2011

Extraction of an Effective Pairwise Potential for Amino Acids

Jie Luo
Iowa State University

Extraction of an effective pairwise potential for amino acids

by

Jie Luo

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Bioinformatics and Computational Biology

Program of Study Committee:

Xueyu Song, Co-major Professor

Robert Jernigan, Co-major Professor

Mei Hong

Iowa State University

Ames, Iowa

2011

TABLE OF CONTENTS

Abstract

Key to successful protein structure prediction is a potential that recognizes the native state from misfolded structures. In this thesis, we introduced a novel way to extract interaction potential functions between the 20 types of amino acids, which used the Modified Hypenetted Chain (MHNC) and the Reverse Monte-Carlo (RMC) method. We extract Radial Distribution Functions (RDFs) from 996 known protein crystal structures from the Protein Data Bank, and using these RDFs we were able to first generate the potential-of-mean-force (PMF) for different pairs of residues, and then we improved these PMFs by including the higher order terms of the Ornstein-Zernike equation using an iteration that starting from the HNC approximation for the pair interaction potential, and in each of the follow step, we conducted Monte-Carlo simulations to generate the RDFs for the updated potential. The updated potentials in each iteration step can be generated either using MHNC or the RMC method. These effective pairwise potentials were then summed up to obtain the total energy score for known protein structures, and their effectiveness was validated by conducting single and multiple decoy set tests using the 'R' Us decoy set.

Chapter I. Introduction

Proteins are the most important biomolecules for biologists. A well defined protein potential function is useful to solve many important protein structure problems. For example, current prediction approaches to protein structure are based on the thermodynamic hypothesis that the native structure is at the lowest free energy state under physiological conditions [1]. A potential that can discriminate between the native and misfolded structures is crucial for any protein structure approaches to be successful.

It is generally accepted that native conformations of proteins correspond to the structures of lowest free energy. As a result, successful potential functions, including most of those based on native structures, should give the lowest free energy for the native conformations. However, it has been shown that classical semi-empirical potentials such as CHARMM [2], cannot always distinguish the non-native folds of proteins from their native structures. Novotny and co-workers [3] also demonstrated that a conventional molecular mechanics potential cannot accurately discriminate native protein structures from misfolded ones. Therefore, developing such a potential (or scoring function) that could successfully discriminate between native structure/non-native structure or correct configurations/incorrect configurations, is still remained a difficult task.

For years, various algorithms have been developed to construct the protein potential energy prediction models. Two different types of potential energy functions are currently in use [4-9]. The first class of potentials, the so-called physical-based potential, is based on the fundamental analysis of forces between atoms [2, 10]. For example, there is the so-called molecular mechanics potential energy functions (MM-PEFs), which incorporate both the ‘bonded’ and ‘non-bonded’ terms. The bonded terms apply to sets of four atoms that are covalently linked, and they serve to constrain bond lengths and angles near equilibrium values. The bonded terms also include a torsional potential that models the periodical energy barriers encountered during bond rotation. The non-bonded terms consist of the Lennard-Jones (LJ) function (which includes van der Waals attraction and repulsion owing to orbital overlap) and Coulomb’s law. The parameters of the bonded and non-bonded terms of an MM-PEF are derived from quantum calculations or from thermodynamic data on a wide range of systems [11, 12]. MM-PEFs have been used predominantly to simulate protein folding and dynamics, but are also used to refine X-ray crystal structures [13].

For physics-based models, the advantage is that they can be derived based on physical laws; the disadvantage is that the calculation of free energy is very difficult because this computation should include an atomic description of the protein and the surrounding solvent. Currently this type of computation is generally still too expensive for structure predictions.

The second class, the so-called knowledge-based potentials, extracts parameters from experimentally solved protein structures. This type of energy function is derived from the database of known protein structures [14, 15, 16]. The probabilities that atomic groups/residue appears in specific configurations or the probabilities that pairs of atomic groups/residuals appear together in a defined relative geometry are calculated. These probabilities are then converted into an effective potential energy function using the Boltzmann probability equation, which will be discussed in more details later in this chapter. The advantage of knowledge-based energy functions is that they can model any behaviors seen in known protein crystal structures, even if a good physical understanding of the behavior does not exist. The disadvantage is that these energy functions are phenomenological and cannot predict new behaviors absent from the training set. Since most knowledge-based models could avoid *ab initio* and atomic level calculations for structure prediction, therefore within today’s computer resources, knowledge-based potentials are generally easier to be used for folding recognition, compared to the physics-based potentials.

While the physics based protein potentials have become fairly standardized, knowledge-based design potentials vary enormously between laboratories [13]. The various terms are typically calibrated and weighted to optimize performance for one type of prediction, such as experimental binding energy [17, 18], or in some other cases, used to produce native-like sequences when redesigning natural proteins [19]. For example, in Dwyer *et al* ‘s paper [20] a de novo triosephosphate isomerase activity was designed using an accurate electrostatics model which included multiple geometry-dependent dielectric constants [21]. Another example is the 93-residue protein with a new α/β fold designed by Kuhlman *et al* [19]. In their potential energy function, an LJ term (with well depths from CHARMM19 and radii fit to match the distribution of distances seen in the PDB) was included, together with a Lazaridis-Karplus empirical solvation term [22], a knowledge-based hydrogen-bonding term [23], a knowledge-based rotamer term and a knowledge-based pairwise interaction term. The scaling factors for each term were adjusted in order to optimize the native sequences when redesigning a training set of 30 proteins.

Some other efforts in knowledge-based potential design include Crippen [24] and Maiorov & Crippen [25], who tried to empirically fit a set of parameters that corresponded to potential energies between certain residue groups which separated on different distances. In their work it was actually shown that the total potential energy of the native structures are lower than the non-native alternatives. Luthy *et al* also developed an empirical method to evaluate the correctness of protein models [26].

For both physics-based and knowledge-based potentials, models were built on different scales. Normally there are two categories of models in terms of atomic detail complexity: all-atom level models and residue-based ones. All-atom model potentials should normally include the interactions between all the atom types and pairs within a protein structure, while for residue-based models, we reduce the protein structure into units of residues/amino acids, or other types of simplified structural units, depending on how the specific models were constructed. Quite often the distribution of pairwise distances is used to extract a set of effective potentials between residues or atoms. In most cases, the knowledge-based potential is built and then used on reduced protein models, i.e., using one or two points for each residue to represent a protein. These points are usually located at the coordinates of the center of mass or geometric center of each side chain. For example, Zou *et al* constructed a protein-protein interaction model with the structure of a protein represented by 20 different types of atom groups [24]; Zhang *et al* developed a residue-specific, 20 residue types potential which was reduced from an all-atom knowledge-based potential (167 atomic types) based on distance-scaled, finite ideal-gas reference state [25]; and more models have been built based on each amino acid being treated as a structural unit. Our potential, falls into this category as well. The advantage for these residue-based, or reduced atom-groups based models is that it is much easier for us to do the structure reduction calculations, and the actual calculations for the potential. In contrast, those all-atom potential models usually cost much more computer time than the residual based ones, for example, in folding recognition or *ab initio* predictions. Nevertheless, several potentials for higher-resolution models had been developed in the hope of providing better discriminatory power than obtained with residue-based potentials. For ranking structures near the native fold, and for protein structure refinement, the detailed interactions between side-chain atoms from different residues may be required to rank correctly low root-mean-square deviation structures [25-28].

For the knowledge-based potentials, a large category of them falls into the use of radial distribution function (RDF) in order to predict the so-called potential-of-mean-force. In Sippl's work, the potentials-of-mean-force were evaluated as a function of distance for two-body interactions between amino acids in protein structures from the radial distribution of amino acids from known protein native structures [29]. The potentials of mean force for the interactions between C^{β} atoms of all amino acid pairs were used to calculate the conformational energies of amino acid sequences in different folds, and it was found that the total energy of the native state is the lowest among all the other non-native ones [30, 31, 32]. Bryant and Lawrence also estimated the pairwise contact potentials depending on inter-residue distance [33]. In Covell and Jernigan's paper, pair contact energies were demonstrated to discriminate successfully between native-like and incorrectly folded conformations in a lattice study of five small proteins [34].

In the following section, we will discuss this type of knowledge-based protein potential in detail, which uses the knowledge of the radial distribution function of amino acids/atomic groups based on known protein structures. Starting from the radial distribution function, which is defined as

$$g_{ij}(r) = \frac{\rho_{ij}(r)}{\rho_{ij}{}^{*}(r)}, \tag{1}$$

where $\rho_{ij}(r)$ and $\rho_{ij}{}^{*}(r)$ are the number densities of the components *i* and *j* which pair at a distance *r* in the experimental structures and in the reference state, respectively. And then, the potential-of mean-force can be represented as the logarithm of the radial distribution function:

$$u_{ij}(r) = -k_B T ln[g_{ij}(r)], \tag{2}$$

Where k_B denotes the Boltzmann constant, *T* stands for the system temperature, *i* and *j* stand for component *i* and *j* from different part of a protein complex. The total energy score is then the sum of all the inter-residual interaction energies:

$$E_{tot} = \sum_{i,j} u_{ij}(r). \tag{3}$$

This method of constructing potential energy between residual pairs in a protein was first developed by Miyazawa and Jernigan in 1985 [35], and then has been reexamined in 1996, with a significantly larger set of protein crystal structures being used for the knowledge extraction. Also, an additional repulsive packing energy term has been added for the 20 amino acids as a function of the number of contacting residues, based on their observed distribution [36].

For the potential-of-mean-force calculated from the above method, this kind of two-body residue-residue potential shows peaks and valleys that correspond to the radial distribution function that used. And these peaks and valleys appear in the potentials definitely represent certain repulsive/attractive areas with the change of inter-residue distances. However, it should be noted that such a potential does not really reflect the actual repulsion or attraction interaction forces between the pair of residues under study, but instead, it is an “effective” potential in a sense that it also includes the average effect due to other residues/environments that act upon this certain pair of residues.

For years, we have seen different variations of the traditional potentials-of-mean-force coming out. Nishikawa and Matsuo devised an empirical potential that was composed of four terms: side-chain packing, hydration, hydrogen bonding and local conformational potentials [31], with the parameters derived from structures of 101 known proteins, where each of the four terms are summed with weights in the total energy score. Their potential used a slightly modified form of Sippl’s potential [33] for the side-chain packing effects in proteins. All the other terms in their potential were evaluated as potentials of mean force. This function was also demonstrated to be an appropriate measure of the compatibility between sequences and structures of proteins.

