Designing of knowledge-based potentials via B-spline basis functions for native proteins detection

Running Title: Designing of knowledge-based potentials via B-spline for proteins

Elmira Mirzabeigi[a], Saeed Mortezazadeh[b], Rezvan Salehi[a], Hossein Naderi-Manesh[b*]

e.mirzabeigi@modares.ac.ir

s.mortezazadeh86@gmail.com

r.salehi@modares.ac.ir

naderman@modares.ac.ir

***Corresponding author**

Email: naderman@modares.ac.ir

Department of Nanobiotechnology, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, P.O. Box 14115-175, Iran

[a] Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, P.O. Box 14115-134, Iran

[b] Department of Nanobiotechnology, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, P.O. Box 14115-175, Iran

ABSTRACT

Knowledge-based potentials are developed to investigate the differentiation of native structures from their decoy sets. This work presents the construction of two different distance-dependent potential energy functions based on two basic assumptions using mathematical modeling. In the first case, according to Anfinsen's dogma, we assumed that the energy of each model structure should be more positive than the corresponding native type. In the second one, we assumed that the energy difference between the native and decoy structures changes linearly with the root-mean-square deviation of structures. These knowledge-based potentials are expressed by the B-spline basis functions of the pairwise distances between C_α-C_α of inter-residues. The potential function parameters in the above two approaches were optimized using the linear programming algorithm on a large collection of Titan-HRD and tested on the remainder. We found that the potential functions produced by Anfinsen's dogma detect native structures more accurately than those developed by the root-mean-square deviation. Both linear programming knowledge-based potentials (LPKP) successfully detect the native structures from an ensemble of decoys. However, the LPKP of the first approach is able to correctly identify 130 native structures out of 150 tested cases with an average rank of 1.67. While the second approach LPKP detects 124 native structures from their decoys. We concluded that linear programming optimization is a promising method in generating knowledge-based potential functions. All the high-resolution structures (training and testing) used for this work are available online and can be downloaded from `http://titan.princeton.edu/HRDecoys`.

Keywords

knowledge-based potential, B-spline basis function, native structure detection, optimization, linear programming

1 Introduction

Proteins are important macromolecules that involved in all cellular processes which theirs functions are directly related to the 3D structure. The prediction of the protein structure is one of the incentive questions in computational biology. Although we know millions of protein sequences, less than a hundred thousand protein structures have been found until now [1]. It can be indirectly concluded that determining the protein structure at the atomic level is overwhelming. Reports on *ab initio* technique developments show significant progress in predicting protein structure in previous years, however, the quality of models does not have enough accuracy to be useful for biologists [2]. Hence, the bioinformatics approach is most widely used to predict the tertiary structure of proteins.

According to Anfinsen's dogma, the native structure is determined by the amino acid sequence of protein which means at the environmental condition that folding occurs, the native structure is formed at the global minimum free energy [3]. Force fields have been developed to calculate the potential energy of molecular systems which refer to the functional form and parameter sets that can be derived from empirical and theoretical studies. The protein structure is considered as atomic or coarse-grained so that potential functions can include pairwise interactions, side-chain orientations, secondary structural preferences, solvent-exposure, and other geometric properties of proteins [4, 5]. The accuracy of these potential functions can be traced by evaluating the ability to detect the native structure from a decoy ensemble. There are two general knowledge-based and physics-based approaches to introduce potential functions [6, 7, 8, 9, 10]. Knowledge-based potentials are simple potentials extracted from the protein data bank designed to improve the quality of protein models whereas, in physics-based potentials, the chemical properties of the molecule are also taken into account. Various methods have been developed to increase the accuracy of these potential functions in recent years [11, 12, 13, 14].

In most cases, knowledge-based potentials are used to sort decoy structures by calculating similarity to the native structure. The inability of the proposed model to identify the native structure as the lowest potential energy is regarded as the error rate of the potential functions. There are two methods to generate knowledge-based potentials. The first method is to use the Boltzmann inversion equation to convert the distribution of geometric properties of known structures to potential energy functions. This method is also called the statistical potential obtained from the ratio of observed frequencies to the reference state. Therefore, several models have been developed for representing reference states such as Sippl's assumption, distance scaled finite ideal-gas reference state (DFIRE), and discrete optimized protein energy (DOPE). In the second method, the potentials are extracted from a training process to quantitatively differentiate between the incorrectly folded models and native structures [1].

Using physical potential functions and considering more than 40 physical constraints in a linear program-

ming problem, Rajgaria et.al [15] have achieved better results than previous studies on the Titan-HRD protein data set. Despite the desirable results, it is inevitable to consider several constraints in the optimization problem using the physical potential functions. In some studies, the least-square problem has been used to optimize knowledge-based potentials. Since the least-square problem is based on norm-2, the energy difference between the native structure and the decoy can give both positive and negative values. Therefore, the use of this method in protein modeling makes the Anfinsen's dogma not applicable to the potential production process. In these studies, choosing the best decoy instead of recognizing the native structure is the main parameter for model evaluation. The best decoy is determined by calculating the minimum distance and the lowest energy difference with the native structure. In addition to the inability to distinguish the native structure from the decoy, there is also a high computational cost for using the least-squares method [1, 16].

