國立交通大學

生物資訊所

碩士論文

蛋白質結構摺疊新能量函式的最佳化

Optimizing New Energy Functions for Protein Folding

研究生：邱一原

指導教授：楊進木　教授

黃鎮剛　教授

中華民國九十四年六月

蛋白質結構摺疊新能量函式的最佳化

Optimizing New Energy Functions for Protein Folding

研 究 生：邱一原　　Student：Yi-Yuan Chiu

指導教授：楊進木　　Advisor：Jinn-Moon Yang

黃鎮鋼　　Jenn-Kang Huang

國 立 交 通 大 學

生 物 資 訊 所

碩 士 論 文

A Thesis Submitted to Institute of Bioinformatics
National Chiao Tung University in partial Fulfillment of the Requirements
for the Degree of Master in
Bioinformatics

June 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年六月

國　立　交　通　大　學

博碩士論文全文電子檔著作權授權書

（提供授權人裝訂於紙本論文書名頁之次頁用）

本授權書所授權之學位論文，為本人於國立交通大學 生物資訊 所 ＿＿＿＿ 組， 93 學年度第 2 學期取得碩士學位之論文。

論文題目：Optimizing New Energy Functions for Protein Folding

指導教授：楊進木教授、黃鎮剛教授

■ 同意　□不同意

本人茲將本著作，以非專屬、無償授權國立交通大學與台灣聯合大學系統圖書館：基於推動讀者間「資源共享、互惠合作」之理念，與回饋社會與學術研究之目的，國立交通大學及台灣聯合大學系統圖書館得不限地域、時間與次數，以紙本、光碟或數位化等各種方法收錄、重製與利用；於著作權法合理使用範圍內，讀者得進行線上檢索、閱覽、下載或列印。

論文全文上載網路公開之範圍及時間：	
本校及台灣聯合大學系統區域網路	■ 中華民國94年7月20日公開
校外網際網路	■ 中華民國94年7月20日公開

授 權 人：邱一原

親筆簽名：邱一原

中華民國　94 年 7 月 17 日

國　立　交　通　大　學

博碩士紙本論文著作權授權書

（提供授權人裝訂於全文電子檔授權書之次頁用）

本授權書所授權之學位論文，為本人於國立交通大學 生物資訊 所 ＿＿＿＿＿＿ 組，93 學年度第 2 學期取得碩士學位之論文。

論文題目：Optimizing New Energy Functions for Protein Folding

指導教授：楊進木教授、黃鎮剛教授

■ 同意

本人茲將本著作，以非專屬、無償授權國立交通大學，基於推動讀者間「資源共享、互惠合作」之理念，與回饋社會與學術研究之目的，國立交通大學圖書館得以紙本收錄、重製與利用；於著作權法合理使用範圍內，讀者得進行閱覽或列印。

本論文為本人向經濟部智慧局申請專利(未申請者本條款請不予理會)的附件之一，申請文號為：＿＿＿＿＿＿＿＿＿＿，請將論文延至＿＿年＿＿月＿＿日再公開。

授 權 人：邱一原

親筆簽名：邱一原

中華民國 94 年 7 月 17 日

國家圖書館博碩士論文電子檔案上網授權書

ID:GT009251504

本授權書所授權之論文為授權人在國立交通大學 生物科技 學院 生物資訊 所 _____ 組 _93_學年度第_2_學期取得碩士學位之論文。

論文題目：Optimizing New Energy Functions for Protein Folding

指導教授：楊進木教授、黃鎮剛教授

茲同意將授權人擁有著作權之上列論文全文（含摘要），非專屬、無償授權國家圖書館，不限地域、時間與次數，以微縮、光碟或其他各種數位化方式將上列論文重製，並得將數位化之上列論文及論文電子檔以上載網路方式，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

※ 讀者基於非營利性質之線上檢索、閱覽、下載或列印上列論文，應依著作權法相關規定辦理。

授權人：邱一原

親筆簽名： 邱一原

民國94年7月17日

1. 本授權書請以黑筆撰寫，並列印二份，其中一份影印裝訂於附錄三之二(博碩士紙本論文著作權授權書)之次頁；另一份於辦理離校時繳交給系所助理，由圖書館彙總寄交國家圖書館。

國立交通大學

生物資訊研究所碩士班

論文口試委員會審定書

本校 生物資訊研究所 碩士班 邱一原 君

所提論文(中文) 蛋白質結構摺疊新能量函式的最佳化

(英文) Optimizing New Energy Functions for Protein Folding

合於碩士資格水準、業經本委員會評審認可。

口試委員： 呂平江 教授 呂平江 教授

黃明經 教授 黃明經 教授

指導教授： 楊進木 教授 楊進木 教授

黃鎮剛 教授 黃鎮剛 教授

所長： 何信瑩 教授

中華民國 94 年 6 月 28 日

蛋白質結構摺疊新能量函式的最佳化

學生：邱一原　　指導教授：楊進木
黃鎮剛

國立交通大學生物資訊所碩士班

摘　要

蛋白質結構預測的方法中，一個可行的策略是產生大量可能的結構，再依據計分函式或能量函式從中挑選出最適當的結構。蛋白質摺疊的結構範圍相當廣闊，直接利用計算方法產生出類似自然結晶結構的結構有相當的難度；在理想的情況下，非自然的摺疊結構或許可以提供相當的助益，利用非自然的摺疊結構來修正計分函式或能量函式，並利用修正過後的計分函式或能量函式來分辨出自然和非自然的結構，也許是個可行的方法。在這篇論文研究中，我們發展了兩個適用於蛋白質摺疊問題上的能量函式，並且在常見的測試評量中有不錯的表現。其中的一個能量函式為 MOLSIM，是以基本物理作用為基礎，所發展出的能量函式。另外一個能量函式是 GEMSCORE，也是以物理作用為基礎，但是卻是以簡化過的物理公式來算能量。與其他利用物理作用所發展出的能量函式不同的是設定的參數數量，一般以物理作用為基礎的能量函式可能會需要幾百甚至幾千個參數設定。而我們的能量函式只針對不同的能量計算項目給予不同的比例參數，並利用演化式方法來最佳化這些參數。我們詳細地分析和比較我們的能量函式與之前其他研究者所發展的能量函式，在六個測試的資料中，包含有 96 種不同的蛋白質，超過 70,000 個結構。MOLSIM 和 GEMSCORE 分別能從中辨識中 70 和 73 個正確的自然結構。我們相信我們的能量函式夠快並且夠簡單，而且應用在結構預測上，可以分辨出自然和非自然的結構。

Optimizing New Energy Functions for Protein Folding

Student: Yi-Yuan Chiu Advisor: Dr. Jinn-Moon Yang
Dr. Jenn-Kang Huang

Institute of Bioinformatics
National Chiao Tung University

Abstract

One strategy for protein structure prediction is to generate a large number of possible structures (decoys) and select the most fitting ones based on a scoring or energy function. The conformational spaces of a protein are huge, and chances are rare that any heuristically generated structures will directly fall in the neighborhood of the native structure. It is desirable that the unfitting decoy structures can provide insights into native structures, so prediction can be made progressively. In this thesis, we develop two simple energy functions for protein folding and show that their good performance with popular benchmarks. One is MOLSIM, a physics-based energy function; another is GEMSCORE, an empirical energy function based on physical mechanisms with simplified model. Instead of hundreds or thousands parameters used in other physics-based energy functions by previous authors, we adopt only few overall weights and use an evolutionary algorithm to optimize the parameters of these two energy functions. Here we present a systematic comparison of our results with the works based on physics-based energy functions by previous authors. Six testing decoy sets, including 96 protein sequences with more 70,000 structures were evaluated. There are 70 and 73 native proteins that identified from these decoy sets with MOLSIM and GEMSCORE, respectively. We believe that our energy functions are fast and simple to discriminate between native and nonnative structures for protein structure prediction.

Acknowledgements

The most appreciation is for my advisors, Dr. Jinn-Moon Yang and Dr. Jenn-Kang Huang. Because of their advices and instructions, I could finish this thesis. I am very grateful for Dr. Jenn-Kang Huang and thanks for his suggestions on energy functions and molecular dynamic. He helped me to understand the problem of protein folding and how to solve mistakes. I also thank goes to my thesis adviser, Dr. Jinn-Moon Yang. Without his suggestions and comments on writing thesis and GEMSCORE, I can't accomplish this work. Thanks my two advisors very much.

I also must thank to everyone in Yang's lab. We work together, and have funs together. Thanks for guys who discuss with me, especially Hung-Chu Chen. Thanks for their company with me during past two years. Finally, I would thank my whole family, for their support during past twenty-four years.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Protein Structure Prediction and Energy Functions

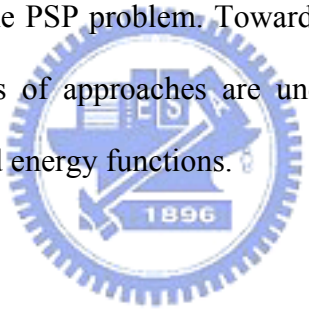
Protein structure prediction (PSP) at atomic detail, which is an important aspect of the protein folding problem, remains one of the fundamental unsolved problems in the field of computational structural biology. The PSP computational method involves two basic critical elements: an efficient search method and a reliable energy function. The former should be able to search a large number of potential structure candidates with reasonable accuracy and speed. Energy functions for PSP should effectively discriminate between the native structure and non-native structures during the search structure candidates.

Many methods have been proposed to for PSP by generating a reliable structure from a given sequence whose structure is unknown. These methods are separated into three categories: homology modeling (comparative modeling), fold recognition (threading), and *ab initio* folding. The main concept of homology modeling and fold recognition is to find out the best known structure as template for a given target sequence. The evolutionary relationship can be deduced from sequence similarity [1-3] or by threading a sequence against a library of structures and selecting the best match [4, 5]. Then generate candidate structures by aligning the target sequence and the template structure. The template selection and alignment are the critical steps for modeling methods.

Unlike homology modeling and fold recognition, *ab initio* folding generates candidate structure only from sequence information. No dependence on database information is needed and predictions are based on sampling of the protein conformational space to optimize an energy function [6-14]. In one possible scenario for selecting the most reliable structure, a large number of candidate structures are generated and evaluated with an energy

function that can distinguish the native and misfolded ones (decoys).

Regardless of which class a prediction method belongs to, an effective energy function is usually required. These functions are typically used in one of two ways: they are either used as optimization criteria to drive conformational search algorithms to sift through the conformational space (folding problem) or they are used as selection criteria to select a conformation from a set of possible structures (reverse folding problem). Energy functions of PSP computational methods are generally rooted in the thermodynamic hypothesis that the native-state conformations is the most stable conformation and, therefore, must occupy the lowest energetic state [15]. Effective energy functions that can accurately depict the energy landscape of protein conformation space are a common requirement for all computational approaches to the PSP problem. Toward the aim of developing such energy functions, three different types of approaches are under investigation: knowledge-based energy functions, physics-based energy functions.



1.2 Related Works

Knowledge-based energy functions are generally derived from distributions of experiment structural data [16, 17]. A previous comparative study showed that knowledge-based functions usually perform better [18]. This is part of the reason that most previous works have been focused on knowledge-based energy functions in structure prediction applications [16, 17, 19-21]. Given a database of high-quality structural information, knowledge-based energy functions can often produce the desired results with far less computational overhead. However, there are a few exceptions. For example, the Park et al. compared various knowledge-based and physics-based functions and found that the knowledge-based functions are not better than physics-based functions at ranking native structures [22]. In addition, the Tobi et al. showed that it is not possible to find a pairwise

knowledge-based potential with a resolution of 1 Å or better [23, 24].

Energy functions derived from experimental structure information are constrained by their underlying statistics, which means that their accuracy and applicability are intrinsically tied to the data source used for parameterization; if the particular data set overrepresents a certain class of structural properties (e.g., helical structures), the resulting energy function would also reflect this statistical bias in its scoring. For the reasons of simplification and fast, reduced representation are usually used in knowledge-based energy functions. These energy functions may contain pseudo-potentials which lack the physical meanings.

On the other hand, physics-based energy functions are based on physical mechanisms. And they do not have such inherent limitations when they are carefully parameterized. Because they are derived from *ab initio* quantum mechanical calculations based on the principles of physics alone, they do not have any intrinsic bias toward any particular structural properties. Physics-based approaches assume that the protein potential energy functions can be broken down into terms of bond stretching, angle bending, torsional, and non-bonded interactions. These parameters in these terms are then fitted to high-level *ab initio* quantum mechanical calculations and small molecule thermodynamic and spectroscopy data. The advantage of a physics-based energy function is the clear physical meaning of each individual term. Great efforts have been invested to understand the driving forces or dominant forces in its discriminative ability.

Despite their perceived advantages, physics-based energy functions have not been widely considered practical for fold recognition or protein structure prediction types of applications. This finding was mostly due to the high-computation cost required and the cumulative inaccuracies introduced in parameterization of the energy functions compounded by the fact that most of the earlier energy functions were calibrated against

rather sparse and often qualitative experimental data. Because of the continued improvement in computer speed and advances in energy function design, this situation has begun to change in recent years; physics-based energy functions are now showing signs of living up to their potential [25-33].

As more new energy functions are being developed, one problem that became apparent was the lack of a standardized benchmark to allow comparisons of performances across different energy functions parameterized by using different properties and methodologies. One of the earliest studies that resembled a benchmarking test for protein potentials was the study carried out by Novotny *et al*. [34], in which two proteins with the same number of residues but different folds were considered and the sequence of one was "threaded" onto the fold of the other. The resulting correct and incorrect models were then evaluated with use of the CHARMM potential. The conclusion of this study was frequently misinterpreted as supporting evidence that modern molecular mechanics-based potentials were not of sufficient accuracy to discriminate between native and non-native folds.

In the spirit of the Novotny test, several groups have created decoy sets (non-native or near-native conformations) as a testing benchmark for evaluating the usefulness of a new scoring function [22, 35-37]. These decoy sets provide an objective common platform on which new energy functions can be evaluated. Furthermore, one may also view the decoys as the products of a previous step in a hierarchical structure prediction scheme; a high-quality scoring function could then be integrated as a filter to select the best candidate from among a set of low-resolution prediction of the native fold.

Another problem for developing a physics-based energy function is that the determination of potential parameters is conceptually and practically difficult. Although it would seem the most deductive and logical, determining potential parameters solely from

electronic structure calculation of small molecules does not necessarily give the best performance for modeling proteins in solvent. Instead of this bottom-up approach, we might ask whether we can infer physical forces from their consequences, that is, the structures of proteins already in the structural database. Recently, structural genomics projects have started to produce thousands of 3D protein structures. If we can use this information to improve protein energy functions, this would yield energy functions that are practically powerful for many purposes and should be conceptually helpful for gaining insight into the physical principles of protein architecture.

In this thesis, we try to develop a new energy function that combine the advantages both of knowledge-based and physics-based energy functions and avoid the disadvantages of them. We used simplified energy terms physical mechanisms to form our energy functions. In order to optimize the parameters of our energy functions for protein folding, we adopt a reduced optimization scheme that to consider the overall weight for each energy term as the parameter of each energy term. For our purpose, we use an evolutionary algorithm and a well-developed decoy sets as our training set to optimize the overall weights.

1.3 Evolutionary Algorithms

An evolutionary algorithm is a generally adaptable concept for problem solving, especially difficult optimization problems. It is based on ideas genetics and natural selection. First, a problem is coded to a chromosome, which is represented as solution and initialed randomly. Then find out batter solution in searching space by mutations and crossovers. Selection is most important step in an evolutionary algorithm. It is decided by the objective function of an evolutionary algorithm to pick up the batter solution. It is finished when find a reliable solution by repeating mutation, crossover, and selection again and again.

Evolutionary algorithms have been used to solve problems involving large search spaces where traditional optimization methods are less efficient. They were applied to many biological problems in varying ways. Currently, there are about three main independently developed but strongly related implementations of evolutionary algorithms: genetic algorithms [38], evolution strategies [39], and evolutionary programming [40].

In general the coding function of genetic algorithms was binary-represented. These genetic algorithms may introduce an additional multimodality, making the combined objective function more complex than the original function. To achieve better performance, real-coded genetic algorithms [41, 42] have been introduced. However, they generally employ random-based mutations and hence, still require lengthy local searches near local minimums. In contrast, evolution strategies and evolutionary programming [43, 44], mainly use real-valued representation and focus on self-adaptive Gaussian mutations. This type of mutation has succeeded in continuous optimization and has been widely regarded as a good operator for local searches. Unfortunately, experiments [45] show that self-adaptive Gaussian mutation leaves individuals trapped near local minimums for rugged functions.

Because none of these three types of original evolutionary algorithms is very efficient, many modifications have been proposed to improve solution quality and to speed up convergence. Such a hybrid approach may make a better tradeoff between computational cost and the global optimality of the solution. However, for existing methods local search techniques and genetic operators often work separately during the search process. Family competition evolutionary algorithm (FCEA) is proposed to improve the above approaches and has a good performance for many problems.

FCEA is a multi-operator approach that combines three mutation operators: decreasing-based Gaussian mutation, self-adaptive Gaussian mutation, and self-adaptive

Cauchy mutation. It incorporates family competition [46, 47] and adaptive rules for controlling step sizes to construct the relationship among these three operators. To balance the search power of exploration and exploitation, each of operators is designed to compensate for the disadvantages of the other.