As mentioned before, it should be noted here that a two-body residue-residue potential of mean force based on the radial distribution of residues will manifest peaks and valleys as a function of distance, even for hard spheres, which are effects of close residue packing. However, these may not be present in actual interaction potentials. That is, such a potential of mean force reflects not only the actual inter-residue interactions, but also includes the average effects of other residues upon the target residue pair, including those interposed between the target pair and especially the significant effects of residue packing in protein structures. There will be an over-counting if the sum of the potential is taken over all residue pairs. Thus, if the residue-residue potential in a protein is approximated by such a potential of mean force only, the sum of the potential over all residue pairs is unlikely to yield the correct value for the total residue-residue interaction energy. In addition, even though these effective potentials have the important characteristics of low energy values for the native folds of proteins, they are unlikely to succeed in representing the actual potential surface far from the native conformation. Therefore, such potentials-of-mean-force may not be appropriate for applications in study of a wider range of conformations, from the denatured state to the native conformation.

Meanwhile, it should be noted that the various knowledge-based potentials that are based on the potential-of-mean-force, as mentioned above, are estimated with the Bethe approximation, i.e., it is assumed that residue-residue contacts in protein structures are treated the same with those in mixtures of unconnected amino acids and other solvent molecules [39]. The Bethe approximation is a well-known second-order approximation to the mean field approximation that used to describe behaviors in a system consisting of mixtures of multi-component molecular species which interact with each other through chemical bonds, or other forms of interactions [40]. Both the mean-field approximation and the Bethe approximation are used to calculate the partition function for such a mixture system of particles interact with each other.

One problem with the potential-of-mean-force based scoring functions is the lack of consideration of higher order expansion terms. From the original Ornstein-Zernike (OZ) equation [41]

$$g(r)-1=c(r)+\rho\int d^3r'c(r')[g(|\vec{r}-\vec{r}'|)-1], \qquad (4)$$

where r is the distance between two amino acids, ρ is the density of amino acids in question, g(r) is the radial distribution function, and $c(r)$ is the direct correlation function, and Eq.(4) can be also written as

$$h(r_{12}) = c(r_{12}) + \rho \int d\overrightarrow{r_3} c(r_{13}) h(r_{23}), \qquad (5)$$

which means that the total correlation $h(r_{12})$ between particles 1 and 2 can be written as a sum of the direct correlation $c(r_{12})$ (that comes from the interaction between particle 1 and particle 2) and the indirect correlation $\rho \int d\overrightarrow{r_3} c(r_{13}) h(r_{23})$, that represents the sum of interactions between particle 1 and all other particles in the space. Now if we write this indirect term as $g_{indirect}(r)$, we will have

$$c(r_{12}) = g_{tot}(r) - g_{indirect}(r). \qquad (6)$$

As mentioned above, we could use the potential-of-mean-force as the estimation for $g_{tot}(r)$:

$$g_{tot}(r) = \exp[-\beta w(r)], \tag{7}$$

where $w(r)$ is the potential-of-mean-force that introduced before. And then, the indirect correlation $g_{indirect}(r)$ becomes

$$g_{indirect}(r) = \exp[-\beta w(r) - \beta u(r)], \tag{8}$$

where $u(r)$ is the true interaction potential between the pair of particles that interact. So we can see that our estimation using the potential-of-mean-force is actually ignoring the term of $\exp[-\beta w(r) - \beta u(r)]$.

If we try to conduct expansions to the term $\exp[-\beta w(r) - \beta u(r)]$, (for example, using a Fourier expansion,) we could improve the accuracy of estimation for the potential-of-mean-force. Several similar ways have been developed in order to deal with this problem. In Pliego-Pastrana's work, the potential-of-mean-force has been improved by applying closure relationships such as the hypernetted chain (HNC) approximation, or the Percus-Yevick (PY) approximation:

$$\beta u(r) = g(r) - 1 - c(r) - \ln \mathrm{g}(r), \tag{9}$$

$$\beta u(r) = \ln\left\{1 - \frac{c(r)}{[h(r) + 1]}\right\}, \tag{10}$$

By doing this, the effective pairs potential *u(r)* can be improved, compared to the potential-of-mean-force which simply take the logarithm of *g(r)* as an estimation of the potential. Pliego-Pastrana *et al* used this idea to compute the effective potentials between amino acid residues in 2003 [41]. These estimated pair potentials, which used the above closure relationships, could better reflect the thermodynamic properties of the system in contrast to the potential of mean force, by including more accentuated sensitivity of the pair interaction potential to the variation of thermodynamic states.

This method has been proved to be useful to characterize some quite different systems, for example, the pairwise interaction among colloidal particles [42, 43]. When applied to the problem of effective pairwise interactions between amino acids, it was showed by Pliego-Pastrana that this method could be able to describe the characteristic lengths in the formation of α and β secondary structures for alanine and glysine [45].

While Pliego-Pastrana's method has improved the potential-of-mean-force to a certain extent, the ignored part is the bridge function (which is hard to be estimated analytically). In this thesis, we will be introducing a new way that further improves the effective pair potential prediction between amino acids, which includes the effects of the bridge function by using an iterative predictor-corrector procedure or a Reverse Monte-Carlo (RMC) calculation. Two methods are given in our work for the iteration step: one is the traditional MHNC method with the predictor-corrector algorithm; and another will be the Reverse Monte Carlo method. The theory and application of these two methods will be introduced in detail in Part II of this thesis; the training protein structure data sets and the procedure to calculate radial distribution functions between amino acids will also be discussed in Part II. The calculated pair potentials will be presented in Part III, together with the results for decoy set tests of the whole protein energy scores; some comments and discussions will be made in Part IV; finally, in Part V, a summary will be given for the entire work.

Chapter II. Materials & Methods

i) Preparation of the training data set

In our study, the proteins that used to collect the pairwise distribution data are from the Protein Data Bank (PDB) [46], and they all satisfy the conditions that:

1) All protein structures used are determined by X-ray analysis with resolution equals to or better than 2.5Å. All protein structures determined by NMR are excluded.

2) Our study is based on high molecular weight, all proteins that used contain at least 1000 amino acids. (The reason for this criteria is that only systems with a large number of elements are expected to attain thermodynamic equilibrium.)

The total number of protein structures used in our calculation is 996. Our list of proteins includes hydrolases, oxidoreductases, atpases, groels, etc. This dataset has been pre-selected so that proteins with redundant/similar sequences have been removed. There are two ways we could do this: 1) using protein representatives that are sufficiently dissimilar to each other in their sequences; or 2) using a different statistical weight for each protein related to its extent of similarity to other sequences. So far, most statistical analyses have used a representative set of proteins. Usually, protein representatives are chosen by specifying an upper limit for sequence identity [47] or structural similarity [48, 49]. However, it is not clear what value is best as an upper limit of similarity in protein representatives. Also, in such a method, many good structures may be discarded. In this work, the second approach has been taken.

ii) Calculation of the Radial Distribution Function

In statistical mechanics, the radial distribution function (RDF) *g(r)* describes how the particles' density varies as a function of the distance from one tagged particle. More precisely, if there is a particle at the origin *O*, and if *n* = *N/V* is the average number density, then the local density at distance *r* from *O* is *ng(r)*.

In this study, we obtained pairwise distribution functions *g(r)* from the 996 structurally distinct proteins, as described above in the previous section. For each protein, we assumed the positions of the centroids of the *N* residues located inside the sphere of volume *V* (the position and size of the sphere are such that big voids are minimized). The corresponding number density is then *ρ*=*N/V*. Pairwise correlation functions of individual proteins were computed on the understanding that $\rho g(r) 4\pi r^2 dr$ is the number of residues between two concentric spheres of radii *r* and *r*+*dr*, respectively, about a central residue [50]. The spatial resolution *dr* was estimated to be 0.2 Å, considering the uncertainties in centroids' coordinates.

Although it is possible to provide accurate approximations to get the effective pair potential *u(r)*, the radial distribution function (RDF) has to be determined with enough precision to minimize errors induced by statistical noise. Thus, to improve the statistics, we averaged the results from proteins of rather close number densities.

In order to calculate *g(r)*, we use the equation,

$$g_{\gamma\mu}(r) = \frac{1}{\rho N \chi_\gamma \chi_\mu} \left\langle \sum_{i=1}^{N_\gamma} \sum_{j=1}^{N_\mu} \delta(r - |\vec{r}_i - \vec{r}_j|) \right\rangle, \qquad (1)$$

where the indices γ and μ refer to species 1 and 2 (for instance, Glycine and Alanine). $\chi_\gamma = N\gamma /N$ is the residue's ratio for each components in the mixture, where $N\gamma$ is the number of particles of species γ, and $N=N\gamma+N\mu$. The angular parentheses denotes an ensemble average over all proteins to be sampled, while r_i is the position of the geometric center of residue i, and $\delta(r)$ is Dirac's delta function. The number density is $\rho=N/V$ with V being the total volume, and this density can actually be estimated by normalizing the above calculated factor $\frac{1}{N\chi_\gamma\chi_\mu}\langle\sum_{i=1}^{N_\gamma}\sum_{j=1}^{N_\mu}\delta(r - |\vec{r_i} - \vec{r_j}|)\rangle$; and However, Eq.(1) is valid only for the case of infinite systems. In order to obtain bulk-like properties from systems of a finite size but large enough to extract a structural or thermodynamic property, an additional normalization procedure could be applied as discussed in [45].

iii) Theory and Method of Predictor-Corrector MHNC

After the radial distribution functions (RDF) have been derived, we could use these RDFs to specify the pair interaction functions between particles. The determination of the inter-particle interaction in the condensed phases of matter is of fundamental importance and this is the so-called inverse-problem, i.e., the deduction of the inter-particle interactions starting from measured structural data as obtained from scattering experiments. It is believed that in a monatomic liquid, there is a one-to-one correspondence between the structure factor for density fluctuations $S(q) =< \rho_q\rho_{-q} >/N$ **(**where ρ_q is the q component of the microscopic density fluctuation) and the pairwise interaction $u(r)$. If the system many-body forces are present, this interaction $u(r)$ will serve as an "effective" two-body interaction which includes the effect of many-body interactions, and it will be state dependent.