Although many influential factors are involved in the production of knowledge-based potentials. But in this study, the modeling of these functions with a mathematical approach is discussed. Given that the purpose of modeling is to reduce the parameters involved in the problem as much as possible. We considered the pairwise distances between the C_α of amino acids so that the knowledge-based potentials are formulated using B-spline basis functions. We introduce two different optimization processes to obtain potential parameters. Both are based on Anfinson's dogma that in the first method only the energy difference of the native and decoy structure is included in the modeling while in the latter method it is assumed that the energy difference is linearly proportional to the distance of structures. We solved them by using MATLAB software and checked the results of the two methods. Finally, our results were cross-checked with previously reported data. Our LPKPs have better detection for native structures compared to other methods. Both LPKPs are able to identify 130 and 124 native structures out of 150 tested cases, respectively.

2 Materials and Methods

In this model, the amino acids are represented by the location of C_α atoms. Hence a protein with m amino acids is expressed by a vector, $C = (c_1, c_2, ..., c_m)$ where c_k is the location of C_α of the k-th amino acid having the following three-dimensional coordinates:

$$c_k = (x_k, y_k, z_k).$$

We denote the native structure of each protein with $N = (n_1, n_2, ..., n_m)$ and decoy structures with $D_i = (d_1, d_2, ..., d_m)$, where i is the decoy number.

2.1 Geometric Distance Between Two Protein Structures

The Euclidean distance between two points is the basis for calculating the geometric distance between two protein structures. Root-mean-square deviation (RMSD) method was used for this purpose so that the first protein structure is considered to be the reference and the second structure is best fitted to the first one using a transformation operator G:

$$RMSD = \min_{G} \sqrt{\frac{\sum_{k=1}^{m} \|n_k - G(d_k)\|^2}{m}},$$

where n_k and d_k are the k-th location of the C_α atoms in native reference structure and decoy structures, respectively, and m is the number of amino acids. The transformation G involves a set of translation and rotation operations to obtain the best structural alignment between two protein structures and it does not change the shape or size of the proteins [17, 18, 19]. RMSD as defined above is a norm-2 metric distance [20]. The norm-2 (also written "L^2-norm") $\|x\|$ is a vector norm defined for a vector $x^T = [x_1, x_2, \dots x_n]$ by

$$\|x\| = \sqrt{\sum_{k=1}^{n} |x_k|^2},$$

where $|x_k|$ denotes the absolute value of x_k [21].

2.2 Knowledge-based Potential Function

According to Anfinsen's dogma, potential energy functions must be obtained in such a way that the energy of the protein's native structure has the lowest value in a set of decoy structures [3]. This hypothesis is shown in the following constraint:

$$E_{D_i} - E_N > \varepsilon, \tag{1}$$

in this equation, E_N is the energy vector of a native structure and E_{D_i} is the energy vector of the i-th decoy structure of the relevant protein [15]. If we have p proteins and each protein has i decoys, then constraint (1) is reformulated as below:

$$\sum_{p,i} (E_{p,i}(X) - E_{p,n}(X)) > \varepsilon, \tag{2}$$

where X is a vector of parameters that indicates the number of interactions between different types of amino acids at different distances of C_α-C_α in the protein structure. Given that the number of natural amino acids is 20, the total number of types of interactions is equal to 210. Also, each type of interaction is expressed with eight parameters representing potential energy at different intervals. Therefore, the

total number of parameters needed to calculate the potential energy between different types of amino acids at different distances is equal to $210 \times 8 = 1680$. In equation (2), ε is an arbitrary number in which different values have been tested and the best of them have been 0.01. In the p-th protein, $E_{p,i}$ is the energy matrix of the i-th decoy, and $E_{p,n}$ is the energy matrix of the native. Now equation (2) can be rewritten as follow, where parameters S_p are positive frail variables:

$$(E_{p,i} - E_{p,n})X - S_p \geq \varepsilon. \tag{3}$$

In the first approach, the knowledge-based potential functions were optimized through the criteria formulated in equation (3). Also, assuming that the energy difference between the native and the decoy structure can be related to the geometric distance between them, a similar condition can be extended to optimize the potential functions [1, 16, 22]. This constraint is shown as below:

$$E_{D_i} - E_N \propto Dis_{D_i,N}, \tag{4}$$

where, $Dis_{D_i,N}$ is the distance between decoy D_i and native N. The equation (4) is rewritten in the following:

$$E_{D_i} - E_N = \alpha.Dis_{D_i,N}. \tag{5}$$

In this equation, α is constant that is considered as 1 in [22], so the constraint (5) is rewritten for p proteins as below, where $D_{p,i,n}$ is the distance between native structure and i-th decoy structure of p-th protein.

$$\sum_{p,i} E_{p,i}(X) - E_{p,n}(X) \leq D_{p,i,n}. \tag{6}$$

The extended uniform cubic B-spline possesses the convex hull property, symmetry, and geometric invariability [23]. These features are convincing to use this function as the basis for the potential functions. We used this function at eight uniform intervals for each of the 210 types of interactions between amino acids shown in Table 1. So, the energy of all decoy and native structures is calculated using the equation below [1, 22]:

$$E(\theta) = \sum_{i<j} \sum_{p} X_p^{aa(i),aa(j)} B_p(r_{i,j}). \tag{7}$$

In this equation, $aa(i) \in \{1, ..., 20\}$ is the amino acid type of the i-th C_α and $B_p(r_{i,j})$ is the p-th B-spline basis function evaluated on the distance between the i-th and j-th C_α, shown in Figure 1. Also, $X_p^{aa(i),aa(j)}$ are the model parameters determined by the optimization according to the following.

