsible for forming the tertiary structures.

For exploring the energy landscape, a variation of genetic algorithm called conformational space annealing (CSA) [13] is adopted. CSA is one of the most powerful global optimization methods that has been applied to generate favorable configurations in many systems, including protein AB models [14, 15], multiple sequence alignment [16], and template based modeling [17], just a few mentioned.

The overall procedure with specific parameters of CSA for this work is as follows. Initially, 50 random configurations are generated and subsequently energy minimized. For the energy minimization, we replace a part (selected in a random fashion) of a model protein A by a fragment from the fragment library for the corresponding region, and accept the new model if it is lower in score than that of A. We try this fragment replacement procedures 1000 times. We call the pool of 50 initial configurations as the first bank.



(A) $S^1_{atom}(X|A)$ (B) $S^1_{profile}(X|A)$

FIGURE 1. The energy landscapes of the one-body potentials (A) with residue specific atom identities, and (B) with new abstract atom types are shown.

The first bank conformations are copied to *bank*, and they are evolved iteratively as follows. Distant 20 unused seed configurations are randomly selected. 15 trial conformations for each seed and thus $20 * 15 = 300$ in total is formed by replacing a consecutive part (up to 40% of the chain length) selected in a random fashion of each seed by the corresponding part of another bank conformation (also selected in a random fashion), and subsequently energy minimized as above. They are utilized for bank updating procedure.

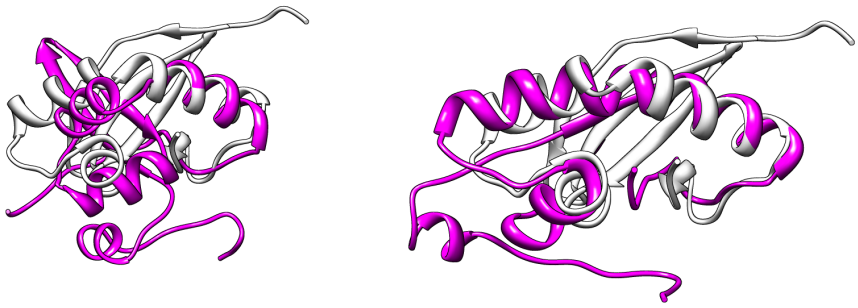
Each trial conformation A is compared with the closest one α in the bank in terms of root mean square deviation (RMSD). If they are similar, A is taken if it is lower than α in energy, and set the seed status as unused. In the case that they are dissimilar, A is compared to the one β with the highest energy in the bank, and A is taken if it is lower in energy than β, and set the seed status as unused. Otherwise A is rejected. In this way, the number of bank conformations is fixed to 50.

If RMSD value between two conformations is less than D_{cut} they are regarded as similar. D_{cut} is set initially to half of the average distance D_{ave} of the first bank conformations, and it is reduced by a factor of $\frac{2}{5}\frac{1}{166}$ at the end of every iteration step. D_{cut} value is fixed to $\frac{D_{ave}}{5}$ after 166 iteration steps. This completes one iteration, and we repeat. The CSA procedure stops automatically if there is no unused seed.

Since the CSA is a stochastic process, we carried out 10 independent runs with different random numbers for each case of the study (see section 4). We note that it took about 20 minutes using 80 cores of 2.40 GHz Intel Xeon processors for 10 independent runs.

4. RESULTS

4.1. **Effect on one-body potential.** Figures 1a and 1b show the energy landscapes of $S^1_{atom}(X|A)$, and $S^1_{profile}(X|A)$, respectively. The Y-axis is the energy value of $S^1_{(\cdot)}(X|A)$ and X-axis is a