In our work, the main task is to generate the effective pairwise potential functions between pairwise amino acids from the known g(r) of protein structures. The history of this inverse problem can be traced back to Johnson, Hutchinson and March [51], and since then there have been different theoretical methods developed to solve this problem, and widely different results have been obtained from the same data [52, 53]. It has become evident that scattering data of very high precision, at least of order of 1% in absolute accuracy, are required over a wide range of moment transfer q. Since in a dense fluid, the RDF g(r) is very insensitive to the exact shape of the pair potential $u(r)$, therefore, the exact pair potential solution in the inverse problem still remains a hard problem to solve [50].

The simulation of model fluids is ideally suited to test whether a theory is adequate for this purpose: Using the RDF obtained from a simulation one should be able to recover the interaction potential used in that computation. However, since the simulation results are statistical in nature, this is a meaningful test only if the statistical noise of simulation is small enough. Therefore, we need to conduct simulations that can give RDFs accurate enough to test theories. In the following, we will introduce an inversion scheme that has been proposed by Reatto *et al* [57] and has been applied to a related problem in the Jastrow theory of Bose quantum fluids [54], which was shown to be successful in the extraction of the pairwise interaction with good accuracy. This scheme is based on the modified hypernetted chain (MHNC) equation and on simulation.

The method starts from the Ornstein-Zernike equation [50]:

$$g(r) - 1 = c(r) + \rho \int d^3r' c(r')[g(|\vec{r} - \vec{r}'|) - 1], \tag{2}$$

where r is the distance between two amino acids, ρ is the density of amino acids in question, and $c(r)$ is the direct correlation function. The pairwise potential $u(r)$ between two amino acids can be found from the MHNC equation

$$\beta u(r) = g(r) - 1 - c(r) - \ln \mathrm{g}(r) + E[r; u(r)], \tag{3}$$

where β is the inverse dimensionless temperature, and $E[r;V(r)]$ is the bridge function. In Eq.(3), in order to get $u(r)$, the only missing part is the bridge function $E[r;V(r)]$. Bridge functions do not have analytical expressions and they have to be approximated with certain closure relations. For example, $E[r;V(r)]=0$ is known as the Hypernetted Closure and $E[r;V(r)]=c(r)-g(r)+1+\ln[g(r)-c(r)]$ leads to the Percus-Yevick (PY) approximation. As mentioned in the previous chapter, works by Pliego-Pastrana *et al* used the above two closure relations to solve Eq.(3), in order to estimate the effective pair potentials between two alanines, two glycines, and pairwise interaction potential between an alanine and a glycine [41, 45]. The results provided by these two approximations are, as reported in [45], satisfactory but not accurate. Now we want to use a predictor-corrector algorithm to improve the estimation for the bridge function. (Another method, which will be using the Reverse Monte-Carlo method, will be discussed in the next section.)

The predictor-corrector approach was initially introduced by Reato *et al* [54] to solve a problem in the theory of Bose quantum fluids and found to converge, and then extended to dense classical liquids [57]. In the following, we will briefly introduce the theory of this method.

Let in the i^{th}-iteration step we know the pairwise interaction potential for the i^{th} iteration to be $u_i(r)$, then the pair potential at the $i+1^{th}$ step can be found by

$$\beta u_{i+1}(r) = g(r) - 1 - c(r) - \ln \mathrm{g}(r) + E[r; u_{i+1}(r)], \tag{4}$$

where the bridge function for the i^{th} iteration $E[r; u_i(r)]$ is found by

$$E[r; u_i(r)] = \beta u_i(r) - g_i(r) + 1 + c_i(r) + \ln[g_i(r)]. \tag{5}$$

The correlation function $g_i(r)$ on the right-hand side of Eq.(5) for the given pair potential $u_i(r)$ will be generated by Monte-Carlo simulation, and the direct correlation function $c_i(r)$ will be calculated as a solution of the Ornstein-Zernike (OZ) equation after a Fourier transformation of (2) [50]:

$$\tilde{c}_\imath(k) = \frac{\tilde{h}_\imath(k)}{1 + \rho \tilde{h}_\imath(k)}. \tag{6}$$

where $\tilde{h}_i(k)$ is the Fourier transformation of the correlation function $g_i(r) - 1$. Now if we conduct reverse-Fourier transformation for $\widetilde{c}_i(k)$, we will get $c_i(r)$ that can be readily used in Eq.(5).

At the beginning of this iteration, we will use the potential-of-mean-force as the starting point; and the radial distribution function $g(r)$ for the starting potential-of-mean-force can be calculated by using Monte-Carlo simulation as well. And then the pair potential for the following step will be estimated using Eq.(4). Repeat this procedure until in two consecutive steps, the $g(r)$ computed become converged to each other within a certain tolerance, i.e., $\|g_{i+1}(r) - g_i(r)\| < \varepsilon$. As long as the radial distribution function $g(r)$ converges, the effective pair potential $u(r)$ will converge as well. And the pair potential $u_{i+1}(r)$ which gives the converging $g(r)$ will be treated as the final effective pair potential we want to estimate.

iv) Reverse Monte-Carlo Method (RMC)

Another approach to solve this problem is by using the reverse Monte-Carlo method. This method also belongs to the general category of solving the inverse problem to get the interaction potential in atomic and molecular systems. Also, it can be applied to more complex systems such as bimolecular systems and organic molecular systems. This method also starts from the radial distribution function $g(r)$, which could be obtained from the experimental structural data, as has been previously discussed. No input potential is required for this method, and the simulation is carried out to minimize differences between calculated and reference averages.

The main objective the Reverse Monte-Carlo method is to provide a method to reconstruct the Hamiltonian from radial distribution functions (RDF). In general, the solution of this problem is not unique; however, if we consider a limited class of Hamiltonians (e.g., those represented by a sum of pair interactions), the solution will be well defined. In the following, we will present this method of automatic adjustment of the pairwise interaction potential, irrespective of its analytical form, to known radial distribution functions.

The idea of this method goes back to the renormalization group Monte-Carlo method for phase transition studies in the Ising model by Swendsen and co-workers [58, 59]. This algorithm was first used to extract the interaction potential for the blocked spins, and now it is shown that the applications of this method could be generalized to a much broader type of systems. It will be shown below that it is possible to renormalize the Hamiltonian of a molecular system of interest [60], and therefore the pair interaction potential could be reconstructed.

Consider a system with a Hamiltonian (potential energy) given as

$$H\{q_i\} = \sum_{\alpha} u_\alpha S_\alpha \{q_i\}, \tag{7}$$

where $S_\alpha\{q_i\}$ are functions of particle coordinates q_i, and u_α are constants which construct the pair interaction potential in the distance section α. The summation in Eq.(7) may also be represented by an integral.

The Hamiltonian of a system with pair interactions can be therefore given as Eq.(7):

$$H\{q_i\} = \sum_{i,k} \Psi\{|q_i - q_k|\} = \sum_{i,k} \int_0^\infty \Psi(\mathrm{r})\delta(\mathrm{r} - |q_i - q_k|)\,\mathrm{dr} \tag{8}$$

$$= \int_0^\infty \Psi(\mathrm{r}) \textstyle\sum_{i,k} \delta(\mathrm{r} - |q_i - q_k|)\, dr.$$

In comparison with Eq.(7), the sum is now replaced by an integral, α is replaced by r, u_α is replaced by Ψ(r), and $S_\alpha\{q_i\}$ replaced by $\sum_{i,k} \delta(\mathrm{r} - |q_i - q_k|)$.

Generalization to particle mixtures is straightforward. We could easily extend the Hamiltonian given in Eq.(7) for systems with three-particle interactions in a similar fashion.

The Hamiltonian in Eq.(7) is defined by a set of parameters u_α . These parameters span a space of Hamiltonians determined by the structural factor $S_\alpha\{q_i\}$, which basically tells how many particles there are in each grid of the coordinates. These Hamiltonians may be considered as equivalent if they have the same canonical averages $\langle S_\alpha\{q_i\} \rangle$ for each α. For systems defined by pair interactions [Eq.(8)], this coincides with the radial distribution functions $g(r)$, due to the fact that $\langle S_r \rangle = 4\pi r^2 g(r)$. The averages $\langle S_\alpha \rangle$ are functions of constants $\{u_\alpha\}$ from the ensemble average $\langle S_\alpha \rangle = S^*_\alpha$. The averages can be calculated from computer simulations (as in our case, the Monte-Carlo simulations) of the whole system.

In the vicinity of an arbitrary point in the space of Hamiltonians, $\{u_\alpha\}$, we can write

$$\Delta\langle S_\alpha \rangle = \sum_\gamma \frac{\partial \langle S_\alpha \rangle}{\partial u_\gamma} \ \Delta u_\gamma + O(\Delta u^2), \qquad (9)$$

where the derivative $\frac{\partial \langle S_\alpha \rangle}{\partial u_\gamma}$ can be further calculated as

$$\begin{aligned} \frac{\partial \langle S_\alpha \rangle}{\partial u_\gamma} &= \frac{\partial}{\partial u_\gamma} \left(\frac{\int dq S_\alpha(q) \exp(-\beta \sum_\lambda u_\lambda S_\lambda(q))}{\int dqexp(-\beta \sum_\lambda u_\lambda S_\lambda(q))} \right), \\ &= -\beta(\langle S_\alpha S_\gamma \rangle - \langle S_\alpha \rangle \langle S_\gamma \rangle) \end{aligned} \qquad (10)$$

and q is the set of degrees of freedom of the reduced system.

Let $u^{(0)}_\alpha$ denote a set of starting values for the parameters u_α for the potential. By carrying out a MC simulation using these values $u^{(0)}_\alpha$, a set of ensemble averages for the structural factor $\langle S^{(0)}_\alpha \rangle$ can be collected in the end of the simulation. The differences between the starting values of $\langle S^{(0)}_\alpha \rangle$ and the reference values are $\Delta\langle S_\alpha \rangle^{(0)} = \langle S^{(0)}_\alpha \rangle - S^*_\alpha$. Then, by solving a set of linear equations for each coordination grid γ as given in Eq.(9), with appropriate coefficients calculated from Eq.(10), and by omitting terms of order $O(\Delta u^2)$, we can obtain the differences $\Delta u_\alpha^{(0)}$ and use them as corrections to the starting potential parameters according to Eq.(11):

$$u_\alpha^{(1)} = u_\alpha^{(0)} + \Delta u_\alpha^{(0)}. \tag{11}$$

The MC simulation is then repeated with this new updated potential $u_\alpha^{(1)}$ to determine a set of corrections $\Delta u_\alpha^{(1)}$. The procedure is repeated until convergence is reached, e.g., when the difference $\Delta\langle S_\alpha\rangle$ becomes vanishingly small for each α within the accuracy of the statistical error of the simulation. The algorithm is similar to that which solves the multidimensional nonlinear equations using the Newton-Raphson method [61].