2.3 Optimization Problems

We designed energy function using an optimization procedure, assuming that for every native structure $N_{n,p}$, we have a set of decoy structures $D_{i,p}$. At each stage, $(E_{i,p} - E_{n,p})$ is the energy difference and $D_{i,n,p}$ is the corresponding distance between the native structure and j-th decoy in the p-th protein. Also, the sum of the frail variables constraint (2) was minimized in each scheme. Since the frail variables are positive, the condition $S_p \geq 0$ was added to each optimization problem. For vector X, we consider the subscription of general constraints [15] for the condition $-4 \leq X \leq 4$.

2.3.1 First optimization scheme

The first optimization problem can be written into linear programming which used Anfinsen's dogma as below, that was named with LPKP[1]:

$$\begin{aligned} \min_{X,S_1,\ldots,S_p} \quad & \sum S_p, \\ s.t \quad & (E_{i,p} - E_{n,p})X - S_p \geq \varepsilon, \\ & S_p \geq 0, \\ & -4 \leq X \leq 4. \end{aligned} \tag{8}$$

where the vector X is a set of parameters of the energy functions, and p is the number of proteins in the training set.

2.3.2 Second optimization scheme

The second optimization problem (called LPKP[2]) can be written into linear programming using Eq.(8) and the relationship between distance and energy difference of two structures as following:

$$\begin{aligned} \min_{X,S_1,\ldots,S_p} \quad & \sum S_p, \\ s.t \quad & (E_{i,p} - E_{n,p})X - S_p \geq \varepsilon, \\ & (E_{i,p} - E_{n,p})X \leq D_{i,n,p}, \\ & S_p \geq 0, \\ & -4 \leq X \leq 4. \end{aligned} \tag{9}$$

This scheme is similar to the first scheme in terms of protein number, optimization method, and the number of parameters X except the constraint (5).

2.4 Training and Test Sets

The high-resolution decoy set contained 1400 protein structures, with 500-1600 decoys for each protein [15], but we deleted proteins in which decoys and native structures do not have the same number of amino acids. Finally, the 1370 proteins remained from the Titan-HRD. The entire set of protein decoy structures has been made available at `http://titan.princeton.edu/HRDecoys/`. From 1370 proteins used for decoy generation, 1220 proteins were randomly selected for training processes. Since each protein has at least 500 decoy structures, therefore, all decoys of each protein were sorted based on their C_α RMSD and then 500 decoys were randomly selected to cover the whole RMSD range. This arrangement of a training set has $500\times\ 1220 = 610000$ decoy structures. Because of computational limitations, it is not possible to include all 610000 decoys in the training step. Therefore, we reduced 500 to 45 decoys per protein to set up the training procedure with 60000 decoy structures. The remaining 150 proteins of Titan-HRD set was used for testing the obtained potential functions. In this step, also 500 decoys were selected using the same technique explained above in generating the training set. A summary of the test and training data sets are given in Table 2.

3 Results and Discussion

The existence of effective factors has been led to the use of various methods for the production of knowledge-based potential functions. In this study, we have been modeled these functions taking into account a limited number of these parameters. The knowledge-based potential functions are designed according to eight uniform cubic B-spline functions and the distances between C_α atoms in protein structures. These potential functions were extracted from the information of 60000 decoy structures. Two sets of energy parameters were constructed based on two different optimization schemes of LPKP^1 and LPKP^2. The main constraint in the LPKP^1 model is the Anfinsen's dogma [3]. Hence, the energy difference between the decoy and the native structure must always be greater than a positive constant value called ε, which are shown in constraint (3). In the LPKP^2 model, the effect of RMSD examined on our scheme by adding constraint (6). The objective function of each scheme has minimized the sum of the frail variables, discussed in the methods section. The zero value for an objective function means that there no infringements. In this study, the objective function has obtained a value close to zero, which represents the very low error of our introduced optimization model. Previously, the optimization schemes (5) and (9) were described, and the method was generally explained in Algorithm 1.