In this thesis, we use the FCEA to look for the most suitable weights of energy terms for protein folding. The next section describes the energy functions and the details of FCEA, such as the scoring function and the representation. In the third section, we analyze the performance on training set and test our energy functions with optimal weights by popular benchmark testing sets. We also discuses the differences of energy functions in many situations. Conclusions and future works are drawn in the fourth section.

Chapter 2 Materials and Methods

For a realizable energy function for protein folding, we adopted two physics-based energy functions which developed for different problems. We modified these energy functions and optimized them using evolution algorithm. One of them is MOLSIM that is based on physical mechanism and originally developed for molecular dynamics. Another is modified from the energy function of GEMDOCK [48, 49] used in protein-ligand docking.

Figure 1 shows the flowchart of training and testing steps. The quality of native structures and decoy structures are not generally well and there are some errors in these structures, like lack of coordinates of atoms. First, we filter these error structures or incomplete structures in training set and testing sets. Second, calculate the energy values of each native and decoy structures in training set using our energy functions. Then, optimize the overall weights of energy terms in energy functions by evolutionary algorithm, FCEA, and get these optimal weights. Repeat above steps until convergence or maximum generations of FCEA. Final, apply these weights into our energy functions and calculate the energy values of the structures in four testing sets and analyze the results and performances. Through these steps, we train the optimal weights of energy terms and judge the quality of our energy functions with benchmarks.

2.1 Energy Functions

2.1.1 MOLSIM : physics-based energy function

MOLSIM is a physics-based energy function developed for molecular dynamics and simulation by Professor Jenn-Kang Huang. This energy function is based on physical mechanisms contain bond stretching, angle bending, torsional, and non-bonded interactions as other force fields for molecular dynamics.

The complete energy function is given as

$$E_{tot} = E_{bond} + E_{angle} + E_{torsion} + E_{elect} + E_{vdW}, \quad (1)$$

where E_{bond} is covalent bond potential, E_{angle} and E_{tosion} are bond angles and torsion angles potentials, respectively. E_{elect} and E_{vdW} are electrostatics potential and van der Waals potential, respectively. MOLSIM uses these physics-based potentials for molecular dynamics and simulation. MOLSIM is all-atom energy function and polar hydrogen atoms are needed in the energy function. Therefore, there are several steps before the energy calculation.

First, MOLSIM program reads the information of protein structure from PDB-format file including coordinates of atoms and residue information. At secondary step, store the structure information with templates of 20 amino acid libraries and check the lacks of atoms and the errors of residues. Then, add the polar hydrogen atoms in specific location defined in the templates. The additional hydrogen atoms may lead to form steric crashes. Energy minimization, steepest descent, is used to reduce steric crashes. After energy minimization, we get the final structure for energy calculation.

Structures may crash or fold to another structure during long molecular dynamic, it could bring some errors and reduce the effect of protein energy functions. The energy values of native and decoy structures are calculated after energy minimization and no more dynamic steps is needed.

Physical mechanisms used in MOLSIM contain bond stretching, angle bending, torsional, and non-bonded interactions. In this thesis, we simplify the original energy function by removing bonded interactions which includes bond stretching, angle bending, and torsional interactions. We consider the energy terms of non-bonded interactions, E_{elect}

and E_{vdW}, and add extra solvation energy (E_{SAS}) to form new MOLSIM energy function. The physics-based energy function, MOLSIM, we used is given as

$$E_{tot} = E_{elect} + E_{vdW} + E_{SAS}. \tag{2}$$

The electrostatic energy (E_{elect}) between atoms i and j is

$$E_{elect} = \sum_{i=1}^{N}\sum_{j=1}^{N} 332 \frac{q_i q_j}{r_{ij}}, \tag{3}$$

where r_{ij} is the distance between the atoms i and j, q_i and q_j are the atomic partial charges of atoms i and j. and 332 is a factor that converts the electrostatic energy into kilocalories per mole. The van der Waals potential used here is Lennard-Jones 6,12 potential:

$$E_{vdW} = \sum_{i=1}^{N}\sum_{j=1}^{N} \left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}\right), \tag{4}$$

where r_{ij} is the distance between the atoms i and j, A and B are constants and predefined by different atom types. For reducing the computational resource, a cutoff of non-bonded interactions, 6.0 Å, is adopted during energy calculation.

The solvent effect is an important potential in protein folding. Many methods are proposed to this issue, for example atomic solvation parameters (ASPs) [50, 51] and generalized Born model (GB) [52] . In this thesis, we use ASPs to approximate to the solvation energy. The solvation energy is calculated by ASPs and solvent-accessible surface area.

The solvation energy, E_{SAS}, is given as

$$E_{sas} = \sum_{i=1}^{N} \sigma_i A_i , \tag{5}$$

where σ_i is the atomic salvation parameter (ASPs) for atom type *i,* and A_i is the Lee and Richards solvent-accessible surface area of atom type *i*. In our calculations, we use a probe radius of 1.4 Å to simulate a water molecule to calculate atomic solvent-accessible surface area. We divide all atoms into six atom types: C, non-charged O, non-charged N, S, charged O, and charged N, used with Eisenberg *et al*. [50, 51]. Table 1 shows the detail the description of each energy term.

2.1.2 GEMSCORE : empirical energy function

In this thesis, we developed another empirical energy function modified from GEMDOCK [48, 49]. The energy function of GEMDOCK has a good performance in flexible protein-ligend docking [48, 49] and screening [53, 54]. GEMDOCK was developed well for docking problem, but there are different between docking problem and protein folding. We need to revise the original energy function of GEMDOCK and redevelop a new empirical energy function by adding other energy terms for protein folding problem.

The energy function of GEMDOCK is given as

$$E_{tot} = E_{inter} + E_{intra} + E_{penal}, \quad (6)$$

where E_{inter} and E_{intra} are the intermolecular and intramolecular energy, respectively, E_{penal} is a large penalty value. We only retain E_{inter} and use ASPs and solvent-accessible surface area as energy term of solvent effect (E_{SAS}). E_{inter} is including electrostatic energy, simplified van der Waals and hydrogen-bonding potentials. As for the requirement of later optimization, we divide the intermolecular energy (E_{inter}) to three terms: electrostatic energy (E_{elect}), van der Waals potential (E_{vdW}) and hydrogen-bonding potential on backbone (E_{bHB}).

The empirical energy function of GEMSCORE is given as

$$E_{tot} = E_{elect} + E_{vdW} + E_{bHB} + E_{SAS}. \quad (7)$$

Electrostatic energy, E_{elect}, is based on general Columbic electrostatic function:

$$E_{elect} = \sum_{i=1}^{N}\sum_{j=1}^{N} 332 \frac{q_i q_j}{4r_{ij}^2}, \quad (8)$$

where r_{ij} is the distance between the atoms i and j, q_i and q_j are the formal charges and 332 is a factor that converts the electrostatic energy into kilocalories per mole. The formal charge of atoms is indicated in Table 2. In order to decrease the incorrect situation, we set the low bound of electrostatic energy to -10 (see Figure 2).

The van der Waals potential, E_{vdW}, and hydrogen-bonding potential on backbone, E_{bHB}, are both simplified atomic pair-wise potential functions (see Figure 2). In this model, these two potentials are calculated by the same function form but with different parameters. The model is given as

$$E_{vdW}, E_{bHB} = \begin{cases} V_6 - \dfrac{V_6 r_{ij}^{B_{ij}}}{V_1} & r_{ij}^{B_{ij}} \le V_1 \\ \dfrac{V_5 (r_{ij}^{B_{ij}} - V_1)}{V_2 - V_1} & V_1 < r_{ij}^{B_{ij}} \le V_2 \\ V_5 & \text{if} \quad V_2 < r_{ij}^{B_{ij}} \le V_3 \\ V_5 - \dfrac{V_5 (r_{ij}^{B_{ij}} - V3)}{V_4 - V3} & V_3 < r_{ij}^{B_{ij}} \le V_4 \\ 0 & r_{ij}^{B_{ij}} > V_4 \end{cases}. \quad (9)$$

$r_{ij}^{B_{ij}}$ is the distance between the atoms i and j with the interaction type B_{ij} forming by the atomic pairwise where B_{ij} is a state of either van der Waals potential or hydrogen-bonding potential on backbone. The parameters of these different potentials, V_1,…, V_6, given in

Figure 2. The value of the hydrogen-bonding potential on backbone should be larger than the one of the van der Waals potential.

In order to calculate the value of the hydrogen-bonding potential, it is must to identify the hydrogen bond. We assigned hydrogen bonds according to the method used with Baker and Hubbard [55]. Figure 3 shows the definition of a hydrogen bond used in GEMSCORE. If the angle of N-H-O (θ_{NHO}) is more than 120° and the distance between hydrogen atom and oxygen atom (r_{OH}) is less than 2.5 Å, a hydrogen bond is assigned.

To decide hydrogen bonds, we have to add extra hydrogen atoms which don't be contained in PDB-format file before energy calculation. Only the hydrogen bonds formed with both hydrogen atoms and oxygen atoms located at backbone are considered, and we ignore the hydrogen bonds located between backbone and side-chain or side-chain and side-chain.

The solvation energy, E_{SAS}, which is the same as MOLSIM is given as

$$E_{sas} = \sum_{i=1}^{N} \sigma_i A_i \,, \tag{10}$$

where σ_i is the ASPs for atom type *i,* and A_i is the Lee and Richards solvent-accessible surface area of atom type *i*. We use a probe radius of 1.4 Å to simulate a water molecule to calculate atomic solvent-accessible surface area. We divide all atoms into six atom types: C, non-charged O, non-charged N, S, charged O, and charged N, used with Eisenberg *et al*. [50, 51]. Table 1 shows the detail description of each energy term.

2.2 Optimization

The core idea of our evolutionary approach was to design multiple operators that cooperate using the family competition model [46, 47], which is similar to a local search

procedure. The Gaussian and Cauchy mutations, continuous genetic operators, search the weights for the energy terms.

2.2.1 Representation and Initiation

Our optimized method works as follows: It randomly generates a starting population with N solutions of weights for energy function. Each solution is represented as a set of 3n-dimensional vectors (x^i, σ^i, ψ^i), where n is the number of energy terms of an energy function and $i = 1,\cdots,N$, where N is the population size. The vector x is the adjustable variables representing a particular weights of a energy term to be optimized. σ and ψ are the step-size vectors of decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. In other words, each solution x is associated with some parameters for step-size control. The initial step size σ is 0.8 and ψ is 0.2.

2.2.2 Family Competition Evolutionary Algorithm (FCEA)

After initializes the solutions, it enters the main evolutionary loop, which consists of 2 stages in everyone iteration: decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. Each stage is realized by generating a new quasi-population (with N solutions) as the parent of the next stage. These stages apply a general procedure "FC_adaptive" with only different working population and the mutation operator. The FC_adaptive procedure employs 2 parameters, namely, the working population (P, with N solutions) and mutation operator (M), to generate a new quasi-population.

The main work of FC_adaptive is to produce offspring and then conduct the family competition. Each individual in the population sequentially becomes the "family father." With a probability p_c, this family father and another solution that is randomly chosen from the rest of the parent population are used as parents for a recombination operation. Then the

new offspring or the family father (if the recombination is not conducted) is operated by differential evolution to generate a quasi-offspring. Finally, the working mutation is operates on the quasi-offspring to generate a new offspring. For each family father, such a procedure is repeated L times, called the family competition length.

Among these L offspring and the family father, only the one with the lowest scoring function value survives. Since we create L children from one "family father" and perform a selection, this is a family competition strategy. This method avoids the population prematureness but also keeps the spirit of local searches. Finally, the FC_adaptive procedure generates N solutions because it forces each solution of the working population to have one final offspring. In the following, genetic operators are briefly described. We use $a = (x^a, \sigma^a, \psi^a)$ to represent the "family father" and $b = (x^b, \sigma^b, \psi^b)$ as another parent. The offspring of each operation is represented as $c = (x^c, \sigma^c, \psi^c)$. The symbol $x^s{}_i$ is used to denote the ith adjustable optimization variable of a solution s, $\forall i \in \{1,\ldots, N\}$.

2.2.3 Recombination and Mutation Operators

We implemented modified discrete recombination and intermediate recombination. A recombination operator selected the "family father (a)" and another solution (b) randomly selected from the working population. The former generates a child as follows:

$$x_i^c = \begin{cases} x_i^a \\ x_i^b \end{cases} \text{ with probability } \begin{matrix} 0.8 \\ 0.2 \end{matrix}. \tag{11}$$

The generated child inherits genes from the "family father" with a higher probability 0.8. Intermediate recombination works as

$$w_i^c = w_i^a + \beta(w_i^b - w_i^a)/2, \tag{12}$$

where w is σ or ψ based on the mutation operator applied in the FC_adaptive procedure. The intermediate recombination only operated on step-size vectors and the modified discrete recombination was used for adjustable vectors (x).

After the recombination, a mutation operator, the main operator of our evolutionary approach, is applied to mutate adjustable variables (x). Gaussian and Cauchy Mutations are accomplished by first mutating the step size (w) and then mutating the adjustable variable x:

$$\begin{aligned} w'_i &= w_i A(\cdot) \\ x'_i &= x_i + w'_i D(\cdot) \end{aligned}, \tag{13}$$

where w_i and x_i are the ith component of w and x, respectively. And w_i is the respective step size of the x_i, where w is σ or ψ. $A(\cdot)$ is evaluated as $\exp[\tau' N(0, 1) + N_i(0,1)]$ if the mutation is a self-adaptive mutation, where $N(0,1)$ is the standard normal distribution, $N_i(0, 1)$ is a new value with distribution $N(0, 1)$ that must be regenerated for each index i. When the mutation is a decreasing-based mutation $A(\cdot)$ is defined as a fixed decreasing rate $\gamma = 0.95$. $D(\cdot)$ is evaluated as $N(0, 1)$ or $C(1)$ if the mutation is, respectively, Gaussian or Cauchy. For example, the self-adaptive Cauchy mutation is defined as

$$\begin{aligned} \psi_i^c &= \psi_i^a \exp[\tau' N(0,1) + \tau N_i(0,1)], \\ x_i^c &= x_i^a + \psi_i^c C_i(t). \end{aligned} \tag{14}$$

We set τ and τ′ to $(\sqrt{2n})^{-1}$ and $(\sqrt{2\sqrt{2n}})^{-1}$, respectively, according to the suggestion of evolution strategies. A random variable is said to have the Cauchy distribution [$C(t)$] if it has the density function: $f(y;t) = (t/\pi)/(t^2 + y^2), -\infty < y < \infty$. In this thesis, t is set to 1. Our decreasing-based Gaussian mutation uses the step-size vector σ with a fixed decreasing rate $\gamma = 0.95$ and works as $\sigma^c = \gamma\sigma^a$ and $x_i^c = x_i^a + \sigma^c N_i(0,1)$.

2.2.4 Objective Function of FCEA

For optimization, the energy function becomes

$$E = \sum_i w_i \mu_i \,, \tag{15}$$

where w_i is the weight of energy term μ_i. Given an energetic weight set w (in this thesis, it's 8 and 9 for MOLSIM and GEMSCORE, respectively), we used FCEA to look for the most suitable energy function by minimizing a well-developed objective function. Eight energy terms of MOLISM are define in Equations 2 and 5 and nine energy terms of GEMDOCK are define in Equations 7 and 10.

A successful energy function not only has to be able to correctly distinguish between native and native-like structures but must also do so convincingly. In the regard, the quality of an energy function is judged by the size of energy gap assigns to the native structure and the average energy of the rest of the non-native structures. A mostly used measure for assessing this quality is the *Z*-score. We defined the *Z*-score as follows:

$$Z = \frac{E_{native} - < E >}{\Delta E}, \tag{16}$$

where

$$\Delta E = \sqrt{\sum_d (E_d - < E >)^2} \,. \tag{17}$$

E_{native} is the energy value of a native structure, E_d is the energy score of a decoy structure d, and $<E>$ is the mean of energy values of all non-native structures in a decoy set. *Z*-score is used for measuring the energy separation between the native structure and the other decoy structures in the units of the standard deviation of the ensemble. The *Z*-score above is only for a single protein. While we seek the weights of an energy function, we makes *Z*-scores of

"all" proteins simultaneously low enough, that is, negative and large in the absolute value.

We need an objective function that reflects the *Z*-scores of the many proteins in the training set. We have many approaches to develop an objective function. For example, we minimize the summation of *Z*-scores over proteins and we obtain an energy function that gives small *Z*-scores for many proteins but large *Z*-scores for a few proteins. This is not desired because the *Z*-scores of "all" proteins need to be small. This problem arises partly because proteins in the training set have different quality decoys and the current energy function allows some proteins to be more easily recognized than others. Another approach may be to minimize the maximum of *Z*-score of proteins in the training set. However, the optimized weights are probably determined by few proteins, which may be exceptional ones that could be structures with some errors. This tends to lead to unreasonable energy weights. Some intermediate approaches have been proposed by Koretke *et al.* [56] as well as by Mirny and Shakhnovich [57].

To avoid the effects of few proteins and minimize the *Z*-scores of "all" proteins, we chose a normalized *Z*-scores. When the *Z*-scores of some proteins are very small, it is difficult to optimize other proteins which have large *Z*-scores. In order to reduce the ill effect mentioned above, we normalize the *Z*-scores with

$$f(z) = \frac{1}{1+\exp^{-z}}, \tag{18}$$

which *f(z)* maps *Z*-scores to the value among 0 and 1.