A similar method has been applied to a study of the critical point region in the Ising model in [59]. In that particular case, the number of constants u_α was finite. In fact, it was in the range from 1 to 7. For molecular systems described with pair interaction potentials, the formal number of constants is infinite because of the integral in Eq.(8). For numerical solutions, we can use a finite grid to approximate a continuous function.

Let R_{cut} be the cutoff radius for the interaction potential in the computer simulation. For example, R_{cut} can be chosen as half of the cubic box length. The interval $[0, R_{cut}]$ can be divided into M small slices, with each slice corresponding to a small region around the distance $r_\alpha=\frac{\alpha R_{cut}}{M}, \alpha = 1, \dots, M.$ Then the Hamiltonian of the system of N particles can be written as

$$H = \sum_{\alpha=1}^{M} u(r_\alpha) S_\alpha, \tag{12}$$

where $u(r_\alpha) = \Psi(r_\alpha)$ is the potential parameter value at the distance r_α, and S_α is the number of pairs between the particles within distances around r_α inside the αth piece of slice. In computer simulations, S_α can be normally estimated with the radial distribution function (r) :

$$g(r_\alpha) = \langle S_\alpha \rangle \frac{V}{2\pi r_\alpha^2 N(N-1)} \,. \tag{13}$$

It can be apparently seen from Eq.(13) that if we know the radial distribution function *g(r)*, we are able to compute the assemble averages $\langle S_\alpha \rangle$. As a trial function or an initial approximation to the effective potential function, we can use, for example, the potential-of-mean-force which was discussed in the previous sections:

$$u_\alpha^{(0)} = -kTlng^*(r_\alpha)\,. \tag{14}$$

v) Comparison between Predictor-Corrector MHNC and the Reverse Monte-Carlo Method

In our study, both the predictor-corrector MHNC and the Reverse Monte-Carlo method were used to extract the effective potential between each pairs of the 20 amino acids. For the predictor-corrector MHNC method, the iterations converged in 10 iteration cycles, on average; for the Reverse Monte-Carlo method, the iteration can get converge within 6 cycles, on average. We compared the resulting effective pairwise potentials obtained by both ways and found that the differences between the resulting potentials from the two methods are within a statistical error range (<1%). Therefore, we can conclude that these two methods will lead to the same results for the effective pair potential, but the Reverse Monte-Carlo method is more efficient in terms of computing time.

vi) Total energy score calculation and the decoy sets

After we have obtained a complete set of effective pair potentials between 20 amino acids, we will sum them up to get the total energy score of the protein:

$$total\ energy = \sum_{i,j} u_{ij}(r) \tag{15}$$

Since the native structure of a protein must be the lowest in its free energy compared with all other conformations of the same chain in order to be almost exclusively populated in solution, a stringent test of energy functions is the extent to which they attribute lower energies to native and near native conformations than to non-native conformations. Indeed, “decoy discrimination” tests have become a widely used approach for testing and validating alternative energy models [62, 63, 64].

An optimal decoy set should (1) contain conformations for a wide variety of different proteins to avoid over-fitting; (2) contain conformations close (<6Å) to the native structure because structures more distant from the native structure may not be in the native structure’s energy basin and thus become impossible to recognize; (3) consist of conformations that are at least near local minima of a reasonable scoring function, so they are not trivially excludable based on obviously non-native protein like features; and (4) be produced by a relatively unbiased procedure that does not use information from the native structure during the conformational search. If (4) is the case, then a method that performs well on the decoy set can immediately be used for structure prediction [65].

In our study, we used the Decoys ‘R’ Us decoy set [66] which has a list of decoy structures whose main use is to test energy or score functions for protein structures. These decoys are computer generated conformations of proteins that possess some characteristics of native proteins, but are not biological real proteins. We apply our extracted potentials to the single and multiple decoy sets available in this dataset. Single decoy sets have one correct and one incorrect conformation given for each native protein structure; multiple decoy sets have a list range of conformations with various root mean square deviations (RMSD) from the native structure. The main goal is to distinguish the non-native conformations from the native one. The results for the decoy set tests will be shown in the next chapter.

Chapter III. Results

i) 210 extracted potentials between 20 amino acids

We extracted 210 effective pairwise potentials between 20 amino acids, using the Reverse Monte-Carlo method. (Since comparisons have been made with those generated by the predictor-corrector MHNC method, and the results turned out to be almost identical. So we chose the Reverse Monte-Carlo method, which was much more efficient in terms of computer time.) 12058 amino acids were used during the Monte-Carlo simulation step. For each type of amino acids, the following numbers were used in the simulation: 1084 Alanines, 650 Arginines, 470 Asparagines, 699 Aspartic acids, 145 Cysteines, 831 Glutamic acids, 410 Glutamines, 952 Glycines, 301 Histidines, 687 Isoleucines, 1096 Leucines, 650 Lysines, 241 Methionines, 482 Phenylalanines, 578 Prolines, 662 Serines, 650 Threonines, 169 Tryptophans, 410 Tyrosines and 891 Valines.

The above mentioned amino acid component ratios were obtained according to the corresponding ratios from the protein training data that we used. For the 12058 residues in the MC, we run approximately five days for each Monte-Carlo cycle; and the Reverse Monte-Carlo get converged in 5-6 iterations on average. As a result, 210 effective pair potentials were extracted after we run Reverse Monte-Carlo iterations. Some of the extracted potential were plotted in Figures 1-6.

From these extracted effective potentials, we could see that they share some common properties. 1) A large number of them have their first minima at around r = 3Å region (as in the case of ALA-ALA pairwise potential, see Fig.1). This is due to the fact that the two consecutive amino acids have a distance of 3.8Å in a polypeptide chain. 2) Some potentials do not have the 3Å minima shown, but instead, they have their first minima shown at around r = 5-6 Å region (as in the case of ALA-ARG pairwise potential, see Fig.2). The reason for this is that these certain pairs of amino acids are not very likely to be found next to each other on a polypeptide chain, at least for our training sample pool of protein structural data.

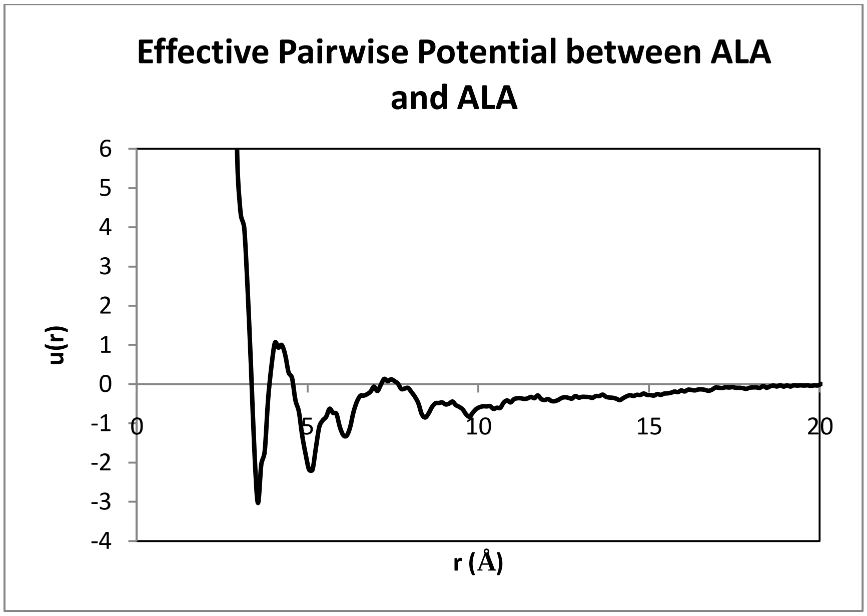

Fig.1. Extracted pairwise potential between ALA and ALA

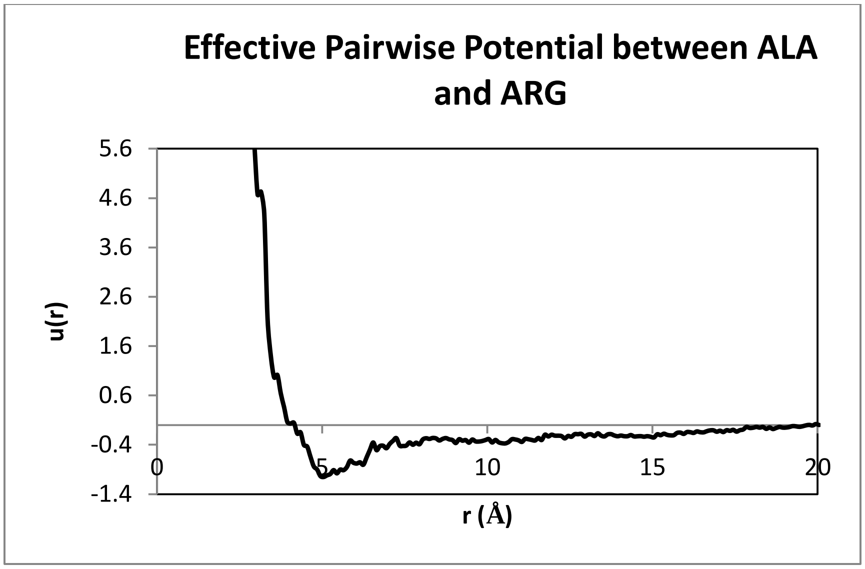

Fig.2. Extracted pairwise potential between ALA and ARG

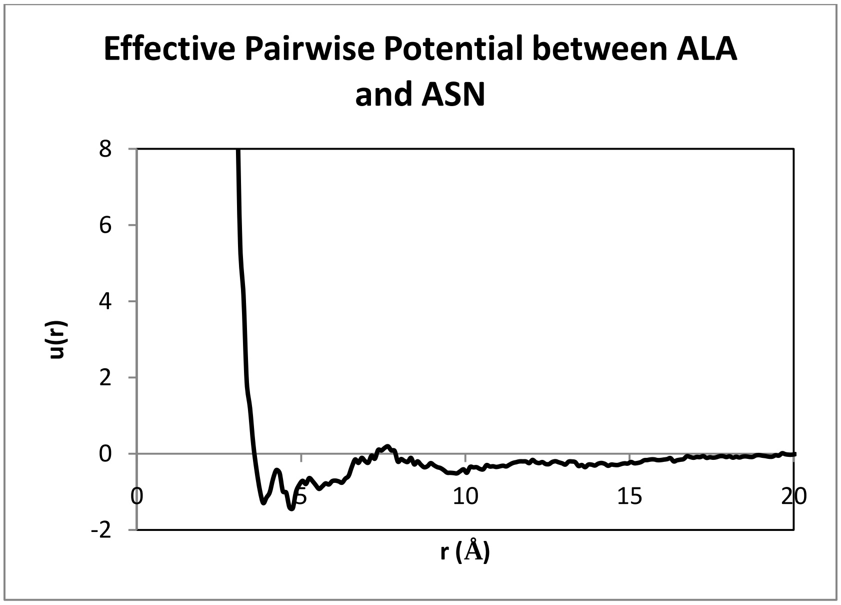

Fig.3. Extracted pairwise potential between ALA and ASN

ii) Calculation of the whole protein potential

After the 210 extracted potentials between each pair of the amino acids were obtained, we sum them up to get the energy of the whole protein:

$$total\ energy = \sum_{i,j} u_{ij}(r) \tag{1}$$

iii) The Decoy Set Test Results

1. Single Decoy Set Test

The single decoy set in the Decoys 'R' Us [66] has misfolded conformations listed for 23 native chains: 1bp2, 1cbn, 1fdx, 1hip, 1lh1, 1p2p, 1ppt, 1rei, 1rhd, 1rn3, 1sn3, 2b5c, 2cdv, 2ci2, 2cro, 2cyp, 2i1b, 2paz, 2ssi, 2tmn, 2ts1, 5pad, 5rxn. The energy scores for the native and the corresponding misfolded conformations are listed in Table1. The energy scores for the native conformations were observed to be consistently lower than the energy scores for the non-native conformations. Using our extracted potential, the native energy scores were lower than the misfolded energy scores for all cases. On average, the total scores of the misfolded conformations were lower than that of the corresponding native conformations by 21.7%.

Protein (PDB id)	Native energy score	Misfolded energy score
1bp2	-593.5	-574.4
1cbn	-1331.5	-1211.6
1fdx	-196.7	-138.6
1hip	-347.8	-337.4
1lhl	-761.8	-359.1
1p2p	-566.7	-505.6
1ppt	-132.0	-102.0

1rei	-1129.7	-1095.8
1rhd	-338.9	-297.2
1rn3	-605.8	-580.5
1sn3	-243.5	-234.5
2b5c	-356.0	-350.2
2cdv	-486.7	-456.2
2ci2	-250.6	-245.3
2cro	-277.8	-256.4
2cyp	-1588.3	-1402.5
2ilb	-721.9	-711.0
1paz	-578.4	-577.0
2ssi	-452.3	-353.5
2tmn	-1729.5	-1515.5
2ts1	-1653.7	-1648.3
5pad	-1085.1	-986.9
5rxn	-861.2	-602.4

Table1. Energy scores for the misfolded single decoy set from the ‘R’ Us database [66]. The native conformation energy scores are lower than their decoys for all case.

2. **Multiple Decoy Set Test**

The Decoys ‘R’ Us decoy set also provides multiple decoy structures for a set of proteins [66]. For each native conformation, there are multiple non-native conformations which fall in a range of root mean square deviations (RSMD) from the native structure. The decoys generated using different methods are classified separately (labeled *lattice_ssft*, *4state_reduced*, *lmds*, *fisa*, and so on). Decoys are generated for a series of native proteins using each method. Rank scores of the native structure among its decoys as well as energy-RSMD plots for the native and decoy structures have been commonly used to test the effectiveness of potential functions. The Rank scores were calculated for all the native and decoy structures based on our extracted potentials. Also, to compare our potential with some of the previously developed potentials, we list the rank scores for the *4state_reduced*, *lattice_ssfit* and *lmds* decoy set calculated using Miyazawa-Jernigan (MJ) potentials as listed in Park and Levitt’s paper [67], and Krishnamoorthy and

Tropsha's four-body potential [68]. We will discuss the comparison results given by our extracted pairwise potential and the other two potentials in the later section.

As shown by the results (Table 1-3), our potential could successfully distinguish the native structures from their decoys in most of the cases. For the *4state_reduced* multiple decoy set, our potential ranks 3 out of 7 proteins as the lowest energy; it also ranks the 2cro protein as the second lowest, and the 1ctf protein as the third lowest. In contrast, Miyazawa Jernigan's potential could only rank 2 out of 7 proteins as the lowest energy in this decoy set; and for those that do not rank as the lowest, our potential also out performs the Miyazawa-Jernigan potential, as shown in Table 1. Krishnamoorthy and Tropsha's four-body potential could also rank 3 out of 7 proteins in this decoy set as the lowest energy; but for the others (that were not rank as the No.1 lowest energy structures), KT's potential ranks the 1r69 as the third lowest and 4rxn as the 5th lowest, but could only rank the 1sn3 protein as 113th.

Protein	Our potential rank	MJ potential rank	KT potential rank
1ctf	3	17	7
1r69	1	9	3
1sn3	35	97	113
2cro	2	1	1
3icb	1	1	1
4pti	1	2	1
4rxn	8	7	5

Table1. Native rank scores for the *4state_reduced* multiple decoy set

Protein	Our potential rank	MJ potential rank	KT potential rank
1beo	1	1	1
1ctf	1	1	1
1dkt-A	36	92	89
1fca	1	2	1
1nkl	1	1	1
1pgb	15	25	14
1trl-A	146	175	1179
4icb	1	1	1

Table2. Native rank scores for the *lattice_ssfit* multiple decoy set

For the *lattice_ssfit* multiple decoy set, our potential could rank 5 out of 8 proteins as the lowest energy scored ones. To compare with our potential, Miyazawa-Jernigan potential rank scores were also listed. It shows that the MJ potential could rank 4 out of 8 proteins as the lowest energy, and ranks the native 1fca protein as the second lowest. For the other three proteins 1dkt-A, 1pgb and 1trl-A, our potential performs better than the MJ potential in the ranking score for all of them. Krishnamoorthy and Tropsha's four-body potential also ranks 5 out of 8 proteins in this decoy set as the lowest energy, but for the other three proteins (1dkt-A, 1pgb and 1trl-A), our potential gives better rank then the KT's potential (as can be seen from Table2).

Protein	Our potential rank	MJ potential rank	KT potential rank
1shf-A	13	15	28
1b0n-B	25	32	488
1bba	36	92	205
1ctf	1	2	1
1dkt	1	1	4
1fc2	15	25	372
1igd	146	175	189
2cro	1	1	1
2ovo	30	55	46
4pti	7	9	7

Table3. Native rank score for the *lmds* multiple decoy set

The *lmds* multiple decoy set is probably the hardest among the three. For this decoy set, our potential could rank 3 out of 10 native structures as the lowest energy among all the decoys, while both the Miyazawa-Jernigan and Krishnamoorthy and Tropsha's potential only successfully placed 2 out of 10 native structures as the lowest energy. For all the other proteins that do not rank as the lowest energy, our potential still perform better than the other potentials, as listed in Table3.

3. RMSD test

In order to study the variation of the total energy scores with the RMSD (Root-Mean-Squared-Distance) of the protein structures, we plotted the energy scores calculated from our extracted potential for the native structure (which is assigned an RMSD value of zero Å) and its decoys against their RMSD values for each of the decoys considered. There is an observable trend in increasing the total scores with increasing RMSD values in most of the cases, as can be seen from the plots below.

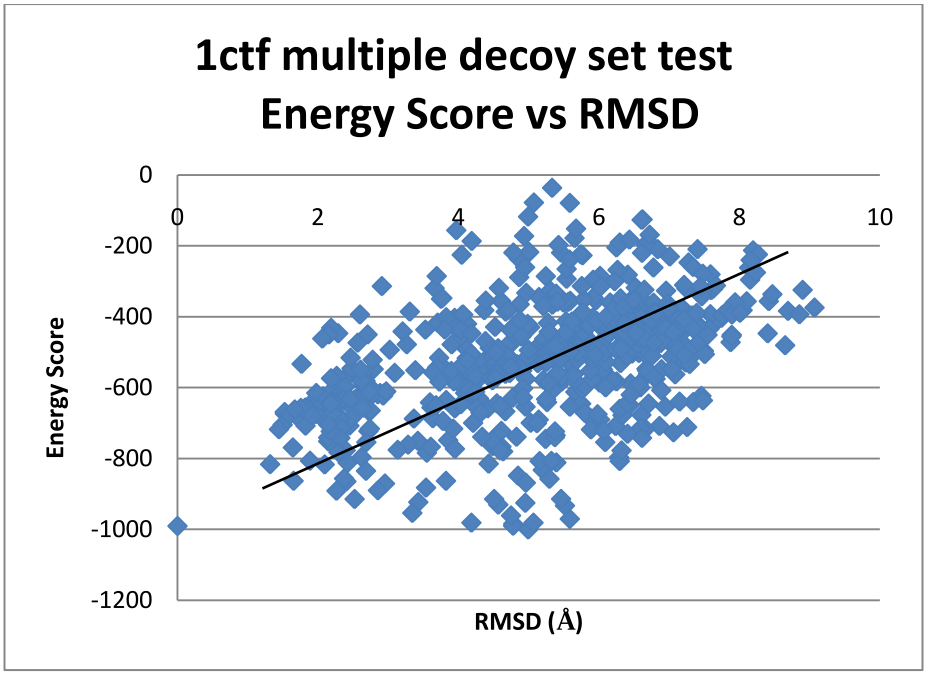

Fig.1 Total energy score for 1ctf and its 630 decoys (from the *4state_reduced* decoy set) as a function of the RMSD values. (The native score is lower than 99.7% of the decoy scores)