At first, the value of ε considered 0.01 in schemes (5) and (9). The optimization schemes implemented with ε and approximately 60000 structures. Since the ability to identify between the native and native-like structures is a considerable standard for any potential function, the optimization results have been

Algorithm 1 Execution of knowledge-based potential function

Input: M, N_{decoys}.
▷ M : Number of Proteins.
▷ N_{decoys} : Number of Decoys in a Similar Protein.
Output: $E_{i,j}$, $D_{i,n,p}$.
▷ $E_{i,j}$: Energy difference between two structures.
▷ $D_{i,n,p}$: RMSD distance between two structures.
▷ θ^1 : The model parameters determined by the optimization (5).
▷ θ^2 : The model parameters determined by the optimization (9).

for $i = 1$ to M **do**
  Read C_α three-dimensional coordinates for native structure.
  Solve $Dis_{i,1}$ for each pair of C_α in native structure.
  Solve problem E_{Native_i} (7) for native structure.
  for $j = 1$ to N_{decoys} **do**
    Read C_α three-dimensional coordinates for decoy structure.
    Solve $Dis_{i,j+1}$ for each pair of C_α in decoy structure.
    Solve problem $E_{Decoy_{i,j}}$ (7) for decoy structure.
    Set $E_{i,j} = E_{Decoy_{i,j}} - E_{Native_i}$.
    Solve RMSD problem $D_{i,n,p}$ (8).
  end for
end for
Solve problem (5) for θ^1.
Solve problem (9) for θ^2.

tested to detect native structure. The LPKP examined 500 decoys for each of the 150 test proteins of the Titan-HRD decoy set. In this examination, the comparative rank of the native among decoy structures has been calculated. An ideal potential function should be able to detect rank 1 for the native structures of all the proteins in the test set. As we know, it is an accepted condition that the test set should not share with the training set since it invalidates the potential energy appraisement. Hence, our test set was carefully selected so that the training and test sets had no overlap. The examination results are presented in Table 6. According to that, 130 native structures were correctly recognized among 150 proteins with an average rank of 1.67 in method (5). Moreover, native structures were distinguished in 124 proteins of scheme (9) with an average rank of 2.83. The results represent that our proposed method is significantly more accurate than previous methods [15, 24, 25, 26].

Other force fields such as HR [15], LKF [24], TE13 [25], and HL [26] have been tested on this set of high-resolution decoys. All these force fields were fundamentally different from each other in their methods of energy estimation. The HR force field is a novel C_α-C_α distance-dependent potential where the interaction distance range 3-9 Åis divided into eight bins. Also, this method has the nearest results to our model with the detection limit of 113 native structures. The LKF force field is a C_α-C_α distance-dependent potential where the interaction distance range 3-9 Åis divided into eight bins. However, the LKF was not successful in the detection of native structures. Moreover, the TE13 force field is also a distance-dependent (13 bin) force field, but the interaction distance is measured between the geometric

centers of the side chain of two interacting residues. The HL force field is a contact-based force field, such that two conditions needed to consider a pairwise amino acid contact. First, they have to be at least five residues apart from each other and secondly, the distance between their non-hydrogen atoms must be less than 4.5 angstroms. The comparison of the energy rankings obtained using different force fields presented in Table 3. As one can see in table 3, according to the ranks, LPKPs are the best in identifying the native structures. Assigning rank 1 to the native structure means that the force field is adept at finding the native structure from an array of its nonnative configurations. therefore, the LPKPs could increase the percentage of native fold recognition from 75.33 to 86.67.

The results of optimization (5) and (9) were tested to calculate the correlation coefficient of each of the proteins in the test Titan-HRD set. We used the below formulation for computing the correlation coefficient:

$$Corr(d, E) = \frac{1}{N-1} \sum_{i=1}^{N} \frac{S_d(X_i) - \langle S_d \rangle}{\sigma(S_d)} \frac{E(X_i) - \langle E \rangle}{\sigma(E)}.$$

The $Corr(d, E)$ defined between values of energy $E(X_i)$ and distance $S_d(X_i)$ for all of the X_i decoy structures. Furthermore, $\langle . \rangle$ and $\sigma(.)$ were determined the mean and standard deviation. The reason for defining the correlation coefficient $Corr(d, E)$ is that it can measure the quality of energy function E with respect to the distance S_d. Initially, by examining the correlation of each protein, we found that the scheme (5) had better than (9) in the correlation coefficient LPKP and RMSD. For the presentation of these results, two proteins randomly were chosen in Figure 2, and have been compared for two optimization methods. The correlation had been approximately 0.7 for the first method and this value reduced to 0.6 in the second.

Considering the correlation graph of all proteins in Figure 3, we found that the scheme (5) had a mean correlation of 0.7. Hence, the scheme (5) had more than 80 proteins with a correlation of 0.6. These results indicate a high correlation in LPKP. Furthermore, by comparing the two graphs in Figure 3, we found that scheme (5) has a higher correlation amount. Also, method (9) had a mean correlation of 0.6 and had more than 70 proteins with a correlation of 0.6.