Even the *Z*-score is low enough; it does not guarantee the native structure can be distinguished from the decoy set, because *Z*-score is based on the different value between energy value of the native structure and the mean of energy values of all non-native structures in a decoy set. To ensure that energy value of a native structure is the lowest

among a decoy set, we added a related measure *Z'*-score [30] given as

$$Z' = \frac{E_{native} - E_{lowest}}{\Delta E}, \tag{19}$$

where E_{lowest} is the lowest energy value of the non-native structure among the decoy set, and ΔE is the same used in *Z*-score (in Equation 17). In contrast to the *Z*-score, the *Z'*-score gives a quantitative measure of how well separated the native structure is from its lowest energy neighbor from within the decoy set. We applied the same normalized approach into *Z'*-score. Finally, our objective function of evolutionary approach is

$$S = \sum_{i}^{N} (f(z) + f(z')), \tag{20}$$

where *i* is the protein *i*, and *N* is the number of proteins in a training set. We minimize the score *S* to find out optimal weights for energy terms of our energy functions.

2.3 Data Sets of Decoys

A popular approach to test energy functions is to partially sample the conformational space utilizing constructed decoy sets. Publicly available decoy sets not only provide a means to evaluate performance of energy functions, but also permit comparisons between different structure discrimination methods [50, 51, 58]. Many of these decoy sets contain a large number of candidate conformations, with varying degrees of similarity to the native conformation. The goal of developing an efficient energy function is to recognize the best conformation that is most similar to the native structure.

2.3.1 Training Sets

To develop our new energy functions to discriminate the native structures and non-native structures, we used well-developed decoy set, Rosetta-Tsai [59], as our training

sets. The decoy set was obtained from the Baker Laboratory website (ftp://ftp.bakerlab.org/pub/decoys/decoys_11-14-01.tar.gz). These Rosetta-Tsai decoy set is modified from Rosetta all-atom decoy set [60] and generated from fragments of 3 to 9 residues from known structures matched to the targets through a multiple sequence alignment process. These fragments were assembled to native-like structures from a fragment insertion simulated annealing procedure which is using Bayesian scoring functions [12, 36, 61, 62]. The decoy sets increase the number of proteins and frequency of near native models building on side-chains and minimizing clashes.

The decoy set contains structural decoys for 41 different and diverse protein targets. For the training step, we preprocess some error structures in the decoy set. We filter out the proteins with incomplete residues which make inaccuracies of energy functions and the proteins that included in the six testing decoy sets. Table 3 lists the final 30 protein targets and the length ranging from 35 to 85 residues. These protein targets are used for training our energy functions to obtain the overall weights of energy terms.

2.3.2 Testing Sets

We tested the performances of our energy functions with optimal weights on six decoy sets. They are the EMBL misfolded set of Holm and Sander [63] , the 4state_reduced set of Park and Levitt [22], local-minima decoy set of Kesar and Levitt [35], the lattice_ssfit set of Samudrala *et al.* [64], and fisa decoy set of Simons *et al*. [61]. These decoy sets are obtained from the Decoys'R'Us database [35] (http://dd.stanford.edu/). The Rosetta all-atom decoy set [60] could be download from the Baker Laboratory website (ftp://ftp.bakerlab.org/pub/decoys/RosettaAllAtomDecoys.tar.gz). The protein lists of decoy sets are summarized in Tables 3. Since these sets contain thousands of conformations generated by different conformational search algorithms, the performance of an energy

function depends on each set. Success in one decoy set does not guarantee success in another [65]. Therefore, testing on a variety and large number of decoy sets is to provide a rigorous evaluation of how well an energy function works.

The EMBL misfolded set [63] is set of intentionally misfolded structures and their corresponding native structures. This decoy set contains structural decoys for 25 protein targets including from 36 to 317 residues and was constructed by swapping side-chains on pairs of crystallographically determined structures with the same number of residues while keeping the backbone geometry unperturbed. The resulting chimeric structures were then subjected to 500 steps of steepest descent energy minimization using the GROMOS program.

The 4state_reduced [22] decoy set contains structural decoys for 7 small proteins including from 54 to 75 residues with varying topological folds. The protein structures have been generated by exhaustively enumerating the backbone rotamer states of 10 selected residues in each protein using an off-lattice model with four discrete dihedral angles states per rotatable bond. Among hundreds of thousands of generated conformations, only those compact structures scoring well by using a variety of scoring functions and having reasonable RMSD from the native conformation have been selected.

The local-minima decoy set (lmds) [35] also contains structural decoys for 10 small proteins including from 36 to 68 residues. These decoys are conformations that occupy minima in a modified classical ENCAD [66-69] force field using united and soft atoms. Initially ten thousand structures were generated by randomly modifying only the dihedral angles in the loop regions. After minimization in torsion space to the local minima in the potential, up to 500 of the lowest energy conformations for each protein were kept to make up the decoy sets.

The lattice_ssfit [64] decoy set contains structural decoys for 7 proteins including from 55 to 98 residues. These decoy conformations have been generated by *ab initio* protein structure prediction methods. The conformational space of a sequence was exhaustively enumerated on a tetrahedral lattice. A lattice-based scoring function was used to select the 10,000 best-scoring conformations. All-atom structures were then generated by fitting an off-lattice four-state torsion model to the lattice conformations, using predicted secondary structure. The 10,000 conformations were further scored with a combination of RAPDF (residue-specific all-atom probability discriminatory function) [70], HCF (hydrophobic compactness function) [64], and the Shell energy function (one-point-per-residue scoring function) [23]. The 2000 best-scoring conformations for each protein were selected for the decoy sets.

The fisa [61] decoy set contains structural decoys for 4 proteins including from 43 to 76 residues. These decoy conformations are generated by the same protocol as in training set. Use a fragment insertion simulated annealing procedure to assemble native-like structures from fragments of unrelated protein structures with similar local sequences (determined through a multiple sequence alignment process) based on Bayesian scoring functions.

The Rosetta all-atom decoy set [60] was generated from the same protocol which used in the training set and fisa decoy set. Fragments of 3 to 9 residues from known structures matched to the targets through a multiple sequence alignment process were selected to assemble native-like structures by fragment insertion simulated annealing procedure which is using Bayesian scoring functions [12, 36, 61, 62]. The terms of the scoring function included those for hydrophobic burial, electrostatics, disulfide bonds, the packing of α-helices and β-strands, and the formation of β-sheets. During annealing only, another term used to promote compactness was added based on the radius of gyration. The number of

decoy structures in the sets ranges from 994 to 1999. The native structures are not been within the downloaded decoy set and we copy the native structures from PDB database. But, the native structures are different from the decoy structures. The residue sequences of decoy structures are incomplete sequences contrast to the sequences of native structures. We filter the incomplete proteins to avoid inaccuracies of energy calculation. The protein which has difference between the number of residues of decoy structure and ones of native structure is small than 10 and the length of sequence of decoy structure is less than 30 residues was filtered off. We then cut the native sequences to fit the length of the residue sequences of decoy structures to reduce the computational error of our energy functions. Final, 42 proteins were left in the decoy set and include from 34 to 149 residues.

Chapter 3 Results and Discussions

3.1 Parameters in optimization

The parameters of an evolutionary algorithm may conduce to different results. Population sizes and maximum generations are usually adjusted for searching best results in an evolutionary algorithm, but difference may result from the same parameters. In this thesis, we examine many different population sizes and maximum generations to ensue least differences. The final values of weights are not slimily because of different initial random values. To observe the variations of weights, therefore, we consider the weight of electrostatic potential as the base and calculate ratios of other energy terms to electrostatic potential.

Table 4 shows the variations of weights of MOLSIM with different population sizes and fixed maximum generations, 200. Each result of different population sizes is the average value of five independent runs. The standard deviation shows that our evolutionary algorithm can find the best weights and are stable in many different runs. Table 5 lists the similar results for different maximum generations and a fixed population size, 200, excluding maximum generations, 100 and 200 steps. Although the variation of maximum generations, 100 and 200, doesn't like other results, the standard deviation values are enough small.

Tables 6 and 7 list the variations of weights of GEMSCORE. For the same reason, the different values of weights from different initial random values, we use the van der Waals potential as the base to calculate the ratios of other energy terms. The results show that similar conclusion. The evolutionary algorithm we used in this thesis is stable for optimizing the weights of our energy functions. To ensure to find the best weights, we use population sizes, 200, and maximum generations, 300, in our evolutionary algorithm. The

final overall weights of energy terms in two energy functions are presented in Table 8. We use these weights for testing popular benchmarks latter.

3.2 Results of training set

As the description in section 2.2.4 of Materials and Methods, the *Z*-score of a native structure is calculated to describe how far the effective energy of the native structure is separated from the energy spectrum of the decoy structures and we judge the size of energy gap between the native structure and the lowest energy non-native structures with *Z'*-score. Results of the *Z*-score and *Z'*-score from the optimized training decoy set are presented in Table 9. Bold faced values indicate the rank of the native structure is larger than 500.

There are 17 X-Ray and 13 NMR structures in the training set. Figure 4 shows the performances of our energy functions on different structure determination by X-ray and NMR on training set. In Figure 4, Rank1 and Rank5 mean that the native structures rank within top 1 and top 5 among its corresponding decoy set after the training process, respectively. The ranks of the native structures are in the top 10 percent (Rank10%, ~186) and the rest 90 percent (Other) among its corresponding decoy set, respectively. In both Rank1 and Rank5, our methods have better performance in X-Ray structures than NMR structures. MOLSIM presents the results in NMR structure as batter as X-Ray structures in Rank10%. GEMSCORE can identify all X-Ray structures completely, but misidentify the about 40 percent of NMR structures.

In general, our energy functions have better ability to find out the native structures which are X-Ray structures. These worse structures are most NMR structures and contain many models in original PDB file. But only one model was considered as the native structure used in training set. It may be part of reason that the bad performances of our energy functions for some NMR structures. Though, our energy functions don't have the

performances for some NMR structures as well as X-Ray structures, native structures of NMR structures are mostly ranked before the 50 percent among its corresponding decoy set in the training set.

3.3 Test with popular benchmarks

After weights of energy terms are yielded, we tested our optimized energy functions with popular benchmarks. The results of EMBL misfolded decoy set are listed in Table 10. There are 24 proteins identified from 25 native-misfolded protein pairs in MOLSIM. GEMSCORE identify all 25 proteins in this decoy set. Only one protein, 2cdv, is misidentified in MOLSIM. There is no Z-score in EMBL misfolded decoy set because of only one decoy structure in the set.

The results of the *Z*-score and *Z'*-score from the other five testing decoy sets, 4state_reduced, lmds, lattice_ssfit, fisa, and Rosetta all-atom decoy set, are presented in Tables 11-15. The results of total six testing sets compared with previous works based on physics-based energy functions by previous authors [27-32] are summarized in Table 16. There are 44 and 46 native proteins that ranked first in total 71 decoy sets with MOLSIM and GEMSCORE, respectively.

Figure 5 plots correlations between the energies and the all-atom root mean square deviation (RMSD) of decoys from their corresponding native structures in 4state_reduced decoy set on 4state_reduced set. Table 17 shows the details of comparison with previous works in 4state_reduced decoy set. *Z'*-score isn't used in general, therefore, we compare the rank and *Z*-score of native structure. The number before slash is the rank of native structure in this decoy sets. Another is the *Z*-score of native structure. Our results are as well as previous works, which used AMBER, or other force-fields.

There is one native structure, 3icb, could not be distinguished in both our energy functions. Protein 3icb is a vitamin D-dependent calcium-binding protein and contains two calcium ions. Figure 6 shows the location of calcium ions in the three-dimension structure of 3icb. The two Ca^{2+} of 3icb locate in the loop section of native structure and that may be one reason for worse result. Lack of Ca^{2+} at loop could make unstable loop of native structure. Therefore, we refine the native structure with two Ca^{2+} and calculate energy with the same weights again. Because of problems of programs, only GEMSCORE energy function was used to re-calculate the refined structure. Native structure is distinguished from non-native structures after put calcium ions back to the native structure. Result is presented in Table 18.

The results of other three testing sets, lmds, lattice_ssfit, and fisa, are listed in Table 19-21. As 4state_reduced set, the performances of our energy functions are good as previous works. Although there are some native structures that can't be identified from decoys, they are 1fc2 in fisa, 1b0n-B, 1bba, 1dtk and 1fc2 in lmds, 1dkt-A, and 1trl-A in lattice_ssfit. Some proteins, like 1bba and 1fc2 in lmds, were ranked as worse energy. They were missed by previous work [27]. The reason for this massive failure is not entirely clear. Perhaps, this is because 1bba is pancreatic hormone with a long loop. 1fc2 and 1b0n-B are both protein complexes and have many missing coordinates (≥15 residues) in their native structures. Figure 8 shows the whole protein structures of 1fc2 and 1b0n-B. The numbers of residues with coordinates of these three proteins are less than 45.

The others proteins, 1dtk in lmds, 1trl-A in lattice_ssfit, which can't be distinguished from decoys are both NMR structures and contains 20 and 8 models, respectively. One reason for the worse results is that the native structure adopted in decoy sets may not fit for our energy functions. The same results that performances of our energy functions are weak for NMR structures are presented in training set. Figure 7, 9, and 10 plot correlations between the energies and the all-atom RMSD of decoys from their corresponding native structures on lmds, lattice_ssfit, fisa decoy sets, respectively.

The same result that has better performance in X-Ray structures than NMR structures on the Rosetta all-atom decoy set. The result is presented in Table 15. There are 20 X-Ray and 22 NMR structures in the decoy set. Native structures are most identified using MOLSIM. 16 and 18 native structures are ranked within top 1 and top 5, respectively. There are 19 native structures ranked first using GEMSCORE and one misidentified structure is ranked four. However, our methods have the worse performance in NMR structures. MOLSIM identified 9 and 12 native structures from decoy structures within top 1 and top 5, respectively. 7 and 10 native structures are identified as top 1 and top 5 using GEMSCORE in total 22 structures.

Figure 11 summaries the results on five testing decoy sets excluded EMBL misfolded decoy set. There are 45 X-Ray structures and 26 NMR structures in these decoy sets. Top1 and Top5 mean that the native structure ranks within top 1 and top 5 among its corresponding decoy set, respectively. The ranks of the native structures are in the top 10 percent (Top10%) and the rest 90 percent (Other) among its corresponding decoy set, respectively. GEMSCORE presents the better power in distinguishing X-Ray structures than MOLSIM. MOLSIM have the better performance to identify NMR structures than GEMSCORE and distinguish X-Ray structures as well as GEMSCORE.

3.4 Hydrogen-bonding interactions in GEMSCORE

Hydrogen-bonding interactions are important interactions during protein folding. In this thesis, however, we only consider the hydrogen-bonding interactions on backbone as a part of GEMSCORE. The original energy function using in GEMDOCK used the hydrogen-bonding interactions as the main potential during protein-ligand docking. Nevertheless, it may not suit to apply the hydrogen-bonding potential used in GEMDOCK for protein folding directly. Table 22 compares the results on six testing decoy sets by

GEMSCORE with hydrogen-bonding potential (E_{HB}) and hydrogen-bonding potential on backbone (E_{bHB}). E_{HB} is a potential of all hydrogen-bonding interactions and calculate energy with the same protocol of E_{bHB}. Definition of hydrogen-bonding potential, E_{bHB}, is described in section 2.1.2 (in Equation 7). The result of GEMSCORE with E_{HB} haves performance as well as GEMSCORE with E_{bHB} excluding 4state_reduced decoy set.

We analyzed the result on 4state_reduced decoy set with different hydrogen-bonding potentials, E_{HB}, E_{bHB}, and E_{nbHB}. Table 23 presents the result on 4state_reduced decoy set. E_{HB} means the potential of all hydrogen-bonding interactions. E_{bHB} and E_{nbHB} mean the potentials of hydrogen-bonding interactions on backbone and hydrogen-bonding interactions which do not locate on backbone, respectively. (i.e., $E_{HB} = E_{bHB} + E_{nbHB}$). Result of E_{nbHB} has terrible performance than E_{bHB}. In the analysis on training set, we can see the same result as the result on 4state_reduced decoy set. Figure 10 shows the percentages of native structures ranked within 1, 5, 100, 500, 1000 and larger 1000 in the training set with different hydrogen-bonding potentials, E_{HB}, E_{bHB}, and E_{nbHB}.

The terrible result on E_{nbHB} may be caused from the definition of hydrogen-bonding interactions used in the energy function of GEMDOCK. In order to reduce computational time of GEMDOCK, a simplify strategy was adopt. It does not to add hydrogen atoms and assign hydrogen bonds depended on nitrogen atoms and oxygen atoms during energy calculation. Hydrogen bonds are assigned based on the distance between nitrogen atom and oxygen atom. However, it is impossible to from hydrogen bond at some angle of N-H-O (θ_{NHO}), even the distance is enough close. When the all hydrogen-bonding interactions are divided to hydrogen-bonding interactions on backbone and not on backbone, E_{nbHB} contains the all inaccuracy hydrogen-bonding interactions described above and has terrible performance.