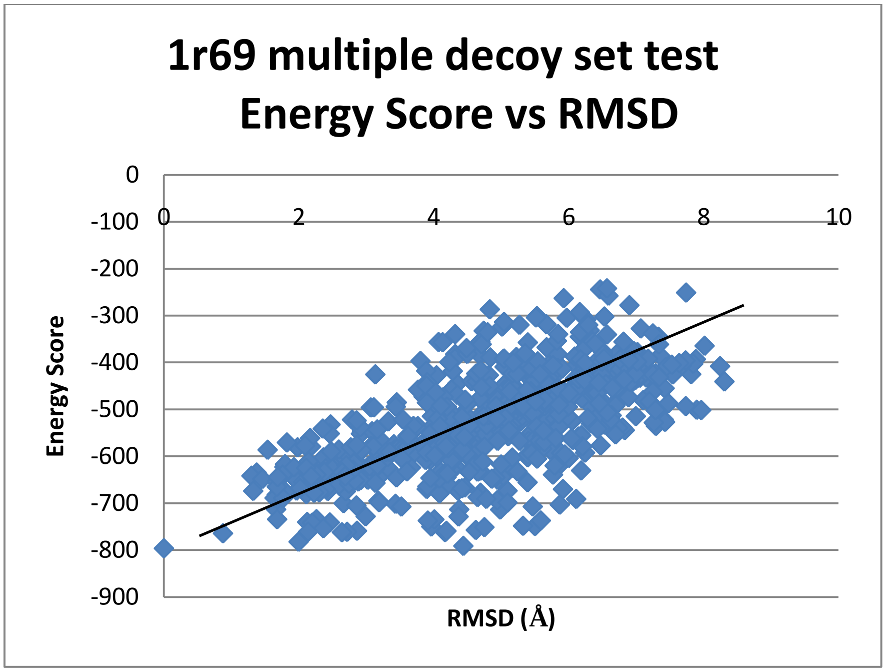

Fig.2 Total energy score for 1r69 and its 675 decoys (from the *4state_reduced* decoy set) as a function of the RMSD values. (The native score is lower than all of the decoy scores)

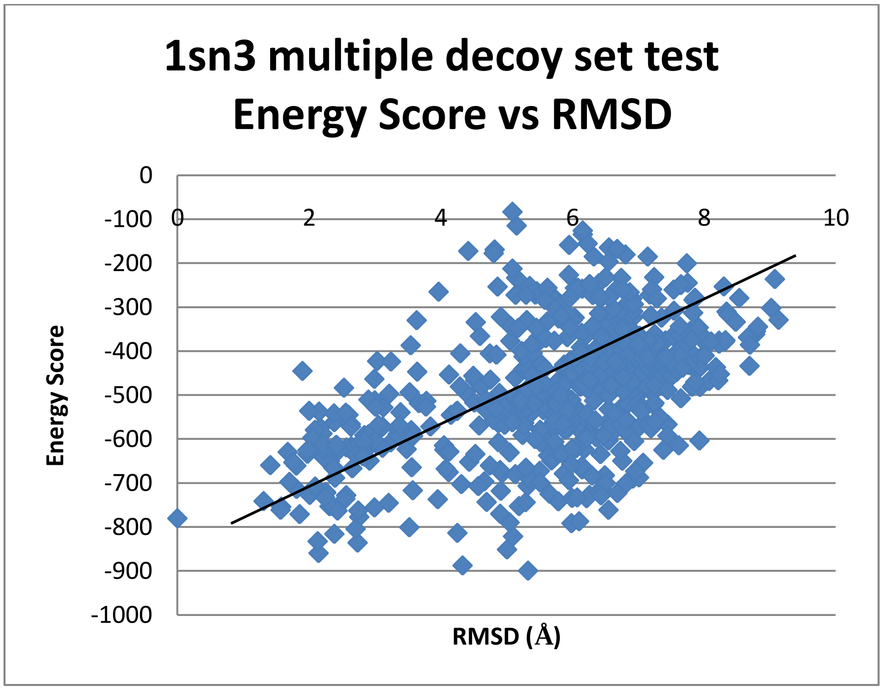

Fig.3 Total energy score for 1sn3 and its 660 decoys (from the *4state_reduced* decoy set) as a function of the RMSD values. (The native score is lower than 94.7% of the decoy scores)

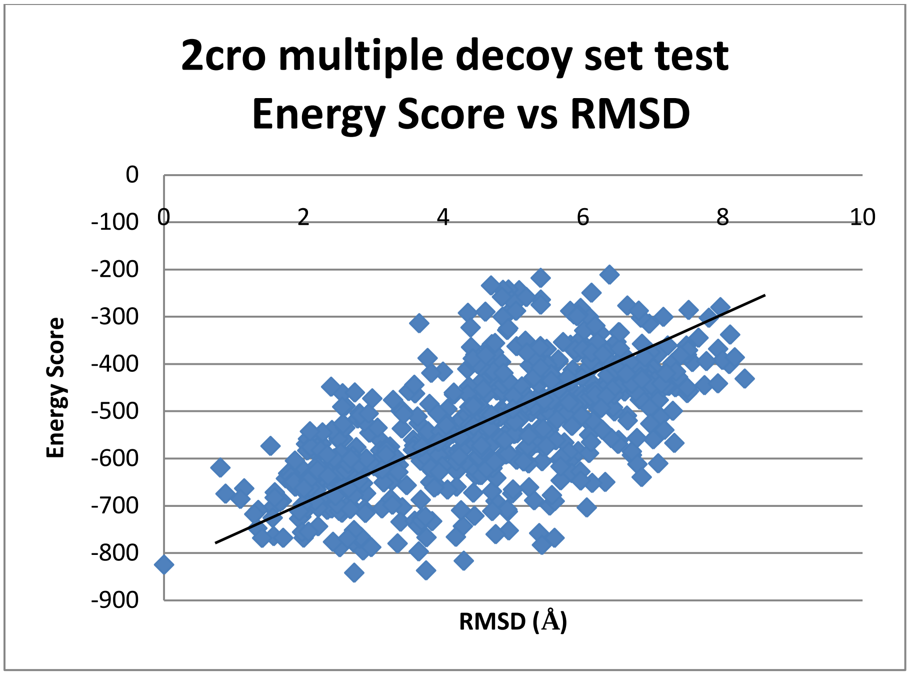

Fig.4 Total energy score for 2cro and its 674 decoys (from the *4state_reduced* decoy set) as a function of the RMSD values. (The native score is lower than 99.8% of the decoy scores)

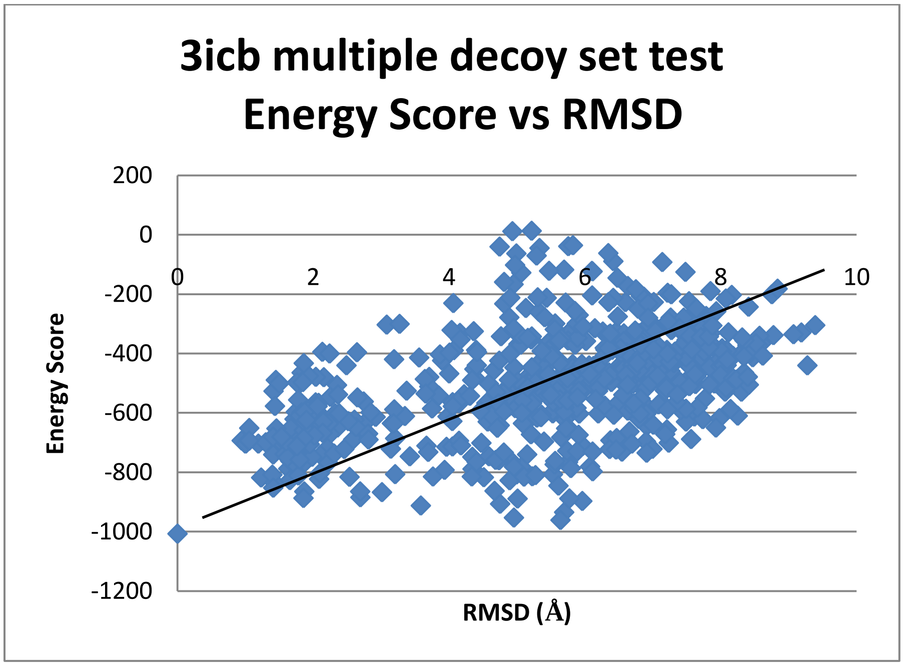

Fig.5 Total energy score for 3icb and its 653 decoys (from the *4state_reduced* decoy set) as a function of the RMSD values. (The native score is lower than all of the decoy scores)

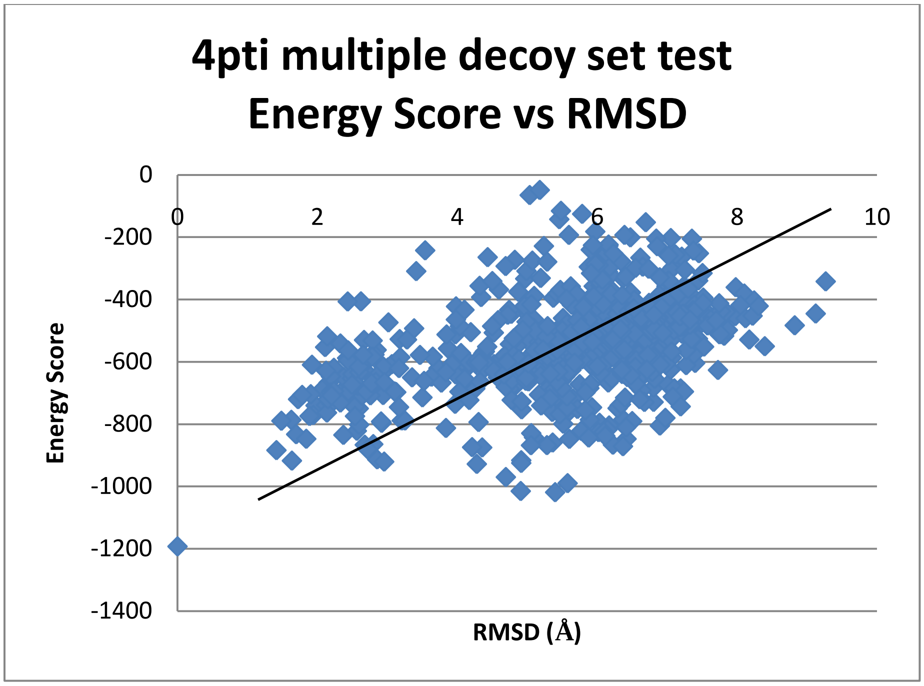

Fig.6 Total energy score for 4pti and its 687 decoys (from the *4state_reduced* decoy set) as a function of the RMSD values. (The native score is lower than all of the decoy scores)