Originally, the number of decoys and the value of epsilon were considered constant in our training set. Since the LPKP[1] brought on better results, in the following, the conditions are discussed for this modeling method. Hence, we considered five different values for the number of training set decoys. In this experiment, we included the values of 15, 25, and 35 instead of 45 for the number of decoys in each protein. The results of identifying native structures for each of these values are summarized in Table 4. We observed that as the number of decoys increased to 35, our model accuracy increased dramatically. However, above this had no significant effect on the rate of detection of the native structure. Furthermore, according to Figure 4, in each protein, the graph of different values for the number of training set decoys is very

similar to the logarithmic function. Considering this diagram, if the number of decoys selected for the training set is more than a certain amount, there is no significant change in the results and only increases the computational volume of the modeling. Therefore, we selected 45 decoy structures per protein for use in the process of developing knowledge-based potentials. In the last step, different values for ε has been considered in the optimization schemes. We realized the slight changes in the amount of ε do not create a significant impact on our modeling by examining these values.

Although our priority was detecting the native structure among decoys as the first structure, and our proposed scheme was able to present adequate detect regardless of elimination of many involved parameters and mathematically modeling, for comparison the LPKPs with more recent methods, the best decoy introduced in [1] is used, which is demonstrated in Table 5.

4 Conclusion

Knowledge-based potentials are developed to derive native structures from their decoy sets. In this work, we constructed two different sets of distance-dependent potential energy functions based on the two basic assumptions. At first, we assumed that the energy of each decoy should be more positive than the corresponding native type. In the next step, it is assumed that the energy difference and the distance between the two structures are linearly dependent. The RMSD was used to calculate the distance between the decoys and native structures. Each of the potential energy functions has terms of pairwise distances between C_α-C_α and is expressed using B-spline basis function. We optimized the parameters of the potential function by using two linear programming problems on a large collection of Titan-HRD decoy set. Furthermore, the obtained results were tested on the remainder of Titan-HRD. We found that the potential functions developed based on Anfinsen's dogma have more accurate detection than those developed by the root-mean-square deviation of structures. However, both linear programming knowledge-based potentials (LPKP) were successful to recognize the native structures from an ensemble of high-resolution decoys. The LPKP in the first scheme was able to identify correctly 130 native structures out of 150 test cases with an average rank of 1.67 and the second LPKP scheme was able to detect 124 native structures with an average rank of 2.83. This indicates that the linear programming optimization is a promising method in generating knowledge-based potential functions. All the structures including training and testing Titan-HRD used for this work are available online and can be downloaded from `http://titan.princeton.edu/HRDecoys`.

References

[1] Carlsen, M., Koehl, P., Røgen, P. (2014), On the importance of the distance measures used to train and test knowledge-based potentials for proteins, *PLoS One*, **9:** 1-18 .

[2] Zhang, Y. (2008), Progress and challenges in protein structure prediction, *Curr. Opin. Struct. Biol.*, **18:** 342-348.

[3] Anfinsen, C. (1973), Principles that govern the folding of protein chains, *Science*, **181:** 223-230.

[4] Zhou, H., Skolnick, J. (2011), GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction, *Biophys. J.*, **101:** 2043-2052.

[5] Skolnick, J. (2006), In quest of an empirical potential for protein structure prediction, *Curr. Opin. Struct. Biol.*, **16:** 166-171.

[6] Summa, C., Levitt, M. (2007), Near-native structure refinement using in vacuo energy minimization, *Proc. Natl. Acad. Sci.(USA)*, **104:** 3177-3182.

[7] Zhu, J., Fan, H., Peiole, X., Honig, B., Mark, A. (2008), Refining homology models by combining replica-exchange molecular dynamics and statistical potentials, *Proteins: Struct. Funct. Bioinf.*, **72:** 1171-1188.

[8] Chopra, G., Kalisman, N., Levitt, M. (2010), Consistent refinement of submitted models at CASP using a knowledge-based potential, *Proteins: Struct. Funct. Bioinf.*, **78:** 2668-2678.

[9] Amautova, Y., Scheraga, H. (2008), Use of decoys to optimize an all-atom forcefield including hydration, *Biophys. J.*, **95:** 2434-2449.

[10] Bhattachary, D., Cheng, J. (2013), 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic level refinement, *Proteins: Struct. Funct. Bioinf.*, **81:** 119-131.

[11] Rohl, C., Strauss, C., Misura, K., Baker, D. (2004), Protein structure prediction using Rosetta, *Methods Enzymol*, **383:** 66-93.

[12] Zhang, Y., Kolinski, A., Skolnick, J. (2003), Touchstone II: A new approach to ab initio protein structure prediction, *Biophys. J.*, **85:** 1145-1164.

[13] Benkert, P., Tosatto, S., Schomburg, D. (2008), QMEAN: A comprehensive scoring function for model quality assessment, *Proteins: Struct Func Bioinfo*, **71:** 261-277.

[14] Zhang, Y., Skolnick, J. (2004), Automated structure prediction of weakly homologous proteins on a genomic scale, *Proc. Natl. Acad. Sci.(USA)*, **101:** 7594-7599.

[15] Rajgaria, R., McAllister, S., Floudas, C. (2006), A novel high resolution Ca-Ca distance dependent force field based on a high quality decoy set, *Proteins: Struct. Funct. Bioinf.*, **65:** 726-741.

[16] Carlsen, M., Røgen, P. (2015), Protein structure refinement by optimization, *Proteins: Struct. Funct. Bioinf.*, **83:** 1616-1624.