(A) $S^1_{atom}(X|A)$ (B) $S^1_{profile}(X|A)$

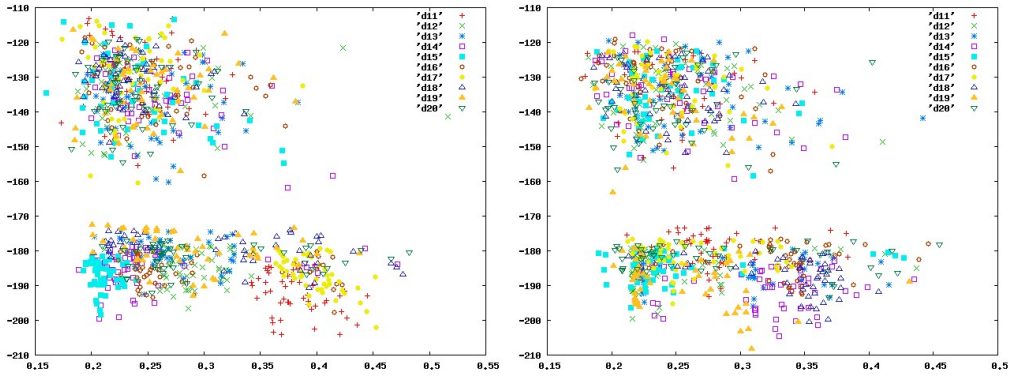
FIGURE 2. The lowest energy models (magenta) generated using (A) $S^1_{atom}(X|A)$ and (B) $S^1_{profile}(X|A)$, respectively, superposed with the native structure (silver) are shown. More kinked alpha helix in the left panel is well adjusted to the native in the right panel.

value for similarity between the generated protein model and the native structure called TM-score. TM-score is between 0 and 1, and 1 for the perfect match of two structures.

In each figure, we observe that first bank conformations with high energies are evolved to form the final bank conformations with low energies. Our analysis is focused on the overall shape of low energy regions. When we use residue specific atom identities ($S^1_{atom}(X|A)$), the lowest energy regions are found between 0.25 and 0.35 as shown in Figure 1a. On the other hand, if profile columns are used as abstract amino acid types, the lowest energy regions are shifted to the right and are found between 0.4 and 0.45 as shown in Figure 1b. Moreover, the overall shape of the energy landscape of $S^1_{profile}(X|A)$ is more funnel shaped than that of $S^1_{atom}(X|A)$, i.e., even though there is a very narrow low energy region between 0.2 and 0.25 in Figure 1b, it is roughly observed that the quality of models is improved as their energy is lowered.

The lowest energy model (magenta) of each case superimposed with the native structure (silver) is shown in Figures 3a and 3b. In the left panel, more kinked alpha helix in front generated using $S^1_{atom}(X|A)$ is well adjusted to the native structure as shown in the right panel by using $S^1_{profile}(X|A)$. Further structural comparison data is shown as below.

Search Score	TM-score	Matched Length	Matched RMSD
$S^1_{atom}(X\|A)$	0.3054	35	2.98
$S^1_{profile}(X\|A)$	0.4191	53	2.80



(A) $S^2_{atom}(X|A)$ (B) $S^2_{profile}(X|A)$

FIGURE 3. The energy landscapes of the two-body potentials (A) with residue specific atom identities and (B) with new abstract atom types are shown.

By using $S^1_{profile}(X|A)$, 58 % of the whole chain is correctly modeled with RMSD of that matched region being 2.80. Well matched region is much extended compared to 38% of the $S^1_{atom}(X|A)$ case.

4.2. **Effect on two-body potential.** Figures 3a and 3b show the energy landscapes of $S^2_{atom}(X|A)$, and $S^2_{profile}(X|A)$, respectively. Using residue specific atom identity for $S^2_{atom}(X|A)$, the lowest energy region is formed between 0.35 and 0.45. On the other hand, it is formed between 0.3 and 0.35 using abstract profile atom types. More detailed comparison of the lowest energy model of each case to the native structure is shown as below.

Score	TM-score	Matched Length	Matched RMSD
$S^2_{atom}(X\|A)$	0.3668	35	1.94
$S^2_{profile}(X\|A)$	0.3086	33	2.61

The aligned length of the lowest energy model of $S^2_{atom}(X|A)$ is 38 % of the whole chain with RMSD of the matched region being 1.94, which is slightly better than the case of $S^2_{profile}(X|A)$ with corresponding values 36 % and 2.61, respectively. Overall, it seems to be better to use residue specific atomic identity for two-body potentials than to use profile columns, although the difference is marginal.

5. DISCUSSION

We derived two new statistical potentials for protein folding from a protein structure database. One-body potential is obtained by calculating the probabilities for a specific residue type

to have n neighboring side-chain center atoms within the half-sphere in the experimental protein structures. The half-sphere is centered at C_α atom of a_i and containing the vector from C_α to side-chain center of it. Two-body potential is obtained by calculating the probabilities for the corresponding side-chain centers of specific residue pairs to be in a distance bin in known protein structures.