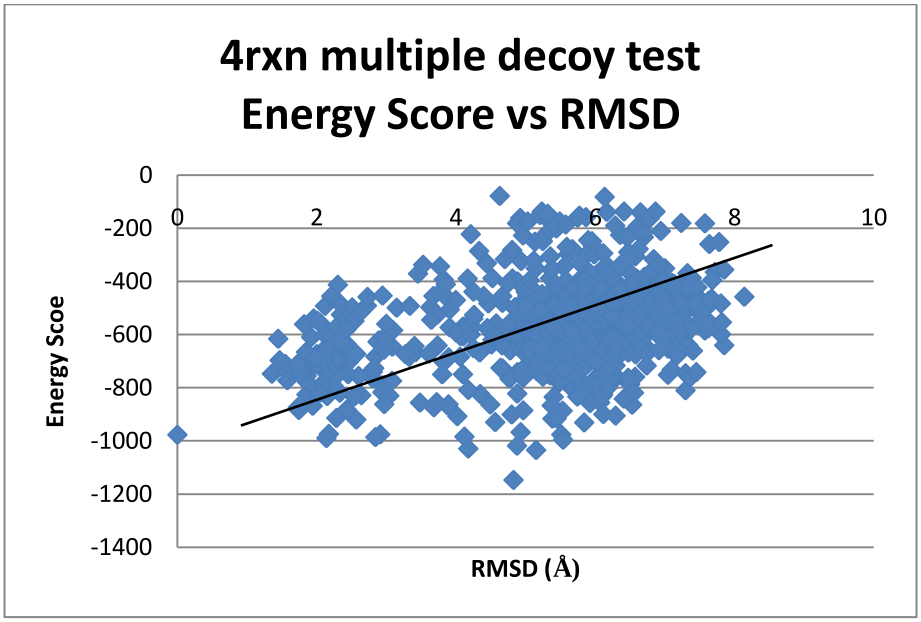

Fig.7 Total energy score for 4rxn and its 677 decoys (from the *4state_reduced* decoy set) as a function of the RMSD values. (The native score is lower than 99.0% of the decoy scores)

Chapter IV. Discussion

i) Discussion for the potential

As mentioned before, the effective pair potentials were extracted from a training set of 996 X-crystal native protein structures collected from the Protein Data Bank, as described in Materials and Methods. The potentials for each pairs of amino acids were generated. The 20 amino acids types result in a total number of 210 inter-residue potentials, because the pair interaction potential between residue *i* and residue *j* cannot be distinguished from the pair potential between residue *j* and residue *i*. The large number of residual pair occurrences for most residual pairs in the large training set guarantees sufficient statistics to derive the pair potentials for these pairs.

Firstly, it is worth noticing that our extracted effective pair potentials share similar shapes with the corresponding potentials of mean force (which were obtained by simply taking the logarithm of the radial distribution function *g(r)*). The peaks and wells appear at similar distance locations for most of the time, but overall our extracted potentials were less attractive, compare with the corresponding potentials of mean force. Also, the effective potential *u(r)* has a rather unusual form. It normally contains of three sharply defined potential wells and two barriers of distinct shapes, but some of them only contain one or two wells; all of them have magnitude of the order of *KbT*. This specific combination of barriers and wells, however, predicts the existence of polypeptide bond and the α-helix structures, which are two most prominent features shown in a protein structure. If we try to reconstruct a protein structure, these characteristic peaks and wells will give the repulsion and attractions needed to regenerate a native structure. Normally, the first minima in our effective pair potential (which appears at around r = 3Å) corresponds to the distance of two consecutive residues that found in the known structures; and the second minima (which appears at around r = 5Å, in most cases) usually represent the existence of some specific structural motifs – such as alpha-helixes, and the positions of the minima usually fix the distance d between the two residues at the *i* and *i+2* locations of the sequence.

In some other cases, the distance between the residues at the *i* and *i+2* locations of certain sequences is equal to the minimum of the third well, which appears at around r = 6 to 7Å. The angle between these two pairs of residues (*i*, *i+1*) and (*i+1*, *i+2*) turn out to be 119 degrees. Not surprisingly, this phenomenon is in concordance with the appearances of the β-sheets and β-strands.

ii) Discussion for the Decoy Set Test

As mentioned before, we used the 'R' Us single and multiple decoy set for our test of the generated residue-based pair potentials for the whole protein energy calculation. Several other studies on protein potential functions have used the same decoy set before. Therefore, we will compare the decoy set test results on our potential with these previously developed potentials. It is worth noticing that those potentials, which have been developed in different other groups, were not all residue-based. Some of them had a much higher level of model complexity, for example, the all-atomic ones. Since those potentials require much more computational time to generate and to be tested against, so we confine our comparison only to those potentials that have a similar our slightly higher complexity levels with ours.

As have been shown in the Results section, our extracted potential was able to distinguish successfully the native structures against the misfolded one in the single decoy structure test. For all the 23 protein structures that have been examined, our potential gives the native structures lower total energy scores than the misfolded structures; and the total energy of the native ones were 21.7% lower than the misfolded ones, on average. Krishnamoorthy and Tropsha's four body potential [68] also report lower energy scores for all these 23 native structures. However, their model of the potential has a much higher level of complexity. We also conducted the same test for the potential of mean force, for the same residue-level pair potential, and it turns out the the potential of mean force can only rank 19 out of 23 of the same native structure as being lower in terms of the total energy scores comparing with the misfolded structure.

For the multiple decoy set test, we also used the structures from the 'R' Us decoy database. We calculated the rank scores and the RSMD for the *lattice_ssft*, *4state_reduced* and *lmds* decoy sets that were generated using different methods. These multiple decoy sets have ranging from 600 to 2000 incorrect structures (to be compared with the native/correct ones) for each protein that listed. As from the results reported in the previous section, we could see that for *lattice_ssft* and *4state_reduced* our potential could always rank the native structure as the No.1 lowest score, for more than half of the proteins listed. This turned out to be better than both the KT's four-state potential and MJ's potential of mean force (for the same residue levels). For those proteins listed that our potential failed to rank the native conformation as the No.1 lowest energy, our rankings were still consistently lower than the ranking that given out by MJ's potential of mean force. This results demonstrated that the improvement made by including the MHNC bridge function by using a predictor-corrector algorithm, or the Reverse Monte-Carlo method, could indeed improve the quality of total energy prediction and the power of correct protein conformation selections.

For the *lmds* multiple decoy set, which was demonstrated to be a much uneasy one among the other multiple decoy sets that have been tested, our extracted pair potentials could still rank 3 out of 10 listed proteins as the lowest energy scored ones. This result also better performed the KT's four body potential and the MJ's potential of mean force. For those proteins which our potential did not give the best rank for the native structure, the ranking scores from ours potential also improved the results from the KT's four body potentials.

For the study of the variation of the total energy score with the RMSD of protein structures in multiple decoy sets *4state_reduced*, we plotted the total energy scores calculated from our potential for the native (correct) structure (which has an RMSD value of 0Å) and the corresponding total energy scores for the decoy (incorrect) structures. Figures for these plots were given in the previous section. We consistently observed strong positive correlation for the RMSD values and the total energy scores: higher RMSD valued decoy structures were observed to have higher total energies, as given by our potential. These results were consistent with several other studies that conducted using the same multiple decoy set [68].

Chapter V. Summary

In this thesis, we have provided a novel method to extract pairwise interaction potentials between 20 types of amino acids. These extracted potentials can be further used to calculate the total energy scores of a given protein, and thus can be applied to structure predictions, and correct protein conformation selections.

Our potential, which was based on the previous knowledge of 996 known protein crystal structures from the Protein Data Bank (PDB), belongs to the general category of knowledge-based protein potentials. This type of potential, in contrast to those so-called "physical-based" ones, normally use reduced structures for the protein geometry representation, and do not need quantum or other *ab initial* calculations that based on physical laws. These properties normally lead to much simpler models, and thus give shorter computing time for the model generation and validation. On the other hand, the knowledge-based potentials, since they need to use known protein crystal structures as input for parameters optimization and correction, will usually become dependent on the "knowledge" structures that being used at least to a certain extent. While this is not unusual for most knowledge-based model discovery types of problems, people working on this area tried to select complete, representative, and unbiased training set of protein as the "knowledge" structures to extract the features of the structure, so that these features, (as in this case, the radial distribution functions (RDFs) of our training sets), become robust and insensitive to any new training protein structures that added into or deleted from the set of known proteins.

Using these "knowledge" protein structures from the PDB, we were able to reduce them into residue-based points, where polypeptide bonds and other atomic/molecular details were all ignored. The obtained reduced structures were then used to extract the radial distribution functions (RDFs) between each different pairs of the amino acids. Since the protein training set we used are all in high molecular weight (>1000 amino acid residues), we could assume that the residues are in thermodynamic equilibrium; therefore, these RDFs should be able to reflect the structural properties between the residues within a protein.

From the RDFs that generated from the training set, we were able to first generate the potential-of-mean-force for different pairs of residues (which was discussed in previous work from Miyazawa and Jernigans); and our work improved this idea by including the higher order terms of the Ornstein-Zernike equation and an iterative way to estimate the bridge function that were ignored by the potential of mean force. Technically, we were using an iteration that starting from the HNC approximation for the pair interaction potential, and in each of the follow step, we conducted Monte-Carlo simulations to generate the RDFs for the updated potential. Here, the updated potentials were calculated using two different ways: one way was using a "predictor-corrector" algorithm in which the difference between the updated potential and the older potential from the previous step was generated by the difference of the RDFs in the two iteration steps; and then the difference, after some transformations and normalizations, was added to the previous step potentials as a "corrector". The iteration ends when in the two consecutive steps, the RDFs, or the corresponding pair potential for the certain pairs of residues get converged (technically we calculated the Euclidean distance between the potentials in the two consecutive steps, and when this distance becomes less than our set up threshold, we treat the two potentials as the same, so that the iteration was assumed to be converged.) The other method, which basically share the same idea but computationally more effective, was called the Reverse Monte-Carlo method. In this method we reconstructed the Hamiltonian in each iteration step after the Monte-Carlo simulation using the updated potential, and then structure factors were calculated to obtain coefficients, that were then used to solve a set of linear equations to get the corrections that needed to update the potential. Using both of these two ways, optimization of pairwise potentials could be obtained, in order for the RDFs of the final updated potential to become concordant with the ones that provided by the protein structures from our knowledge base.