[17] McLachlan, A. (1979), Gene duplications in the structural evolution of chymotrypsin, *J. Mol. Biol.*, **128:** 49-80.

[18] Horn, B. (1987), Closed form solution of absolute orientation using unit quaternions, *J Opt Soc Am*, **4:** 629-642.

[19] Coutsias, E., Seok, C., Dill, K. (2004), Using quaternions to calculate RMSD, *J Comp Chem*, **25:** 1849-1857.

[20] Kaindl, K., Steipe, B. (1997), Metric properties of the root-mean square deviation of vector sets, *Acta Cryst A*, **53**: 809.

[21] Horn, R.A., Johnson, C.R. (1990), Norms for Vectors and Matrices, *Cambridge, England: Cambridge University Press.*

[22] Røgen, P., Koehl, P. (2013), Extracting knowledge from protein structure geometry, *Proteins: Struct. Funct. Bioinf.*, **81:** 841-851.

[23] Hollig, K., Horner, J. (2013), Approximation and Modeling with B-Splines, *siam, USA*.

[24] Loose, C., Klepeis, J.L., Floudas, C.A. (2004), A new pairwise folding potential based on improved decoy generation and side-chain packing. *Proteins: Struct. Funct. Bioinf.*, **54:** 303-314.

[25] Tobi, D., Elber, R.(2000), Distance-dependent, pair potential for protein folding: results from linear optimization, *Proteins: Struct. Funct. Bioinf.*, **41:** 40-46.

[26] Hinds, D.A., Levitt, M.(1994), Exploring conformational space with a simple lattice model for protein structure, *J. Mol. Biol.*, **243:** 668-682.

Figure 1: The eight cubic B-spline basis functions B_1,..., B_8 used in the knowledge-based potentials.

Figure 2: Distribution of correlation coefficient between energy and RMSD for all 150 test cases.

Figure 3: Energy-RMSD plot for two test cases with two methods LPKP^1 and LPKP^2.

Figure 4: The effect of number of decoys in the training set on native structure detection and logarithmic function.

Table 1: The bin width distances for eight B-spline basis functions.

ID	C_α Distance (Å)
1	2.2 - 4.6
2	2.8 - 5.2
3	3.4 - 5.8
4	4.0 - 6.4
5	4.6 - 7.0
6	5.2 - 7.6
7	5.8 - 8.2
8	6.4 - 8.8

Table 2: Properties of the different protein decoy sets used in this study.

Decoy set	Nprot [a]	Nres [b]	Ndecoys [c]	RMSD
Titan-HRD	1220	111.80	500	2.46
Titan-HRD*	150	103.95	500	2.51

Training set (Titan-HRD) and test set (Titan-HRD*) from the Titan high resolution decoy set at http://titan.princeton.edu/HRDecoys/. RMSD is the distance measures between the decoys and the corresponding native structures averaged over all decoys and all proteins.

[a] Number of proteins in each set.

[b] The average number of residues in each set.

[c] The average number of decoys in each set.

Table 3: Testing force fields on 150 proteins of the Titan-HRD decoy set.

Method Name	Ave Rank	No of Firsts	Ave RMSD
LPKP[1]	1.67	130 (86.67%)	2.291
LPKP[2]	2.83	124 (82.67%)	2.291
HR[a]	1.87	113 (75.33%)	0.451
LKF[b]	39.45	17 (11.33%)	1.721
TE13[c]	19.94	92 (62.16%)	0.813
HL[d]	44.93	70 (46.67%)	1.092

LPKP[1] and LPKP[2] are the results of schemes (5) and (9), respectively.

[a] is extracted from [15].

[b] is extracted from [24].

[c] is extracted from [25] and TE13 force field was only tested on 148 cases.

[d] is extracted from [26].

Table 4: The effect of number of decoys in training set on native structure detection.

Number of Decoys	No of Firsts
15	52
25	110
35	128
45	130
50	130

Number of decoys for each protein in the training set.

No of firsts is the number of native structures detected in all the test set proteins.

Table 5: Assessing the best decoys selected by energy functions on Titan-HRD dataset.

	Best	**PPD**[a]	**PPE**[a]	**LPKP**[1]	**LPKP**[2]
Titan-HRD	1.11	1.70	1.60	1.62	1.81

Average value over the test set.

LPKP[1] and LPKP[2] are the results of schemes (5) and (9), respectively.

[a] is extracted from [1].

Table 6: Rankings of the native structures using the LPKP[1] and LPKP[2] on Titan-HRD test set.