This strategy works fine in protein-lignad docking because of the few hydrogen-bonding interactions and space restrictions. Nevertheless, many hydrogen bonds from secondary structures to stable protein structure during protein folding. It is effective to consider the hydrogen-bonding potential for developing energy function. The result (Figure 12) shows that the major contribution of hydrogen-bonding potential is from hydrogen-bonding potential on backbone. It also proves why the performance of potential of all hydrogen-bonding interactions (E_{HB}) as well as potential of hydrogen-bonding interactions on backbone (E_{bHB}). It is enough to cover the error caused from the potential of hydrogen-bonding interactions which do not locate on backbone (E_{nbHB}).

3.5 Solvent effects in MOLSIM/GEMSCORE

The original energy functions we used in this thesis don't contain solvation potential. However, solvent effect plays an important role to stable the protein structure in the native situation. We added extra solvation potential to improve our energy functions. Table 24 present the performances of MOLSIM and GEMSCORE. Original and Optimized mean the original energy function without optimization and the optimized energy function without solvation energy, respectively. Optimized+E_{SAS} means the optimized energy function with solvation energy. We can see a lot of improvement in MOLSIM energy function. Not only average *Z*-scores but also the number of native structures ranked first in a decoy set has best performance with optimized energy function that contains solvation energy function.

GEMSCORE also presents the same result in average *Z*-scores. But the number of native structure ranked number one doesn't have variation excluding EMBL misfolded and 4state_reduced decoy sets. The reason for this is not entirely clear. It may the solvation energy we adopted couldn't be suitable for GEMSCORE energy function. Nevertheless, additionally solvation potential improve the performances of our energy functions whether

MOLSIM or GEMSCORE.

3.6 MOLSIM vs. GEMSCORE

In the thesis, we show two energy functions. One is simplified physics-based energy function, MOLSIM, and another is empirical energy function, GEMSCORE. We optimized these energy functions for protein folding. Both of them show the ability that identifies the native structure from lots of non-native structures. These two energy functions have similar energy terms and optimized by the same optimization strategy, however, it still has some differences between them.

Computational resources and time are main discrepancies. MOLSIM is all-atom energy function and need extra information of hydrogen atoms during energy calculation. Although, only specified hydrogen atoms are needed in MOLSIM, evaluate the coordinates of these hydrogen atoms and energy minimization after adding hydrogen atoms require more computational resources and time than GEMSCORE. And these extra hydrogen atoms also increase the computation when energy value is evaluated. For example, GEMSCORE need 1.80 seconds in calculating energy values of 1ctf which contains 65 residues and 486 atoms. MOLSIM costs 3.31 seconds with 1ctf. When evaluate a large protein, like 2tmn that includes 316 residues and 2431 atoms, GEMSCORE (10.7 seconds) still fast then MOLSIM (18.8 seconds).

Nevertheless, MOLSIM has better performance then GEMSCORE in most decoy sets. MOLSIM also presents better ability when identifies NMR structures. Some researches believe that molecular dynamic is needed with all-atom energy function; especially extra hydrogen atoms are added. Molecular dynamic with a short time can provide reliable structure. In this thesis, no molecular dynamic is adopted in MOLSIM developed originally for molecular dynamics. Molecular dynamic may have more accuracy, but more

computational resources and time are needed during energy calculation. That is disadvantage of most physics-based energy functions. However, GEMSCORE, an empirical energy function based on physical mechanisms with simplified model, presents a good performance with popular benchmarks and costs less computational resources and time then physics-based energy function. It may be more suitable for protein structure prediction system to select the best structure from thousands and thousands candidate structures.

Chapter 4 Conclusions and Future Works

4.1 Conclusions

In this thesis, we develop two simple energy functions for protein folding and have well performance on six popular benchmarks. We adopt only few weights to optimize the energy functions with few terms and the performance as well as previous works. However, most of previous used AMBER or other force fields. The costs of computational resources and time are huge and parameters used to optimize the force fields during optimization are hundreds and thousands. The difficulty and time are as more as numbers of parameters in optimization.

The parameters of energy function based on physics mechanical are usually optimized from chemical compounds or small proteins. The optimal parameters may not reflect the native protein structures. We use the information from well-developed decoy set and obtain these optimal parameters of energy function using evolutionary algorithm. This would yield energy functions that are practically powerful for many purposes and should be conceptually helpful for gaining insight into the physical principles of protein architecture.

4.2 Future Works

In the near future, we will apply our derived energy functions for protein structure prediction with CASP6 (Critical Assessment of Techniques for Protein Structure Prediction, http://predictioncenter.llnl.gov/casp6/Casp6.html) targets which many researches have submitted in the world. By this progress, we hope to prove the ability of our energy functions which selects the best structures from many different candidate generation methods. After examination, we will integrate our energy functions with other methods that generate candidate structures and pick up the best candidate structure from thousands and

thousands of ones. Finally, we will build a reliable system to predict protein structure from only protein sequences.

Besides building a reliable system to predict protein structure, many shortcomings can be reformed in our energy functions. A better training set may increase the performance of our methods. Sovation energy used in this thesis is ASPs. In past years, GB was proved to have good performance for solvent effect. It may have better results by combining GB and ASPs. More accuracy and high-speed are also needed for calculating atomic solvent-accessible surface area.

Table 1. The energy terms and descriptions in MOLSIM and GEMSCORE

Term Name	Description
ELC	Electrostatic energy
VDW	van der Waals potential
BHB	Hydrogen-bonding potential on backbone
C	Summation of surface area of each C atom
S	Summation of surface area of each S atom
O	Summation of surface area of each O atom
nO	Summation of surface area of each negative charged O atom in ASP (OD1 and OD2), GLU (OE1 and OE2)
N	Summation of surface area of each N atom
pN	Summation of surface area of each positive charged N atom in HIS (ND1 and NE2), ARG (NH1 and NH2), LYS (NZ)

Table 2. Atomic formal charge used in GEMSCORE

Formal charge	Atom	Description
0.5	N	N atom in HIS (ND1 and NE2)
0.5	N	N atom in ARG (NH1 and NH2)
1.0	N	N atom in LYS (NZ)
-0.5	O	O atom in ASP (OD1 and OD2)
-0.5	O	O atom in GLU (OE1 and OE2)
0	Other atoms	

Table 3. The training set and testing sets

Training Set				
Name	No. of Decoys	No. of X-Ray [a]	No. of NMR [b]	List of Protein ID (PDB ID)
RosettaTasi	1867(30) [c]	17	13	1a32, 1aa3, 1afi, 1ail, 1am3, 1bq9, 1bw6, 1cc5, 1cei, 1dol, 1hyp, 1kjs, 1lfb, 1msi, 1mzm, 1nkl, 1nre, 1pou, 1ptq, 1res 1sro, 1tif, 1utg, 1uxd, 1vcc, 1vif, 2ezh, 2fow, 2fxb, 2ptl
Testing Sets				
Name	No. of Decoys	No. of X-Ray	No. of NMR	List of Protein ID (PDB ID)
EMBL misfold	1(25)	24	1	1bp2(2paz) [d], 1cbh(1ppt), 1fdx(5rxn), 1hip(2b5c), 1lh1(2ilb), 1p2p(1rn3), 1ppt(1cbh), 1rei(5pad), 1rhd(2cyp), 1rn3(1p2p), 1sn3(2ci2), 1sn3(2cro), 2b5c(1hip), 2cdv(2ssi), 2ci2(1sn3), 2ci2(2cro), 2cro(1sn3), 2cro(2ci2), 2cyp(1rhd), 2i1b(1lh1), 2paz(1bp2), 2ssi(2cdv), 2tmn(2ts1), 2ts1(2tmn), 5pad(1rei)
4state_reduced	665(7)	7	0	1ctf, 1r69, 1sn3, 2cro, 3icb, 4pti, 4rxn
lmds	434(10)	8	2	1b0n-B, 1bba, 1ctf, 1dtk, 1fc2, 1igd, 1shf-A, 2cro, 2ovo, 4pti
lattice_ssfit	2000(7)	6	2	1beo, 1ctf, 1dkt-A, 1fca, 1nkl, 1pgb, 1trl-A
fisa	500(4)	4	0	1fc2, 1hdd-C, 2cro, 4icb
RosettaAll	1024(42)	20	22	1aa2, 1acf, 1ag2, 1aj3, 1apf, 1ark, 1ayj, 1bdo, 1bor, 1btb, 1ctf, 1dec, 1erv, 1fbr, 1fwp, 1gb1, 1gpt, 1gvp, 1ksr, 1kte, 1mbd, 1pdo, 1r69, 1ris, 1svq, 1tit, 1tul, 1vls, 1vtx, 1who, 1wiu, 2acy, 2cdx, 2erl, 2ezk, 2fdn, 2gdm, 2ktx, 2ncm, 2pac, 4fgf, 5pti

[a, b] No. of X-Ray and No. of NMR are the numbers of the native structures determination by X-ray and NMR, respectively

[c] Average numbers of decoy sets and the values in the parentheses are the number of proteins in the decoy set

[d] The PDB code of a crystal structures and the PDB IDs in the parentheses are their corresponding intentionally misfolded structures

Table 4. Variations of parameters of MOLSIM with different population sizes and the maximum number of generations of FCEA is set to 200

Energy Term [a]	Pop 100	Pop 200	Pop 300	Pop 400	Pop 500
	AVG [b] ± STD [c]	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
ELC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
VDW	3.189 ± 0.001	3.189 ± 0.002	3.188 ± 0.001	3.189 ± 0.001	3.189 ± 0.001
C	0.392 ± 0.001	0.392 ± 0.001	0.391 ± 0.001	0.392 ± 0.001	0.391 ± 0.001
S	0.779 ± 0.000	0.779 ± 0.000	0.779 ± 0.000	0.779 ± 0.000	0.779 ± 0.000
O	-0.161 ± 0.000	-0.161 ± 0.000	-0.161 ± 0.000	-0.161 ± 0.000	-0.161 ± 0.000
nO	-0.085 ± 0.000	-0.085 ± 0.000	-0.085 ± 0.000	-0.085 ± 0.000	-0.085 ± 0.000
N	-0.161 ± 0.000	-0.161 ± 0.000	-0.161 ± 0.000	-0.161 ± 0.000	-0.161 ± 0.000
pN	-0.592 ± 0.000	-0.592 ± 0.000	-0.591 ± 0.001	-0.592 ± 0.000	-0.591 ± 0.001

[a] These terms defined in Table 1

[b, c] The average (AVG) and standard deviation (STD) were calculated from five independent runs.

Table 5. Variations of parameters of MOLSIM with different maximum number of generations and the population size is set to 200

Energy Term [a]	Gen 100	Gen 200	Gen 300	Gen 400	Gen 500	Gen 1000
	AVG [b] ± STD [c]	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
ELC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
VDW	3.191 ± 0.007	3.188 ± 0.001	3.189 ± 0.000	3.189 ± 0.000	3.189 ± 0.000	3.189 ± 0.000
C	0.391 ± 0.001	0.391 ± 0.001	0.392 ± 0.000	0.392 ± 0.000	0.392 ± 0.000	0.392 ± 0.000
S	0.779 ± 0.002	0.779 ± 0.000	0.779 ± 0.000	0.779 ± 0.000	0.779 ± 0.000	0.779 ± 0.000
O	-0.161 ± 0.001	-0.161 ± 0.001	-0.161 ± 0.000	-0.161 ± 0.000	-0.161 ± 0.000	-0.161 ± 0.000
nO	-0.085 ± 0.001	-0.085 ± 0.000	-0.085 ± 0.000	-0.085 ± 0.000	-0.085 ± 0.000	-0.085 ± 0.000
N	-0.161 ± 0.001	-0.161 ± 0.001	-0.161 ± 0.000	-0.161 ± 0.000	-0.161 ± 0.000	-0.161 ± 0.000
pN	-0.593 ± 0.002	-0.591 ± 0.001	-0.592 ± 0.000	-0.592 ± 0.000	-0.592 ± 0.000	-0.592 ± 0.000

[a] These terms defined in Table 1

[b, c] The average (AVG) and standard deviation (STD) were calculated from five independent runs.

Table 6. Variations of parameters of GEMSCORE with different population sizes and the maximum number of generations of FCEA is set to 200

Energy Term [a]	Pop 100	Pop 200	Pop 300	Pop 400	Pop 500
	AVG [b] ± STD [c]	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
ELC	5.846 ± 0.006	5.849 ± 0.003	5.844 ± 0.003	5.845 ± 0.006	5.843 ± 0.003
VDW	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
BHB	3.245 ± 0.003	3.249 ± 0.005	3.247 ± 0.003	3.246 ± 0.003	3.245 ± 0.003
C	0.406 ± 0.000	0.407 ± 0.000	0.406 ± 0.001	0.406 ± 0.001	0.406 ± 0.000
S	1.132 ± 0.001	1.134 ± 0.001	1.133 ± 0.001	1.132 ± 0.002	1.132 ± 0.001
O	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000
nO	-0.338 ± 0.001	-0.337 ± 0.000	-0.337 ± 0.001	-0.338 ± 0.001	-0.338 ± 0.000
N	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000
pN	-0.678 ± 0.001	-0.678 ± 0.001	-0.678 ± 0.001	-0.678 ± 0.000	-0.678 ± 0.000

[a] These terms defined in Table 1

[b, c] The average (AVG) and standard deviation (STD) were calculated from five independent runs.

Table 7. Variations of weights of GEMSCORE with different maximum number of generations and the population size is set to 200

Energy Term [a]	Gen 100	Gen 200	Gen 300	Gen 400	Gen 500	Gen 1000
	AVG [b] ± STD [c]	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD	AVG ± STD
ELC	5.860 ± 0.039	5.846 ± 0.003	5.845 ± 0.000	5.845 ± 0.000	5.845 ± 0.000	5.845 ± 0.000
VDW	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
BHB	3.284 ± 0.034	3.244 ± 0.006	3.244 ± 0.000	3.244 ± 0.000	3.244 ± 0.000	3.244 ± 0.000
C	0.411 ± 0.002	0.406 ± 0.000	0.406 ± 0.000	0.406 ± 0.000	0.406 ± 0.000	0.406 ± 0.000
S	1.144 ± 0.008	1.132 ± 0.002	1.132 ± 0.000	1.132 ± 0.000	1.132 ± 0.000	1.132 ± 0.000
O	-0.298 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000
nO	-0.336 ± 0.008	-0.338 ± 0.000	-0.338 ± 0.000	-0.338 ± 0.000	-0.338 ± 0.000	-0.338 ± 0.000
N	-0.298 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000	-0.297 ± 0.000
pN	-0.684 ± 0.007	-0.678 ± 0.001	-0.678 ± 0.000	-0.678 ± 0.000	-0.678 ± 0.000	-0.678 ± 0.000

[a] These terms defined in Table 1

[b, c] The average (AVG) and standard deviation (STD) were calculated from five independent runs.

Table 8. The final weights of MOLSIM and GEMSCORE for energy terms

Energy Term [a]	MOLSIM	GEMSCORE
ELC	1.000	5.845
VDW	4.207	1.000
BHB	—	3.244
C	0.403	0.406
S	0.841	1.132
O	-0.144	-0.297
nO	-0.254	-0.338
N	-0.144	-0.297
pN	-0.586	-0.678

[a] These terms defined in Table 1

Table 9. Content of the training decoy set

PDB ID	Type	Structure	N_{res} [a]	N_{decoys} [b]	MOLSIM			GEMSCORE		
					Z [c]	Z' [d]	Rank [e]	Z	Z'	Rank
1a32	a	X-RAY	65	1610	-0.64	2.53	398	-2.04	0.53	20
1ail	a	X-RAY	67	1807	**-0.47**	**1.92**	**572**	-3.72	-0.69	1
1am3	a	X-RAY	57	1898	**-0.47**	**2.87**	**606**	-1.33	1.69	157
1bq9	b	X-RAY	53	1825	-3.07	-0.81	1	-4.89	-1.47	1
1cc5	a	X-RAY	76	1892	-1.82	1.48	52	-1.49	1.63	124
1cei	a	X-RAY	85	1897	-1.63	0.48	11	-4.84	-1.86	1
1dol	ab	X-RAY	62	1871	-3.91	-1.11	1	-4.27	-1.19	1
1hyp	a	X-RAY	75	1893	-1.99	1.01	43	-2.12	1.10	31
1lfb	a	X-RAY	69	1893	-1.04	1.55	279	-1.40	1.91	167
1msi	b	X-RAY	60	1894	-3.79	-0.95	1	-3.61	-1.03	1
1mzm	a	X-RAY	71	1934	-0.99	1.94	311	-2.31	0.39	15
1ptq	ab	X-RAY	43	1885	-1.93	0.67	9	-1.89	1.22	55
1tif	ab	X-RAY	59	1849	-3.87	-0.63	1	-4.86	-1.69	1
1utg	a	X-RAY	62	1897	**0.04**	**2.65**	**1009**	-1.56	1.90	111
1vcc	ab	X-RAY	77	1857	-4.38	-1.51	1	-5.58	-2.48	1
1vif	b	X-RAY	48	1896	-2.28	0.37	4	-3.31	0.12	2
2fxb	ab	X-RAY	81	1800	-3.52	-0.68	1	-4.16	-0.85	1
1aa3	ab	NMR	56	1865	-2.68	0.40	4	-1.02	1.85	289
1afi	ab	NMR	72	1824	-0.77	-0.02	1	-3.81	-0.84	1
1bw6	a	NMR	56	1900	-2.01	1.70	38	-1.74	1.54	72
1kjs	a	NMR	74	1893	-0.70	2.31	480	**-0.05**	**2.96**	**914**
1nkl	a	NMR	70	1898	-1.04	1.92	290	-3.48	-0.67	1
1nre	a	NMR	66	1893	-1.88	1.52	38	-4.06	-0.68	1
1pou	a	NMR	70	1898	-1.51	1.29	84	-1.77	1.08	66
1res	a	NMR	35	1723	**-0.47**	**2.57**	**563**	**0.63**	**3.47**	**1288**
1sro	b	NMR	66	1881	-1.51	1.04	76	-1.20	1.86	225
1uxd	a	NMR	43	1896	-2.59	-0.01	1	-2.63	-0.33	1
2ezh	a	NMR	65	1893	-1.73	1.35	45	-1.31	1.37	156
2fow	ab	NMR	66	1834	-1.11	1.91	231	**0.06**	**3.08**	**982**
2ptl	ab	NMR	60	1835	-2.61	-0.19	1	-1.92	0.96	32
	Averages		64	1867	-1.88	0.92		-2.52	0.50	