After these effective pairwise potentials were extracted for 210 different pairs of amino acids, we were able to sum up the individual ones to obtain the total energy score for known protein structure. We used the ‘R’ Us single and multiple decoy sets to validate our potentials: results from these decoy set test shown that own extracted potential could successfully distinguish the native structure with lower total potential energy scores, compared with the misfolded one, for the single decoy test. For the multiple decoy test, our knowledge-based potential also out perform Miyazawa-Jernigan’s potential of mean force and Krishnamoorthy and Tropsha’s four body potential in terms of overall ranking scores.

From the above results, we could conclude that our work provided a new set of residue-level effective potentials for protein potential energy calculation, and it could be successfully used for native protein structure selections and predictions. At the meanwhile, it provides a way that improves the Miyazawa-Jernigan’s potential of mean force and Pliego-Pastrana’s potential, which used the HNC and PY approximations to include to some extent the higher-order term information from the Ornstein-Zernike equation. Since an iterative way was used in this work, we could eventually obtain pair potentials between different types of amino acids that get concordant with the corresponding radial distribution functions that extracted from know protein structure training sets. Monte-Carlo simulation shown that our potentials could get back to the original RDFs that used as the starting point of the potential-of-mean-force calculations. Future works in this topic that we are planning to conduct includes using these extracted pair potentials to predict protein structures using Monte-Carlo simulations, and further optimization/validation of the potentials with different training sets. The eventual goal of this work is to be able to generate reasonable protein structure using this set of pairwise residue potential without input for the chain connectivity knowledge, which we are currently still working on.

References

1. H. Lu and J. Skolnick, Proteins: Structure, Function, and Genetics **44**,223-232 (2001)
2. B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, M. Karplus, J. Comp. Chem. **4**, 187-193 (1983)
3. J. Novotny, R. Bruccoleri and M. Karplus, J. Mol. Biol. **177**, 787–818 (1984)
4. S. Vajda and M. Sippl, Novotny J. Curr. Opin. Struct. Biol. **7**,222-238 (1997)
5. J. Moult, Curr. Opin. Struct. Biol. **7**,194-199 (1997)
6. L. Mirny and E. Shakhnovich, J. Mol. Biol. **264**, 1164-1179 (1996)
7. M. Hao and H. Scheraga, Curr. Opin. Struct. Biol. **9**, 184-188 (1999)
8. S. Miyazawa and R. Jernigan, Proteins **36**, 357-369 (1999)
9. T. Lazaridis and M. Karplus, Curr. Opin. Struct. Biol. **10**, 139-145 (2000)
10. W. Cornell, P. Ciepak, C. Bayly, I. Gould, K. Merz, D. Frguson, D. Spelleyer, T. Fox, J. Caldwell, P. Kollman, Biochemistry **117**, 5179-5197 (1995)
11. A.D. Mackerell Jr., J. Comp. Chem. **25**, 1584-1604 (2004)
12. W.L. Jorgensen and J. Tirado-Rives, Proc. Natl. Acad. Sci. USA **102**, 6665-6670 (2005)
13. F.E. Boas and P.B. Harbury, Curr. Opin. Struct. Biol. **17**, 199-204 (2007)
14. T. Lazaridis and M. Karplus, Curr. Opin. Struct. Biol. **10**, 139-145 (2000)
15. D. Mohanty, B.N. Dominy, A. Kolinski, C.L. Brooks III and J. Skolnick, Proteins 35, 447-452 (1999)
16. S. Tanaka and H.A. Scheraga, Macromolecules **9**, 945–950 (1976)
17. T. Kortemme and D. Baker, Proc. Natl. Acad. Sci. USA **99**, 14116-14121 (2002)
18. T. Kortemme, L.A. Joachimiak, A.N. Bullock, A.D. Schuler, B.L. Stoddard and D. Baker, Nat. Struct. Mol. Biol. **11**, 371-379 (2004)
19. B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard and D. Baker, Science **302**, 1364-1368 (2003)
20. M.A. Dwyer, L.L. Looger and H.W. Hellinga, Science **304**, 1967-1971 (2004)
21. M.S. Wisz and H.W. Hellinga, Proteins **51**, 360-377 (2003)
22. T. Lazaridis and M. Karplus, Proteins **35**, 133-152 (1999)
23. T. Kortemme, A.V. Morozoc and D. Baker, J. Mol. Biol. **326**, 1239-1259 (2003)
24. G.M. Crippen, Biochemistry **30**, 4232-4237 (1991)
25. V.N. Maiorov and G.M. Crippen, J. Mol. Biol. **227**, 876-888 (1992)
26. S.Y. Huang and X. Zou, Proteins **72**, 557-579 (2008)
27. C. Zhang, S. Liu, H. Zhou and Y. Zhou, Protein Science **13**, 200-411 (2004)
28. S. Debolt, J Skolnick, Protein Eng. **9**, 637-655 (1996)
29. M. Sippl, MOrtner, M Jaritz, P. Lackner and H. Flockner, Folding Design **1**, 288-298 (1996)
30. R. Samudrala and J. Moult, J. Mol. Biol. **275**, 895-916 (1998)
31. F. Melo and E. Feytmans, J. Mol. Biol. **267**, 207-222 (1997)

32. R. Luthy, J.U. Bowie and D. Eisenberg, Nature **356**, 83-85 (1992)
33. K. Nishikawa and Y. Matsuo, Protein Eng. **6**, 811-820 (1993)
34. M.J. Sippl, J. Mol. Biol. **213**, 859-883 (1990)
35. M. Hendlich, P. Lackner, S. Weitckus, H. Floechner, R. Froschauer, K. Gottsbachner, G. Casari and M.J. Sippl, J. Mol. Biol. **216**, 167-180 (1990)
36. M.J. Sippl and S. Weitckus, Proteins: Struct. Funct. Genet. **13**, 258-271 (1992)
37. D.T. Jones, W.R. Taylor and J.M. Thornton, Nature **358**, 86-89 (1992)
38. S.H. Bryant and C.E. Lawrence, Proteins: Struct. Funct. Genet. **16**, 92-112 (1993)
39. D.G. Covell and R.L. Jernigan, Biochemistry **29**, 3287-3294 (1990)
40. S. Miyazawa and R.L. Jernigan, Macromolecules **18**, 534-552 (1985)
41. S. Miyazawa and R.L. Jernigan, J. Mol. Biol. **256**, 623-644 (1996)
42. T.L. Hill, *Statistical Mechanics*. Addison-Wesley, Reading, M.A. (1960)
43. P. Pliego-Pastrna and M.D. Carbajal-Tinoco, Phys. Rev. E **68**, 011903 (2003)
44. M.D. Carbajal-Tinoco, F. Castro-Roman and J.L. Arauz-Lara, Phys. Rev. E **53**, 3745 (1996)
45. S.H. Behrens and D.G. Grier, Phys. Rev. E **64**, 050401 (2001)
46. C.-H. Sow, K. Harada, A. Tonomura, G. Crantree and D.G. Grier, Phys. Rev. Lett. **80**, 2693 (1998)
47. P. Pliego-Pastrna and M.D. Carbajal-Tinoco, J. Phys. Chem. B **110**, 24728-24733 (2006)
48. F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, J. Mol. Biol. **112**, 535-542 (1977)
49. U. Hobohm, M. Scharf, R. Schneider and C. Sander, Protein Sci. **1**, 409-417 (1992)
50. C.A. Orengo, T.P. Flores, W.R. Taylor and J.M. Thornton, Protein Eng. **6**, 485-500 (1993)
51. D. Fischer, C.J. Tsai, R. Nussinov and H.J. Wolfson, Protein Eng. **10**, 981-997 (1995)
52. J.P. Hansen and I.R. McDonald, *Theory of Simple Liquids*, 2nd ed. (Academic, New York, 1986)
53. M.D. Johnson, P. Hutchinson and N.H. March, Proc. Roy. Soc. London, Ser. A **282**, 283 (1964)
54. W.S. Howells and J.E. Enderby, J. Phys. C **5**, 1277 (1972)
55. N.K. Ailawadi, P.K. Banerjee and A. Choudy, J. Chem. Phys. **60**, 2571 (1974)
56. L. Reatto, Phys. Rev. B **26**, 130 (1982)
57. G.L. Masserini and L. Reatto, Phys. Rev. B **30**, 5367 (1984)
58. M.W.C. Dharma-Wardana and G.C. Aers, Phys. Rev. B **28**, 1701 (1983)
59. L. Reatto, D. Levesque and J.J. Weis, Phys. Rev. A **33**, 3451 (1986)
60. R.H. Swendsen, Phys. Rev. Lett. **42**, 859 (1979)
61. G.S. Pawley, R.H. Swendsen, D.J. Wallace and K.G. Wilson, Phys. Rev. B **29**, 4030 (1984)
62. A.P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995)
63. G. Dalquist and A. Bjorck, *Numerical Methods* (Prentice-Hall, Englewood Cliffs, NJ, 1974)

64. C. Hardin, T.V. Pogorelov and Z. Luthey-Schulten, Curr. Opin. Struct. Biol. **12**, 176-181 (2002)
65. A.K. Felts, E. Gallicchio, A. Wallqvist and R.M. Levy, Proteins **48**, 404-422 (2002)
66. B.H. Park, E.S. Huang and M. Levitt, J. Mol. Biol. **2664**, 831-846 (1997)
67. J. Tsai, R. Bonneau, A.V. Morozov, B. Kuhlman, C.A. Rohl and D. Baker, Proteins **52**, 76-87 (2003)
68. R. Samudrala and M. Levitt, Protein Sci. **9**, 1399-1401 (2000)
69. B. Park and M. Levitt, J. Mol. Biol. **258**, 367-392 (1996)
70. B. Krishnamoorthy and A. Tropsha, Bioinformatics **19**, 1540-1548 (2003)