ID	LPKP[1]	LPKP[2]	ID	LPKP[1]	LPKP[2]	ID	LPKP[1]	LPKP[2]	ID	LPKP[1]	LPKP[2]
1em9A	1	1	1fb1A	1	1	1chd-	1	1	1faq-	1	1
1iioA	1	1	1hf9A	1	2	2drpA	1	1	1b0yA	1	1
1qqvA	1	1	1b1bA	1	1	1ci5A	1	1	1tmy-	1	2
1b9lA	1	1	1bik-	1	1	1k3bB	1	1	1gd5A	1	1
1ap0-	1	1	1hgvA	39	35	1fxkC	1	1	1hd0A	1	1
1ghc-	1	1	1tyfA	1	1	2ech-	1	1	1aplD	1	1
1u2fA	1	1	1ag4-	1	1	1af8-	1	1	3monB	1	4
1eptB	2	1	1fe6A	1	1	1cjgA	1	12	1quqB	1	2
1aq3A	1	4	1b2iA	2	1	1g10A	4	1	1eqiA	1	1
1c3kA	1	1	1cm0B	1	1	1n72A	1	1	1cl3A	1	1
1l0oC	1	2	1dpuA	1	1	1k5yR	1	23	1imt-	1	1
1etpA	1	1	1b2pA	1	1	1eal-	1	1	1rof-	1	1
1jq4A	1	1	1ahjA	1	1	1jpyA	1	1	1cmaA	1	6
1a10I	1	1	1hjrA	1	1	1qc7A	1	1	1d7bA	1	1
1lcl-	1	1	1hks-	1	1	1bccH	1	3	1a14H	1	1
1hlb-	1	1	1ec5A	1	1	1ndoB	1	1	1qckA	1	1
1gnf-	1	1	1c6vX	1	1	1k99A	9	1	1irqA	3	2
1lghB	1	39	1c7uA	3	1	1i7kA	1	1	1scjB	1	1
1hnr-	1	1	1ecsA	3	5	1b44D	1	1	1ai9A	1	1
1bpr-	1	1	1dujA	1	1	1g31A	1	1	1auz-	1	1
1dax-	3	1	1aalA	1	1	1i8nA	1	21	1exg-	2	1
1b4rA	1	1	1aiw-	1	1	1hp8-	1	1	1qgeE	1	1
1a2b-	1	1	1xbd-	7	1	1kbhA	1	1	1dc7A	1	1
1a2kA	1	4	1ab1-	1	1	1cqkA	1	1	1j5eP	1	1
1o7bT	1	1	1be9A	1	1	1b4uA	1	1	1bdyA	1	1
1j7dB	1	1	1occE	1	1	1cqqA	1	1	1ly7A	1	1
1jacA	1	1	1hs7A	1	3	1ibxB	1	1	1jajA	1	1
1occJ	1	2	1qj8A	1	1	1occH	1	6	2sob-	1	1
1tafB	1	1	1b6q-	1	1	1akjD	2	1	1a4aA	1	1
1ha8A	1	1	1olgA	21	42	1icfA	1	1	1a4yB	1	1
1yuf-	6	13	1qjzA	1	1	1jy2N	2	36	1csbA	2	1
1dhn-	1	1	1cdzA	1	12	1qkfA	1	1	1hbiA	1	1
1f7lA	1	1	1pcfA	1	1	1g84A	2	1	1jh3A	1	1
1cfaA	1	1	1kilE	1	1	1fpzA	1	1	1ehxA	1	1
1b01A	3	13	1hykA	2	1	1am9A	1	1	1a6l-	1	1
1dk7A	1	1	1hyp-	1	4	1amx-	1	1	1jhcA	1	1
1perL	1	4	1qmtA	1	1	1cg5B	1	1	3lriA	4	1
1fw9A	17	1	1gd7A	1	1						
For	150		150								
Ave	1.67		2.83								
First	130		124								

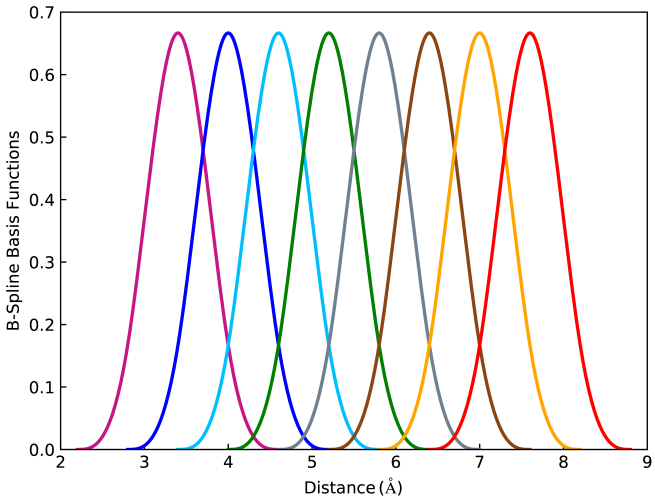

Figure 1: The eight cubic B-spline basis functions B_1,..., B_8 used in the knowledge-based potentials.