[a, b] N_{res}, number of residues in the protein; N_{decoys}, number of decoys in the decoy set

[c, d] Z-score of native structure and Z'-score of native structure in the decoy set

[e] Rank, the ranking of the native structure in the decoy set

Bold faced values indicate the rank of the native structure is larger than 500

Table 10. Content of the EMBL misfolded decoy set

Sequence PDB ID	Backbone PDB ID	RMSD	N_{res} [a]	MOLSIM		GEMSCORE	
				E_{native} [b]	$E_{misfold}$ [c]	E_{native}	$E_{misfold}$
1bp2	2paz	15.79	123	-5031.52	-4642.28	-2286.71	-1439.91
1cbh	1ppt	10.62	36	-1228.36	-868.79	-282.12	-61.01
1fdx	5rxn	9.52	54	-1747.71	-1388.31	-262.63	50.99
1hip	2b5c	14.62	85	-3166.54	-2509.60	-1154.69	-297.39
1lh1	2ilb	17.81	153	-6081.44	-5609.86	-3235.29	-1849.85
1p2p	1rn3	18.97	124	-4900.35	-4523.35	-1873.18	-1461.13
1ppt	1cbh	10.87	36	-1134.21	-995.96	-631.69	-151.09
1rei	5pad	18.65	107	-9081.74	-7721.60	-4476.66	-2504.34
1rhd	2cyp	22.21	293	-12212.07	-10728.13	-5224.37	-4255.46
1rn3	1p2p	18.77	124	-5388.80	-4710.54	-2490.50	-1891.59
1sn3	2ci2	13.64	65	-2451.59	-2108.86	-910.46	-415.65
1sn3	2cro	11.25	65	-2451.59	-1971.55	-910.46	-388.97
2b5c	1hip	14.74	93	-3478.60	-3127.09	-1961.70	-1138.25
2cdv	2ssi	14.60	107	**-3745.05**	**-4026.40**	-1048.87	-946.45
2ci2	1sn3	13.43	83	-2282.23	-2111.40	-1073.66	-746.40
2ci2	2cro	11.68	83	-2282.23	-2173.94	-1073.66	-907.07
2cro	1sn3	11.89	71	-2473.87	-2258.42	-1232.25	-733.32
2cro	2ci2	11.24	71	-2473.87	-2152.53	-1232.25	-687.41
2cyp	1rhd	21.91	294	-13070.15	-11356.26	-7111.14	-4372.74
2i1b	1lh1	17.87	153	-6628.59	-5327.57	-3426.45	-2160.80
2paz	1bp2	15.59	123	-5038.94	-4135.44	-2344.92	-1189.98
2ssi	2cdv	15.10	113	-3717.92	-2829.38	-834.46	-226.07
2tmn	2ts1	23.06	317	-14650.89	-12035.76	-8156.72	-4917.97
2ts1	2tmn	23.12	317	-13753.93	-12888.26	-7482.59	-6405.27
5pad	1rei	19.09	214	-9125.48	-7472.12	-3744.01	-2659.26
	Averages	15.84	132				

[a] N_{res}, number of residues in the protein

[b, c] E_{native}, energy value of native structure in the decoy set; $E_{misfold}$, energy value of misfolded structure in the decoy set

Bold faced values indicate the error results

Table 11. Content of the 4state_reduced decoy set

PDB ID	Description	N_{res} [a]	RMSD Range	N_{decoys} [b]	MOLSIM			GEMSCORE		
					Z [c]	Z' [d]	Rank [e]	Z	Z'	Rank
1ctf	C-terminal domain of ribosomal protein L7/L12	68	2.2–10.2	631	-3.43	-0.57	1	-5.41	-2.73	1
1r69	N-terminal domain of phage 434 repressor	63	2.3–9.5	676	-2.68	-0.21	1	-3.14	-0.37	1
1sn3	Scorpion toxin variant	65	2.5–10.5	661	-2.72	-0.09	1	-2.95	-0.21	1
2cro	Phage 434 Cro protein	65	2.1–9.7	675	-3.41	-0.37	1	-3.42	-0.08	1
3icb	Vitamin D-dependent calcium-binding protein	75	1.8–10.7	654	-2.11	0.31	3	-1.40	1.16	39
4pti	Trypsin inhibitor	58	2.8–10.8	688	-4.10	-1.30	1	-4.24	-1.45	1
4rxn	Rubredoxin	54	2.6–9.3	678	-3.83	-1.16	1	-4.56	-1.94	1
	Averages	64		666	-3.18	-0.48		-3.59	-0.80	

[a, b] N_{res}, number of residues in the protein; N_{decoys}, number of decoys in the decoy set

[c, d] Z-score of native structure and Z'-score of native structure in the decoy set

[e] Rank, the ranking of the native structure in the decoy set

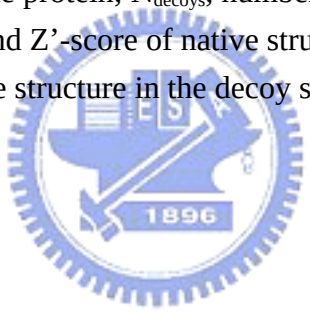


Table 12. Content of the local minima decoy set

PDB ID	Description	N_{res} [a]	RMSD Range	N_{decoys} [b]	MOLSIM			GEMSCORE		
					Z [c]	Z' [d]	Rank [e]	Z	Z'	Rank
1b0n-B	Sinr protein/Sini protein complex	31	2.45–6.03	498	-1.17	1.36	73	-0.83	1.98	110
1bba	Bovine pancreatic polypeptide	36	2.78–8.91	501	5.23	8.03	501	2.91	5.60	498
1ctf	C-terminal domain of ribosomal protein L7/L12	68	3.59–12.5	498	-3.98	-1.79	1	-6.25	-3.59	1
1dtk	DendrotoxinK	57	4.32–12.6	216	-1.76	0.51	7	-1.41	1.38	17
1fc2	Immunoglobulin Fc and Fragment B Of Protein A Complex	43	3.99–8.45	501	1.91	5.67	487	2.16	5.31	494
1igd	ProteinG	61	3.11–12.6	501	-4.10	-1.31	1	-5.89	-3.31	1
1shf-A	Fyn proto-oncogene tyrosine kinase	59	4.39–12.3	438	-4.84	-2.13	1	-5.23	-2.16	1
2cro	434 cro protein	65	3.87–13.5	501	-6.32	-3.22	1	-5.49	-2.53	1
2ovo	Ovomucoid third domain	56	4.38–13.4	348	-3.22	-0.29	1	-3.20	-0.48	1
4pti	Trypsin inhibitor	58	4.94–13.2	344	-5.30	-2.50	1	-5.08	-2.33	1
	Averages	53		435	-2.36	0.43		-2.83	-0.01	

[a, b] N_{res}, number of residues in the protein; N_{decoys}, number of decoys in the decoy set

[c, d] Z-score of native structure and Z'-score of native structure in the decoy set

[e] Rank, the ranking of the native structure in the decoy set

Table 13. Content of the lattice_ssfit decoy set

PDB ID	Description	N_{res} [a]	RMSD Range	N_{decoys} [b]	MOLSIM			GEMSCORE		
					Z [c]	Z' [d]	Rank [e]	Z	Z'	Rank
1beo	Beta-cryptogein	98	7.0–15.6	2001	-5.36	-2.68	1	-8.44	-5.46	1
1ctf	C-terminal domain of ribosomal protein L7/L12	68	5.5–12.8	2001	-4.90	-1.35	1	-8.38	-4.75	1
1dkt-A	Type 1 human cyclin-dependent kinase subunit	72	6.7–14.1	2001	-2.57	0.25	3	-3.31	-0.28	1
1fca	Ferredoxin from clostridium Acidurici	55	5.1–11.4	2001	-5.43	-1.90	1	-5.61	-2.64	1
1nkl	Nk-lysin from pig	78	5.3–13.6	2001	-4.49	-1.85	1	-8.35	-4.92	1
1pgb	ProteinG (B1 lgG-binding domain)	56	5.8–12.9	2001	-5.84	-3.15	1	-8.94	-5.76	1
1trl-A	Thermolysin fragment	62	5.4–12.5	2001	-2.12	0.30	13	-2.90	0.47	8
4icb	Calbindin-binding Protein	65	4.7–12.9	2001	-2.82	-0.54	1	-3.35	0.25	5
	Averages	69		2001	-4.39	-1.48		-6.16	-2.89	

[a, b] N_{res}, number of residues in the protein; N_{decoys}, number of decoys in the decoy set

[c, d] Z-score of native structure and Z'-score of native structure in the decoy set

[e] Rank, the ranking of the native structure in the decoy set

Table 14. Content of the fisa decoy set

PDB ID	Description	N_{res} [a]	RMSD Range	N_{decoys} [b]	MOLSIM			GEMSCORE		
					Z [c]	Z' [d]	Rank [e]	Z	Z'	Rank
1fc2	Immunoglobulin Fc and Fragment B Of Protein A Complex	43	2.1–10.3	501	-0.37	1.72	184	0.15	2.47	299
1hdd-C	Engrailed Homeodomain	57	2.8–12.9	501	-3.06	-0.10	1	-2.37	0.84	4
2cro	Phage 434 Cro protein	65	4.3–12.6	501	-3.54	-1.16	1	-3.82	-1.45	1
4icb	Calbindin-binding Protein	76	4.8–14.1	501	-4.87	-2.40	1	-4.12	-1.55	1
	Averages	60		501	-2.96	-0.49		-2.54	0.08	

[a, b] N_{res}, number of residues in the protein; N_{decoys}, number of decoys in the decoy set

[c, d] Z-score of native structure and Z'-score of native structure in the decoy set

[e] Rank, the ranking of the native structure in the decoy set

Table 15. Content of the Rosetta all-atom decoy set

PDB ID	Structure	N_{nres} [a]	N_{dres} [b]	N_{decoys} [c]	MOLSIM			GEMSCORE		
					Z [d]	Z' [e]	Rank [f]	Z	Z'	Rank
1aa2	X-RAY	109	105	999	-4.79	-2.65	1	-6.50	-3.78	1
1acf	X-RAY	126	123	1999	-5.65	-2.93	1	-7.99	-4.99	1
1bdo	X-RAY	81	75	999	-5.15	-2.36	1	-6.85	-3.87	1
1ctf	X-RAY	69	67	999	-3.34	-0.64	1	-6.08	-3.14	1
1erv	X-RAY	106	105	999	-5.34	-2.82	1	-7.12	-4.56	1
1gvp	X-RAY	88	82	997	-1.18	1.42	102	-3.63	-0.25	1
1kte	X-RAY	106	100	998	-3.59	-1.28	1	-6.05	-3.45	1
1mbd	X-RAY	154	147	999	-4.91	-2.37	1	-8.62	-5.86	1
1pdo	X-RAY	130	121	999	-4.74	-2.12	1	-7.60	-4.21	1
1r69	X-RAY	64	61	999	-2.49	0.31	3	-3.17	-0.52	1
1ris	X-RAY	98	92	999	-3.60	-0.91	1	-6.34	-3.91	1
1tul	X-RAY	103	97	999	-4.67	-2.38	1	-5.11	-2.69	1
1vls	X-RAY	147	143	999	-2.87	0.06	2	-4.95	-2.04	1
1who	X-RAY	95	88	999	-4.00	-1.87	1	-6.59	-4.07	1
2acy	X-RAY	99	92	994	-5.02	-2.68	1	-7.71	-4.92	1
2erl	X-RAY	41	35	999	-1.29	1.19	93	-2.49	0.22	4
2fdn	X-RAY	56	55	999	-3.89	-0.96	1	-3.91	-1.10	1
2gdm	X-RAY	154	149	999	-0.85	-0.32	1	-8.16	-5.55	1
4fgf	X-RAY	125	121	999	-5.73	-3.43	1	-7.88	-4.89	1
5pti	X-RAY	59	55	999	-3.79	-1.37	1	-4.33	-1.30	1
1ag2	NMR	104	97	998	-3.71	-1.68	1	-6.29	-3.67	1
1aj3	NMR	99	95	999	**0.94**	**3.52**	**831**	-1.34	1.65	74
1apf	NMR	50	47	999	-2.23	0.41	6	-2.53	0.43	3
1ark	NMR	61	55	998	-2.81	0.00	1	-1.55	1.43	51
1ayj	NMR	51	46	999	-1.47	1.53	53	-0.94	1.88	166
1bor	NMR	57	52	999	-2.51	0.05	3	-1.80	1.25	23
1btb	NMR	90	89	999	-4.72	-2.34	1	-5.82	-3.00	1
1dec	NMR	40	35	999	-2.57	0.23	3	-3.69	-0.73	1
1fbr	NMR	94	93	999	-4.34	-1.50	1	-5.26	-2.27	1
1fwp	NMR	70	66	999	-0.57	0.00	1	-1.21	1.33	101
1gb1	NMR	57	54	999	-1.94	0.81	15	-2.60	0.46	3
1gpt	NMR	48	47	999	-1.71	0.74	28	-2.17	0.91	12
1ksr	NMR	101	92	999	-3.53	-0.77	1	-2.00	0.89	15
1svq	NMR	95	90	999	-2.14	0.08	3	-1.40	1.24	69
1tit	NMR	90	85	999	-3.99	-1.02	1	-2.82	-0.06	1

Table 15. Content of the Rosetta all-atom decoy set (cont.)

PDB ID	Structure	N_{nres} [a]	N_{dres} [b]	N_{decoys} [c]	MOLSIM			GEMSCORE		
					Z [d]	Z' [e]	Rank [f]	Z	Z'	Rank
1vtx	NMR	43	36	999	-0.11	2.43	476	**0.51**	**3.18**	**718**
1wiu	NMR	94	90	999	-4.19	-0.73	1	-5.18	-1.31	1
2cdx	NMR	61	54	997	-1.58	1.13	34	-1.80	0.98	26
2ezk	NMR	94	93	998	-1.71	0.90	22	-2.12	0.63	7
2ktx	NMR	39	34	999	-0.92	1.85	192	-2.15	0.19	4
2ncm	NMR	100	96	998	-5.02	-2.28	1	-6.88	-3.55	1
2pac	NMR	83	77	999	-0.90	1.88	176	-0.90	1.88	195
Averages		86	81	1022	-3.06	-0.54		-4.31	-1.46	

[a, b] N_{nres}, number of residues in the native structure; N_{dres}, number of residues in the decoy structures

[c] N_{decoys}, number of decoys in the decoy set

[d, e] Z-score of native structure and Z'-score of native structure in the decoy set

[f] Rank, the ranking of the native structure in the decoy set

Bold faced values indicate the rank of the native structure is larger than 500

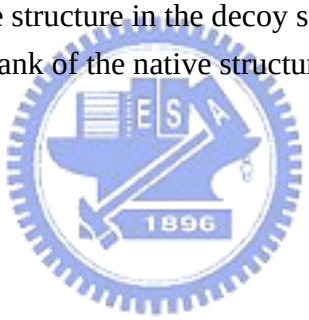


Table 16. Summary of the our results and comparison with previous works on six testing data sets

Decoy Set	MOLSIM	GEMSCORE	Zhou *et al.* [27]	Zhu *et al.* [28]	Fujitsuka *et al.* [29]	Lee *et al.* [30]	Hsieh *et al.* [31]	Zhang *et al.* [32]	
								DFIRE-SCM	DFIRE-allatom
misfold	24/25 [a] (N/A [b])	25/25 (N/A)	25/25 (N/A)			24/25 (N/A)	18/19 (N/A)		
4state	6/7 (-3.18)	6/7 (-3.59)	6/7 (3.49)	N/A (-3.96)	5/7 (-2.36)	N/A (-4.95)	6/7 (-3.43)	6/7 (3.94)	6/7 (3.49)
lmds	6/10 (-2.36)	6/10 (-2.83)	7/10 (0.67)	N/A (-1.75)	4/6 (-3.85)	N/A (-4.49)		3/10 (2.56)	7/10 (0.90)
lattice	6/8 (-4.19)	6/8 (-6.16)	8/8 (8.94)	N/A (-4.08)		N/A (-6.75)		8/8 (6.19)	8/8 (9.47)
fisa	3/4 (-2.96)	2/4 (-2.54)	3/4 (4.49)	N/A (-3.09)		N/A (-2.09)		3/4 (4.70)	3/4 (4.80)
RosettaAll	25/42 (-3.06)	26/42 (-4.31)							

[a] The first number is the number of native structures ranked number one; the second number is total number of proteins in the decoy set.