In the derivation of the potentials, we used both residue specific atomic identity and profile columns as an abstract amino acid type, and revealed its influence on the quality of statistical potentials by performing folding simulations of 91 residue long A chain of protein 2gpi. We found that profile columns used in the one-body potential improved the quality of the potential much compared to the case of using atomic identity. On the other hand, for two-body potentials, it was better to use atomic identity, although the difference is marginal.

The position specific substitution matrices (PSSM) are successfully applied for protein secondary structure prediction [18]. This work is an attempt to use PSSM to predict tertiary structures. Since PSSM is an effective abstraction for an environment of an amino acid, it is understandable it improves the quality of one-body statistical potential, which describes preferable local environment. However, it did not improve the quality of two-body potential. It means that the residue specific atomic identity itself is more influencing factor for the long range interactions than the local environment of a position. In other words, the long range interaction between two spots having environment types A and B was not preserved among various protein structures. This observation could be effectively used when one develops more accurate potentials for structure prediction, which are usually involved in merging various one-body and many-body potentials.

ACKNOWLEDGMENTS

This work was supported by National Research Foundation of Korea Grant funded by the Korean Government (KRF-2008-313-C00122)

REFERENCES

[1] A. Neumaier, *Molecular modeling of proteins and mathematical prediction of protein structure*, SIAM Rev, **39** (1997), 407–460.
[2] C.B. Anfinsen, *Principles that govern the folding of protein chains*, Science, **181** (1973), 223–230.
[3] R. Samudrala and J. Moult, *An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction*, J Mol Biol, **275** (1998), 895–916.
[4] M.J. Sippl, *Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins*, J Mol Biol, **213** (1990), 859–883.
[5] H. Zhou and Y. Zhou, *Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction*, Protein science, **11** (2002), 2714–2726.
[6] T. Hamelryck, *Potentials of mean force for protein structure prediction vindicated, formalized and generalized*, PLoS ONE, **5** (2010), 1–11.
[7] M. Gribskov, A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins*, Proc Natl Acad Sci, **84** (1987), 4355–8.
[8] Y. Zhang, *Template-based modeling and free modeling by I-TASSER in CASP7*, Proteins, **69(Suppl 8)** (2007), 108–117.
[9] D. Chandler, *Interfaces and the driving force of hydrophobic assembly*, Nature, **437** (2005), 640–647.

[10] T. Hamelryck, *An amino aicd has two sides: a new 2D measure provides a different view of solvent exposure*, Proteins, **59** (2005), 38–48.
[11] G. Wang and R.L.Dunbrack, *PISCES: a protein sequence culling server*, Bioinformatics, **19** (2003), 1589–1591.
[12] R. Das and D. Baker, *Macromolecuar modeling with rosetta*, Annu.Rev.Biochem., **77** (2008), 363–382.
[13] J. Lee, H.A.Scheraga, and S.Rackovsky, *New optimization method for conformational energy calculations on polypeptides: conformational space annealing*, J. Comput. Chem., **18** (1997), 1222–1232.
[14] S-Y Kim, S.J. Lee, and J. Lee, *Conformational space annealing and an off-lattice frustrated model protein*, J. Chem. Phys., **119** (2003), 10274–10279.
[15] J. Lee, K. Joo, S-Y Kim, and J. Lee, *Re-examination of structure optimization of off-lattice protein AB models by Conformational space annealing*, J. Comput. Chem., **29** (2008), 2479–2484.
[16] K. Joo, J. Lee, I. Kim, S.J.Lee, and J. Lee, *Multiple sequence alignment by conformational space annealing*, Biophysical J., **95** (2008), 4813–4819.
[17] K. Joo, J. Lee, J-H Seo, K. Lee, B-G Kim, and J. Lee, *All-atom chain-building by optimizing MODELLER energy function using conformational space annealing*, Proteins., **75** (2009), 1010–1023.
[18] DT Jones, *Protein secondary structure prediction based on position-specific scoring matrices*, J. Mol. Biol., **292** (1999), 195–202.