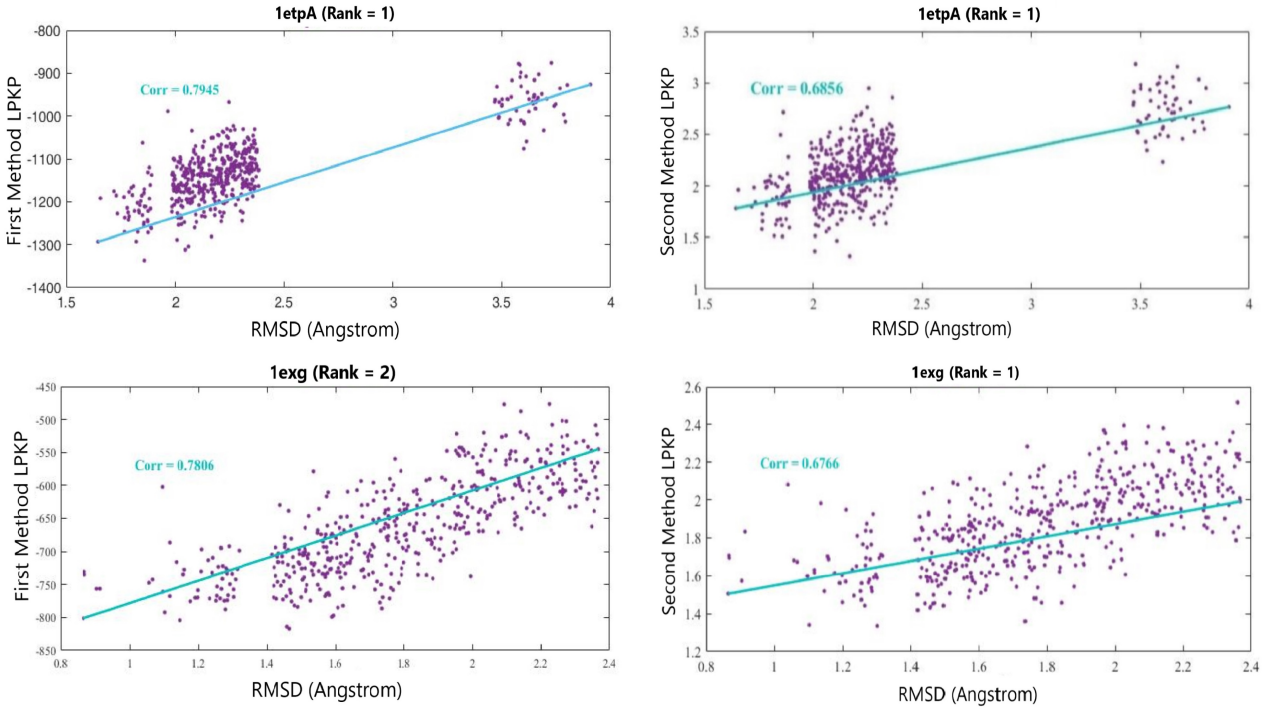

Figure 2: Distribution of correlation coefficient between energy and RMSD for all 150 test cases.

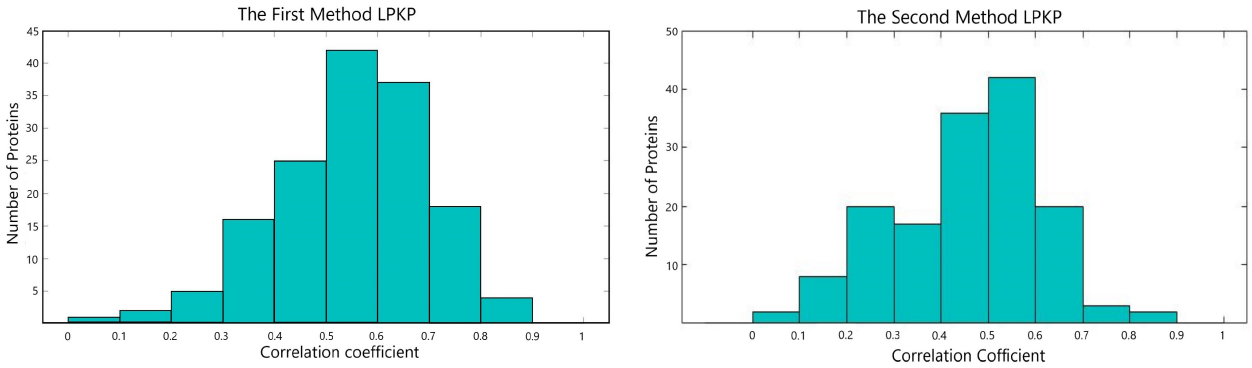

Figure 3: Energy-RMSD plot for two test cases with two methods LPKP[1] and LPKP[2]

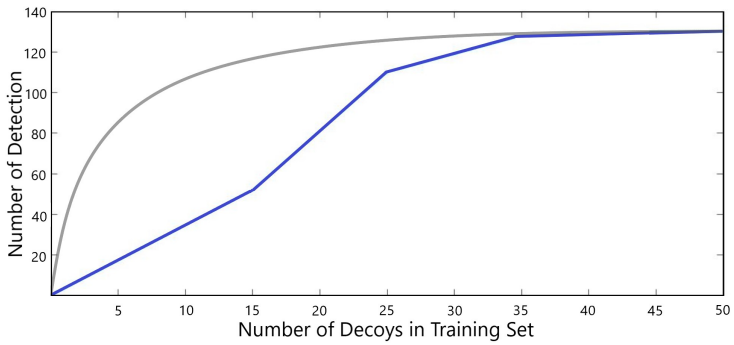

Figure 4: The effect of number of decoys in the training set on native structure detection and logarithmic function