[b] The numbers in parentheses are the average Z-scores. There is no average Z-score in EMBL misfolded decoy set because of only one decoy structure in the set

Table 17. Comparison of our methods with previous works on 4state_reduced decoy set

PDB ID	MOLSIM	GEMSCORE	Zhou *et al.* [27]	Zhu *et al.* [28]	Fujitsuka *et al.* [29]	Lee *et al.* [30]	Hsieh *et al.* [31]
1ctf	1/-3.43[a]	1/-5.41	1/3.86	-/-3.33	1/-2.50	-/-4.26	1/-2.76
1r69	1/-2.68	1/-3.14	1/4.23	-/-3.63	1/-2.50	-/-5.35	1/-3.01
1sn3	1/-2.68	1/-2.95	1/3.79	-/-5.70	1/-3.20	-/-6.33	1/-4.88
2cro	1/-3.41	1/-3.42	1/3.29	-/-3.55	1/-2.30	-/-5.11	1/-2.58
3icb	3/-2.11	39/-1.40	4/2.28	-/-1.97	3/-1.60	-/-2.86	21/-2.06
4pti	1/-4.03	1/-4.24	1/3.62	-/-5.09	1/-2.70	-/-5.35	1/-4.77
4rxn	1/-3.83	1/-4.56	1/3.33	-/-4.43	12/-1.70	-/-5.36	1/-3.95

[a] The first number is the rank of native structure; the second number is Z-score of native structure in the decoy set. The absolute value of Z-score is the larger the better

Table 18. Comparison of our methods with previous works on 4state_reduced decoy set with refined 3icb

PDB ID	MOLSIM	GEMSCORE	Zhou *et al.* [27]	Zhu *et al.* [28]	Fujitsuka *et al.* [29]	Lee *et al.* [30]	Hsieh *et al.* [31]
1ctf	1/-3.43[a]	1/-5.41	1/3.86	-/-3.33	1/-2.50	-/-4.26	1/-2.76
1r69	1/-2.68	1/-3.14	1/4.23	-/-3.63	1/-2.50	-/-5.35	1/-3.01
1sn3	1/-2.68	1/-2.95	1/3.79	-/-5.70	1/-3.20	-/-6.33	1/-4.88
2cro	1/-3.41	1/-3.42	1/3.29	-/-3.55	1/-2.30	-/-5.11	1/-2.58
3icb	3/-2.11	1/-3.78	4/2.28	-/-1.97	3/-1.60	-/-2.86	21/-2.06
4pti	1/-4.03	1/-4.24	1/3.62	-/-5.09	1/-2.70	-/-5.35	1/-4.77
4rxn	1/-3.83	1/-4.56	1/3.33	-/-4.43	12/-1.70	-/-5.36	1/-3.95

[a] The first number is the rank of native structure; the second number is Z-score of native structure in the decoy set. The absolute value of Z-score is the larger the better

Table 19. Comparison of our methods with previous works on local minima decoy set

PDB ID	MOLSIM	GEMSCORE	Zhou *et al.* [27]	Zhu *et al.* [28]	Fujitsuka *et al.* [29]	Lee *et al.* [30]	Hsieh *et al.* [31]
1b0n-B	73/-1.17[a]	110/-0.82	430/-1.17	-/-2.55	N/A	-/-0.61	N/A
1bba	501/ 5.22	498/ 2.91	501/-16.3	N/A	N/A	-/ 4.99	N/A
1ctf	1/-3.98	1/-6.25	1/ 3.54	-/-4.38	1/-4.70	-/-5.12	N/A
1dtk	7/-1.76	17/-1.41	1/ 2.62	-/-3.51	2/-2.30	-/-6.10	N/A
1fc2	487/ 1.91	494/ 2.16	501/-5.72	-/-0.22	N/A	-/-3.38	N/A
1igd	1/-4.10	1/-5.89	1/ 5.16	-/-5.80	1/-6.20	-/-6.16	N/A
1shf-A	1/-4.83	1/-5.22	1/ 6.68	-/-7.53	N/A	-/-8.26	N/A
2cro	1/-6.32	1/-5.49	1/ 4.70	-/-5.97	1/-4.00	-/-8.03	N/A
2ovo	1/-3.22	1/-3.20	1/ 3.21	-/-4.51	1/-4.10	-/-6.00	N/A
4pti	1/-5.30	1/-5.08	1/ 3.96	-/-7.04	17/-1.80	-/-6.24	N/A

[a] The first number is the rank of native structure; the second number is Z-score of native structure in the decoy set. The absolute value of Z-score is the larger the better

Table 20. Comparison of our methods with previous works on lattice_ssfit decoy set

PDB ID	MOLSIM	GEMSCORE	Zhou *et al.* [27]	Zhu *et al.* [28]	Fujitsuka *et al.* [29]	Lee *et al.* [30]	Hsieh *et al.* [31]
1beo	1/-5.36[a]	1/-8.44	1/12.09	-/-4.86	N/A	-/-7.95	N/A
1ctf	1/-4.90	1/-8.38	1/10.05	-/-3.22	N/A	-/-6.98	N/A
1dkt-A	3/-2.55	1/-3.31	1/ 6.87	-/-5.89	N/A	-/-6.40	N/A
1fca	1/-5.43	1/-5.61	1/ 7.18	-/-5.89	N/A	-/-8.30	N/A
1nkl	1/-4.49	1/-8.35	1/ 9.29	-/-3.97	N/A	-/-2.60	N/A
1pgb	1/-5.84	1/-8.94	1/11.87	-/-2.66	N/A	-/-9.55	N/A
1trl-A	13/-2.12	8/-2.91	1/12.09	-/-4.27	N/A	-/-5.49	N/A
4icb	1/-2.82	5/-3.35	1/10.05	-/-1.85	N/A	N/A	N/A

[a] The first number is the rank of native structure; the second number is Z-score of native structure in the decoy set. The absolute value of Z-score is the larger the better

Table 21. Comparison of our methods with previous works on fisa decoy set

PDB ID	MOLSIM	GEMSCORE	Zhou *et al.* [27]	Zhu *et al.* [28]	Fujitsuka *et al.* [29]	Lee *et al.* [30]	Hsieh *et al.* [31]
1fc2	184/-0.37[a]	299/ 0.15	254/0.23	-/-1.94	N/A	-/ 0.23	N/A
1hdd-C	1/-3.06	4/-2.37	1/4.50	-/-2.86	N/A	-/-3.62	N/A
2cro	1/-3.54	1/-3.82	1/6.33	-/-4.54	N/A	-/-5.31	N/A
4icb	1/-4.87	1/-4.12	1/6.91	-/-3.00	N/A	-/-2.26	N/A

[a] The first number is the rank of native structure; the second number is Z-score of native structure in the decoy set. The absolute value of Z-score is the larger the better

Table 22. Comparison of GEMSCORE with different hydrogen-bond potentials, E_{HB} and E_{bHB}, on six testing sets

Decoy Set	GEMSCORE	
	E_{HB} [a]	E_{bHB} [b]
misfold	25/25 [c] (N/A [d])	25/25 (N/A)
4state	3/7 (-3.12)	6/7 (-3.59)
lmds	6/10 (-3.22)	6/10 (-2.83)
lattice	6/8 (-4.87)	6/8 (-6.16)
fisa	2/4 (-2.97)	2/4 (-2.54)
RosettaAll	29/42 (-4.92)	26/42 (-4.31)

[a, b] E_{HB}, potential of all hydrogen-bonding interactions; E_{bHB} (in Equation 7), potential of hydrogen-bonding interactions in backbone

[c] The first number is the number of native structures ranked number one; the second number is total number of proteins in the decoy set.

[d] The numbers in parentheses are the average Z-scores. There is no average Z-score in EMBL misfolded decoy set because of only one decoy structure in the set

Table 23. Comparison of different hydrogen-bonding potentials, E_{HB}, E_{bHB}, and E_{nbHB} on 4state_reduced decoy set

PDBID	E_{HB} [a]			E_{bHB} [b]			E_{nbHB} [c]		
	Z [d]	Z' [e]	Rank [f]	Z	Z'	Rank	Z	Z'	Rank
1ctf	-1.09	1.66	105	-3.15	-0.76	1	1.37	4.43	575
1r69	0.98	3.88	568	-0.89	1.89	135	2.23	5.07	663
1sn3	-0.49	2.65	200	-2.57	0.96	7	1.26	4.52	592
2cro	-1.80	1.01	26	-1.95	1.35	22	-0.67	2.54	168
3icb	0.14	2.65	356	0.03	2.59	362	0.17	2.91	365
4pti	0.91	4.32	567	-1.85	1.32	26	2.41	5.35	684
4rxn	-0.25	3.95	276	-2.87	0.42	5	1.37	5.36	615
Average	-0.23	2.87		-1.89	1.11		1.16	4.31	

[a, b, c] E_{HB}, potential of all hydrogen-bonding interactions; E_{bHB}, potential of hydrogen-bonding interactions on backbone; E_{nbHB}, the potential of hydrogen-bonding interactions which do not locate on backbone (i.e., $E_{HB} = E_{bHB} + E_{nbHB}$)

[d, e] Z-score of native structure and Z'-score of native structure in the decoy set

[f] Rank, the ranking of the native structure in the decoy set

Table 24. Solvation effects on MOLSIM and GEMSCORE

Decoy Set	MOLSIM			GEMSCORE		
	Original [a]	Optimized [b]	Optimized + E_{sas} [c]	Original	Optimized	Optimized + E_{sas}
misfold	20/25[d] (N/A[e])	23/25(N/A)	24/25 (N/A)	18/25 (N/A)	18/25 (N/A)	25/25 (N/A)
4state	2/7 (-1.53)	0/7 (-0.79)	6/7 (-3.18)	1/7 (-1.98)	3/7 (-2.14)	6/7 (-3.59)
lmds	2/10 (-0.73)	8/10 (-3.22)	6/10 (-2.36)	8/10 (-4.28)	8/10 (-4.34)	6/10 (-2.83)
lattice	6/8 (-2.45)	6/8 (-1.22)	6/8 (-4.19)	6/8 (-3.76)	6/8 (-4.07)	6/8 (-6.16)
fisa	3/4 (-1.53)	3/4 (-0.80)	3/4 (-2.96)	2/4 (-3.12)	2/4 (-3.29)	2/4 (-2.54)

[a, b, c] Original, the original energy function without optimization; Optimized, the optimized energy function without solvation energy; Optimized + E_{sas}, the optimized energy function with solvation energy

[d] The first number is the number of native structures ranked number one; the second number is total number of proteins in the decoy set.

[e] The numbers in parentheses are the average Z-scores. There is no average Z-score in EMBL misfolded decoy set because of only one decoy structure in the set

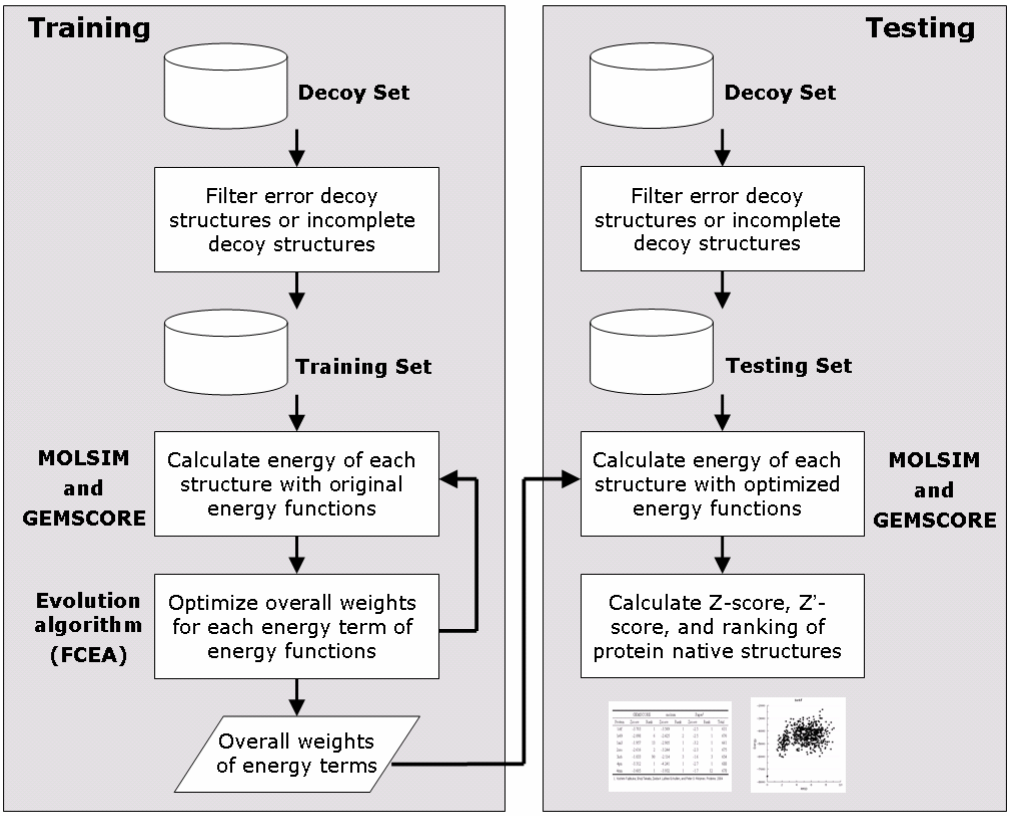


Figure 1. The flowchart of training step and testing step, including decoy sets preparation, energy calculation, and optimization.

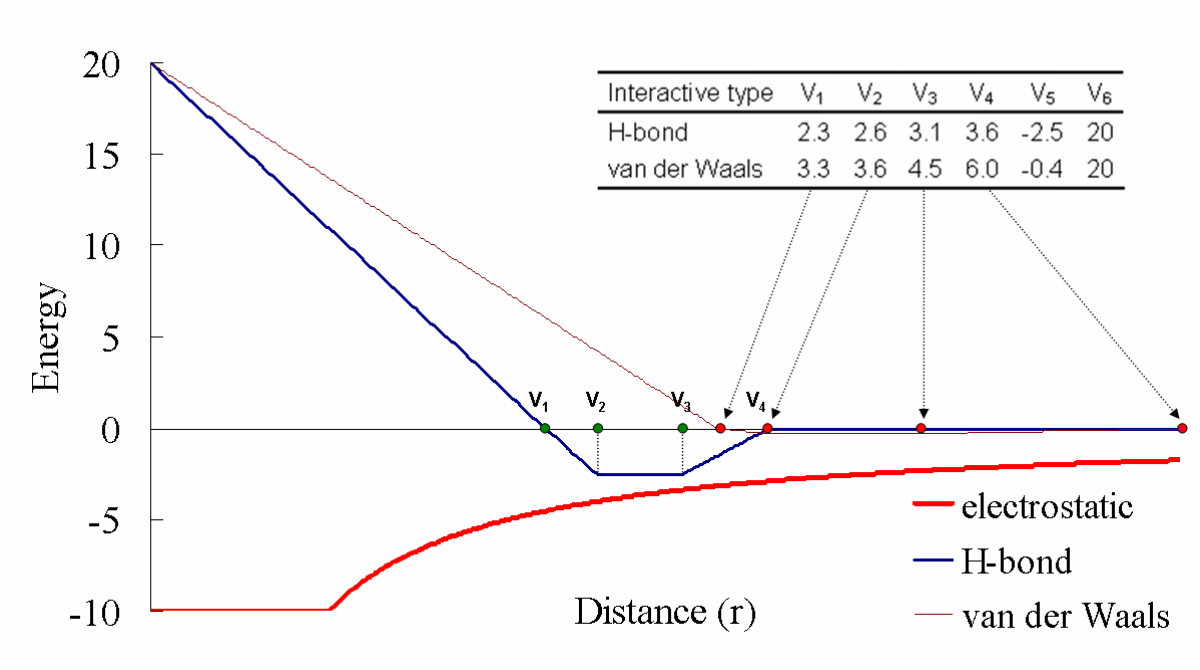


Figure 2. The linear energy function of the pairwise atoms for the van der Waals interactions and hydrogen bonds, and electrostatic potential in GEMSCORE.

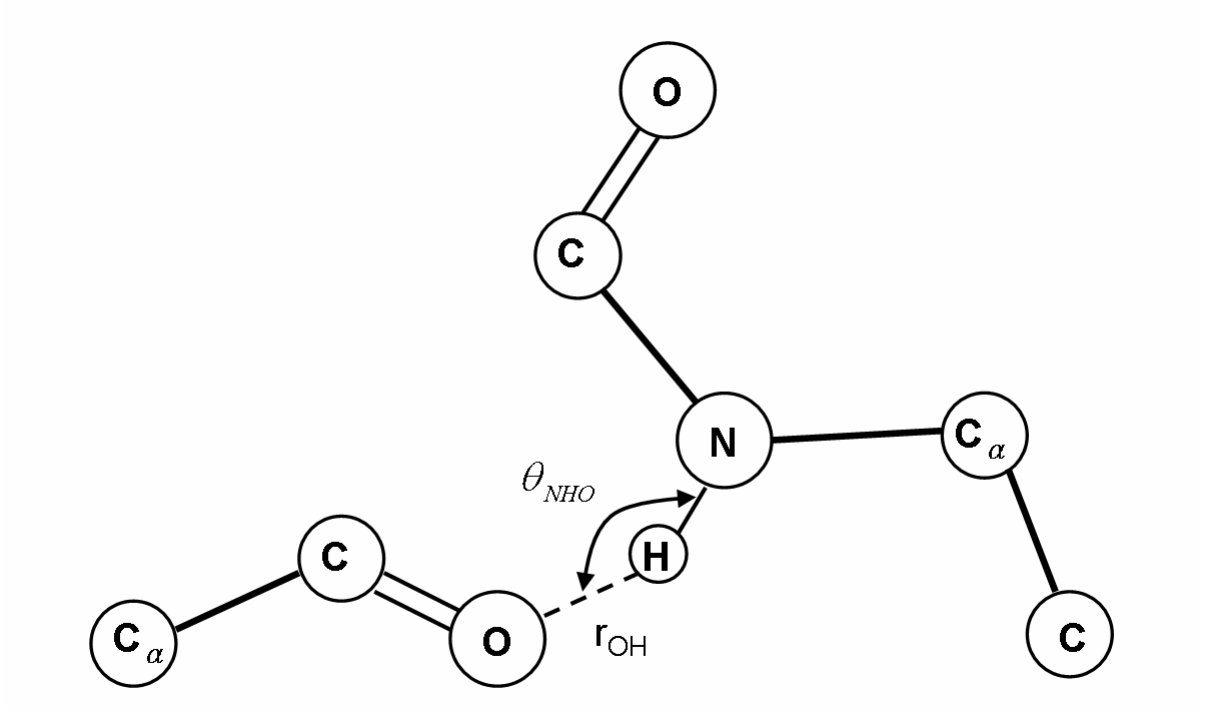


Figure 3. The definition of a hydrogen bond used in GEMSCORE. A hydrogen bond is assigned if the angle of N-H-O (θ_{NHO}) is more than 120° and the distance between hydrogen atom and oxygen atom (r_{OH}) is less than 2.5 Å.

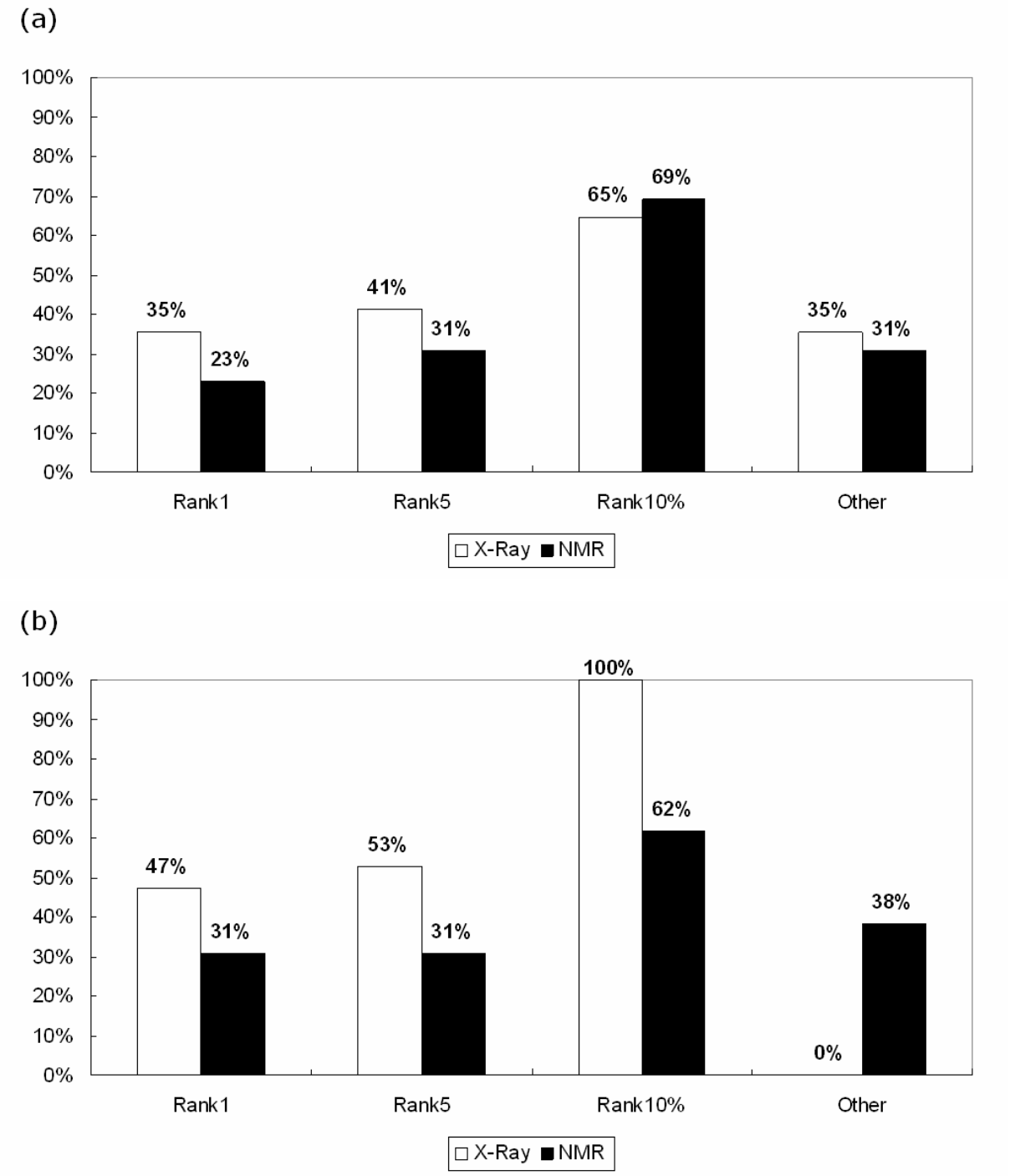


Figure 4. Performances of (a) MOLSIM and (b) GEMSCORE on different structure determination by X-ray and NMR on training set, including 17 X-Ray structures and 13 NMR structures. Rank1 and Rank5 mean that the native structures rank within top 1 and top 5 among its corresponding decoy set after the training process, respectively. The ranks of the native structures are in the top 10 percent (Rank10%) and the rest 90 percent (Other) among its corresponding decoy set after the training process, respectively.

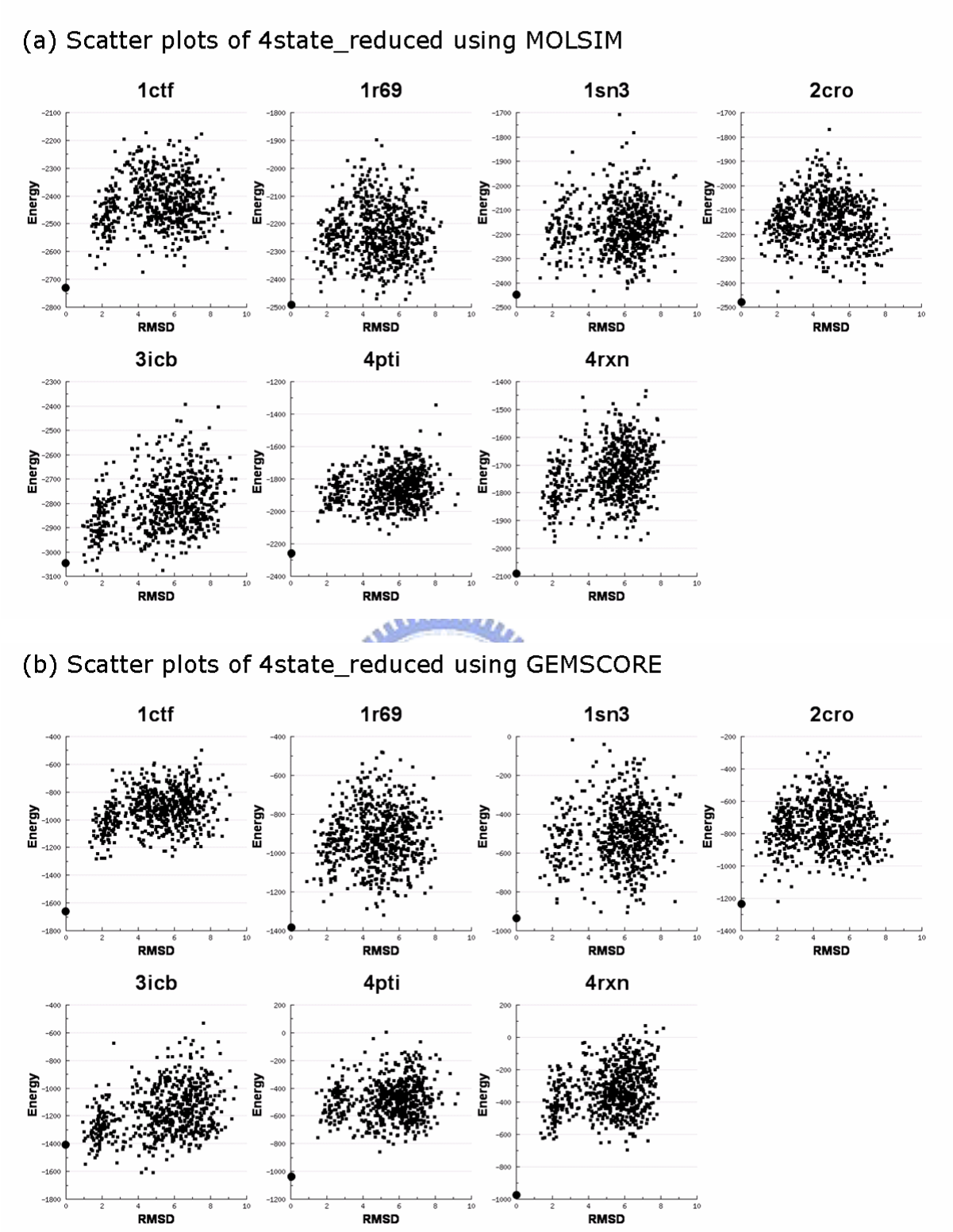


Figure 5. The scatter plots of correlations between the energies and the all-atom RMSD of decoys from their corresponding native structures in 4state_reduced decoy set using (a) MOLSIM energy function and (b) GEMSCORE energy function. The value of RMSD is zero means the native structure of the protein target.

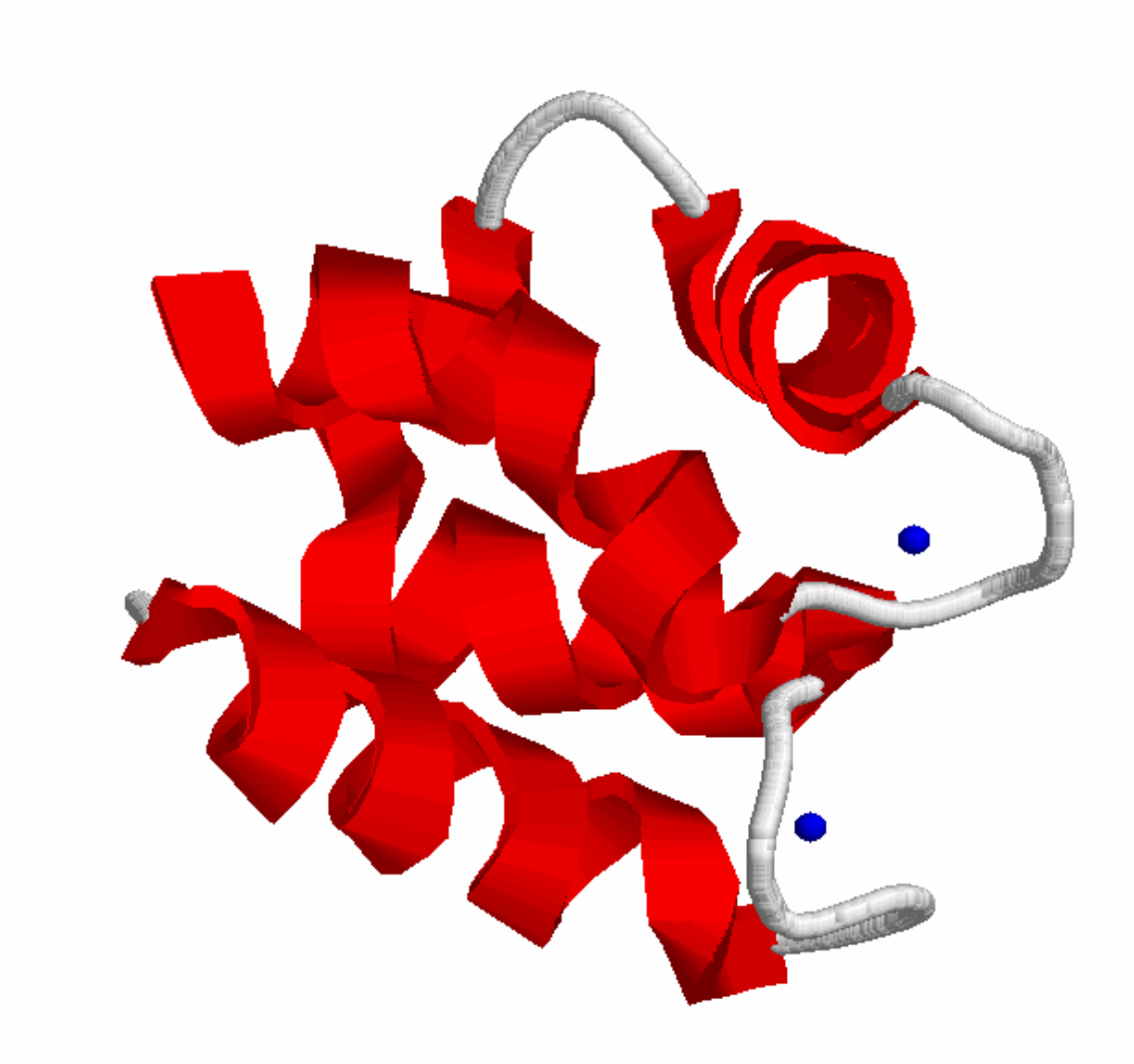
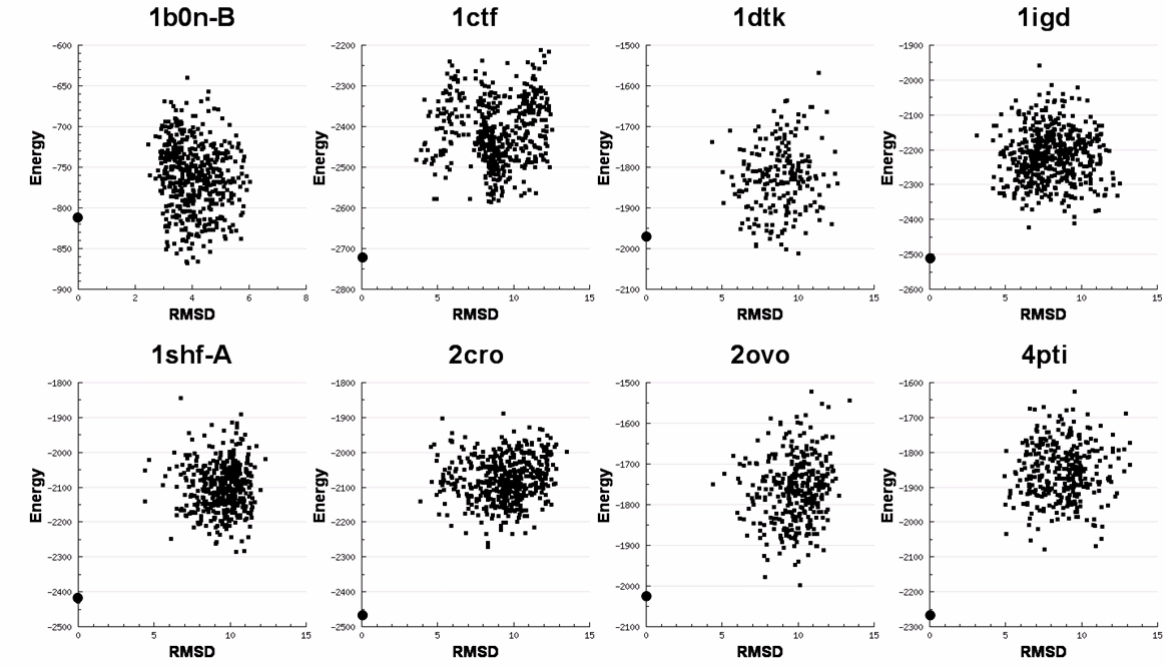


Figure 6. The location of calcium ions in the three-dimension structure of protein 3icb in 4state_reduced decoy set. These two Ca^{2+} (blue) stay at the loop between helixes. One reason for undistinguished native structure may be lack of Ca^{2+} at loop to stable the native structure.

(a) Scatter plots of lmds using MOLSIM



(b) Scatter plots of lmds using GEMSCORE

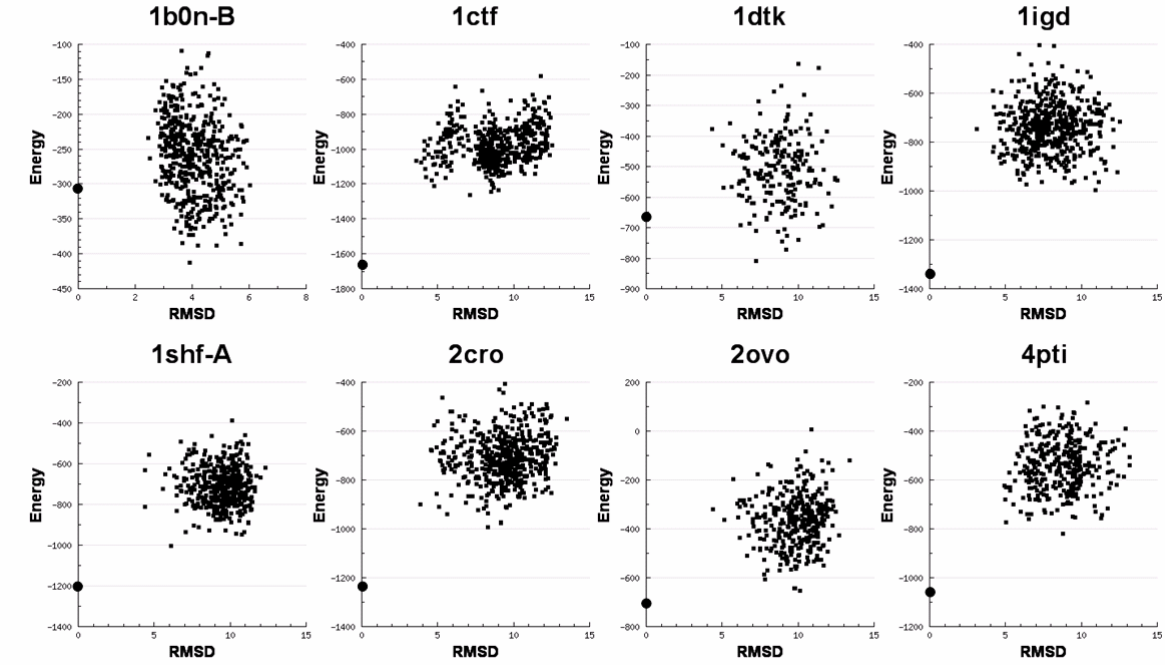


Figure 7. The scatter plots of correlations between the energies and the all-atom RMSD of decoys from their corresponding native structures in lmds decoy set using (a) MOLSIM energy function and (b) GEMSCORE energy function. The value of RMSD is zero means the native structure of the protein target.

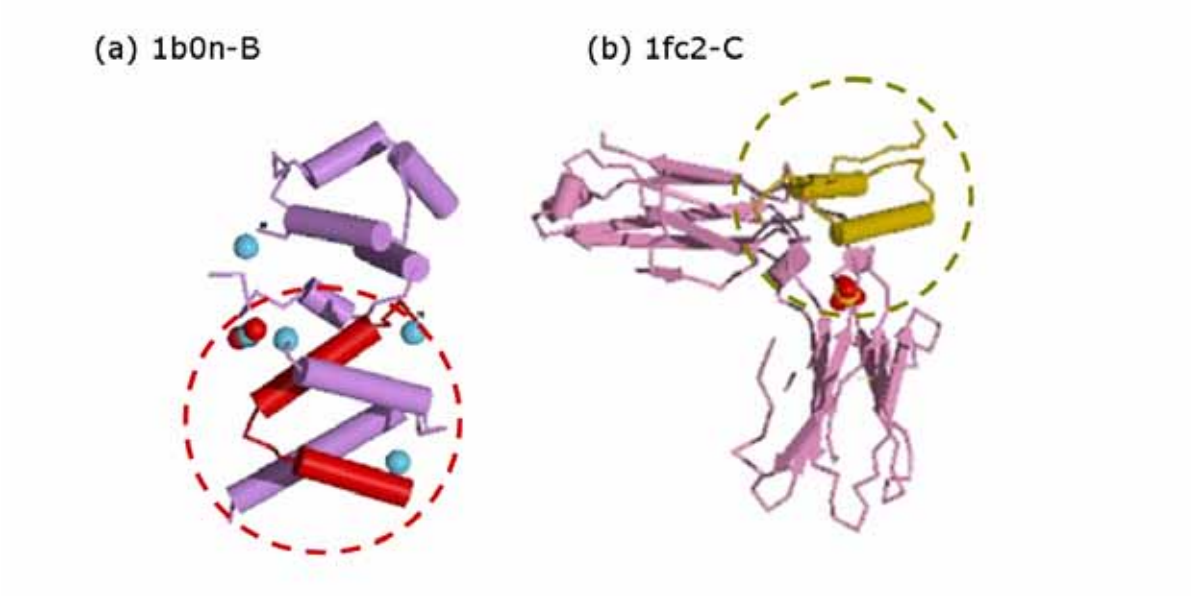
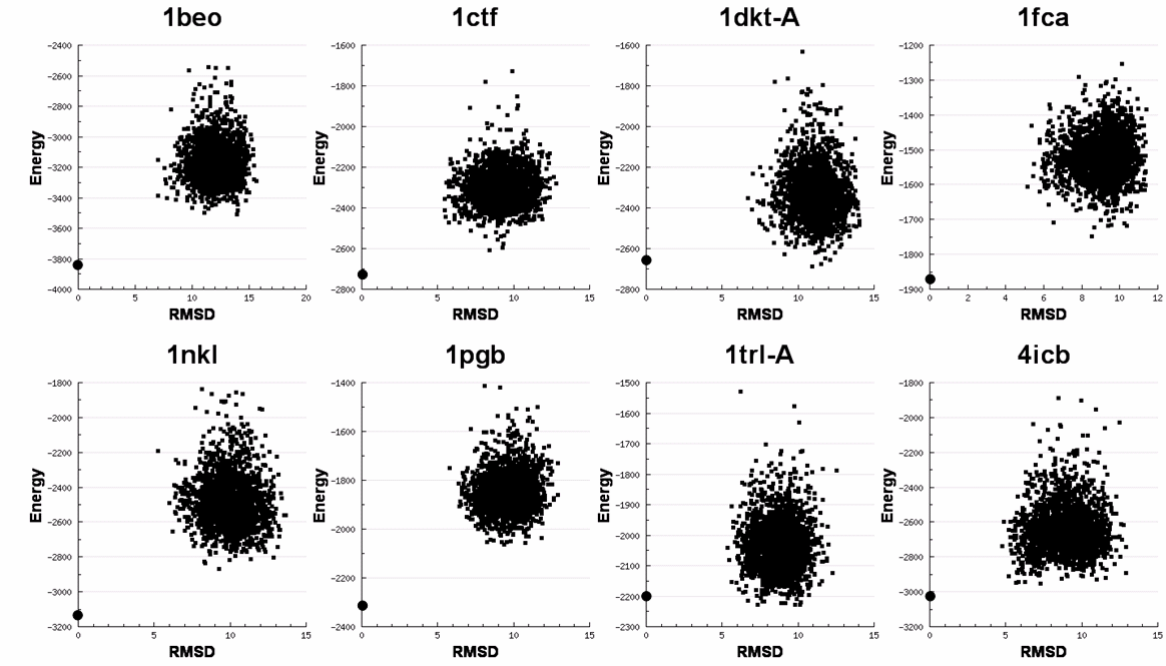


Figure 8. The native structures of misidentified targets, (a) 1b0n-B and (b) 1fc2-C, in local minima decoy set (lmds). They are both protein complexes and the misidentified target is one chain of whole protein. It causes the errors during energy calculation with only single chain. This may be the reason that our energy functions misidentify these protein targets.

(a) Scatter plots of lattice_ssfit using MOLSIM



(b) Scatter plots of lattice_ssfit using GEMSCORE

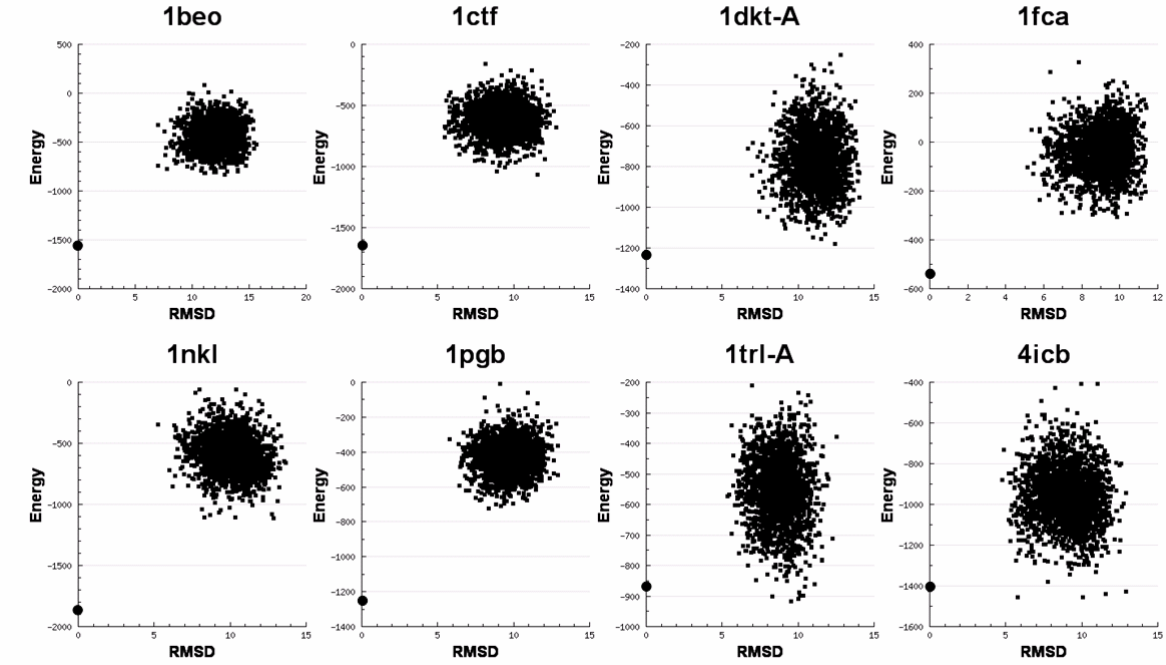
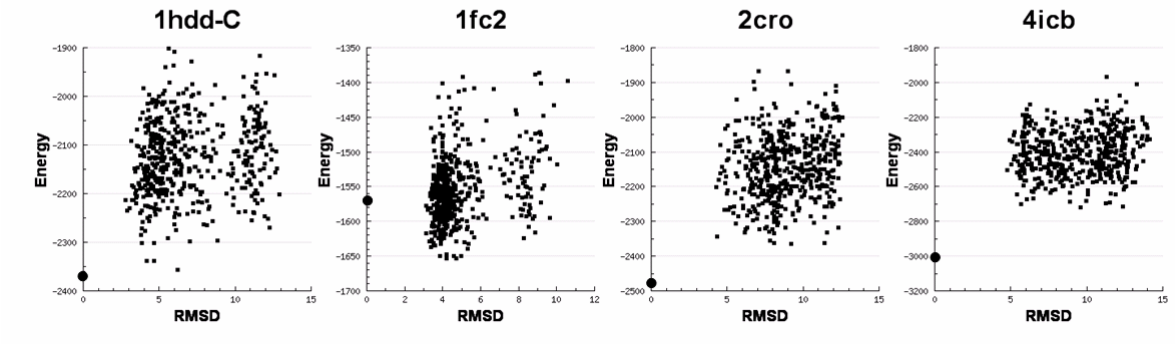


Figure 9. The scatter plots of correlations between the energies and the all-atom RMSD of decoys from their corresponding native structures in lattice_ssfit decoy set using (a) MOLSIM energy function and (b) GEMSCORE energy function. The value of RMSD is zero means the native structure of the protein target.

(a) Scatter plots of fisa using MOLSIM



(b) Scatter plots of fisa using GEMSCORE

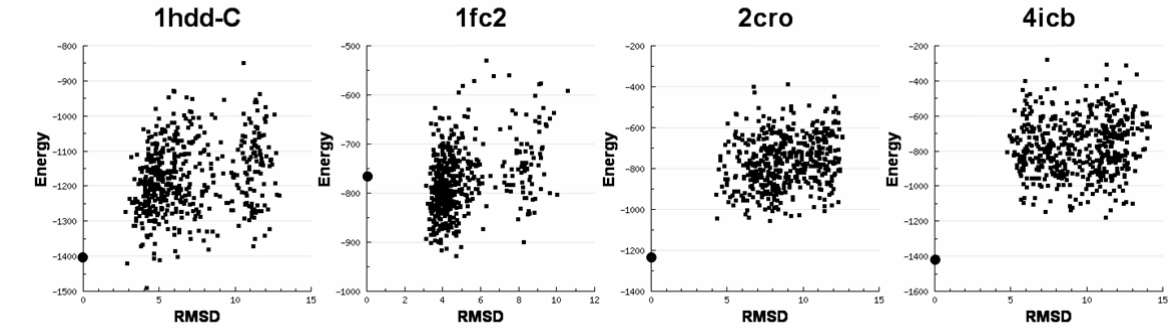


Figure 10. The scatter plots of correlations between the energies and the all-atom RMSD of decoys from their corresponding native structures on fisa decoy set using (a) MOLSIM energy function and (b) GEMSCORE energy function. The value of RMSD is zero means the native structure of the protein target.

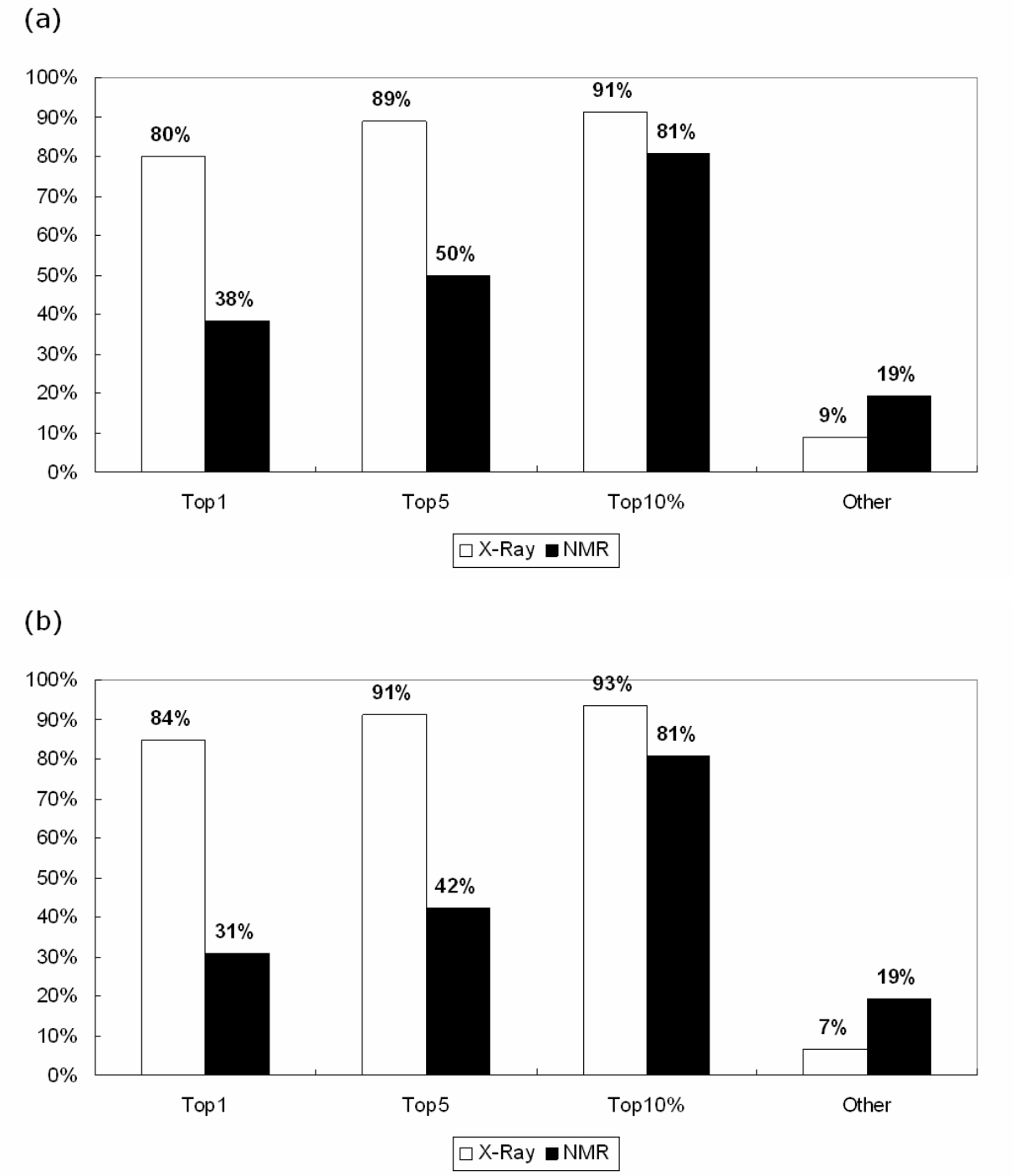


Figure 11. Performances of (a) MOLSIM and (b) GEMSCORE on different structure determination by X-ray and NMR on five testing sets, including 45 X-Ray structures and 26 NMR structures. Top1 and Top5 mean that the native structure ranks within top 1 and top 5 among its corresponding decoy set, respectively. The ranks of the native structures are in the top 10 percent (Top10%) and the rest 90 percent (Other) among its corresponding decoy set, respectively.

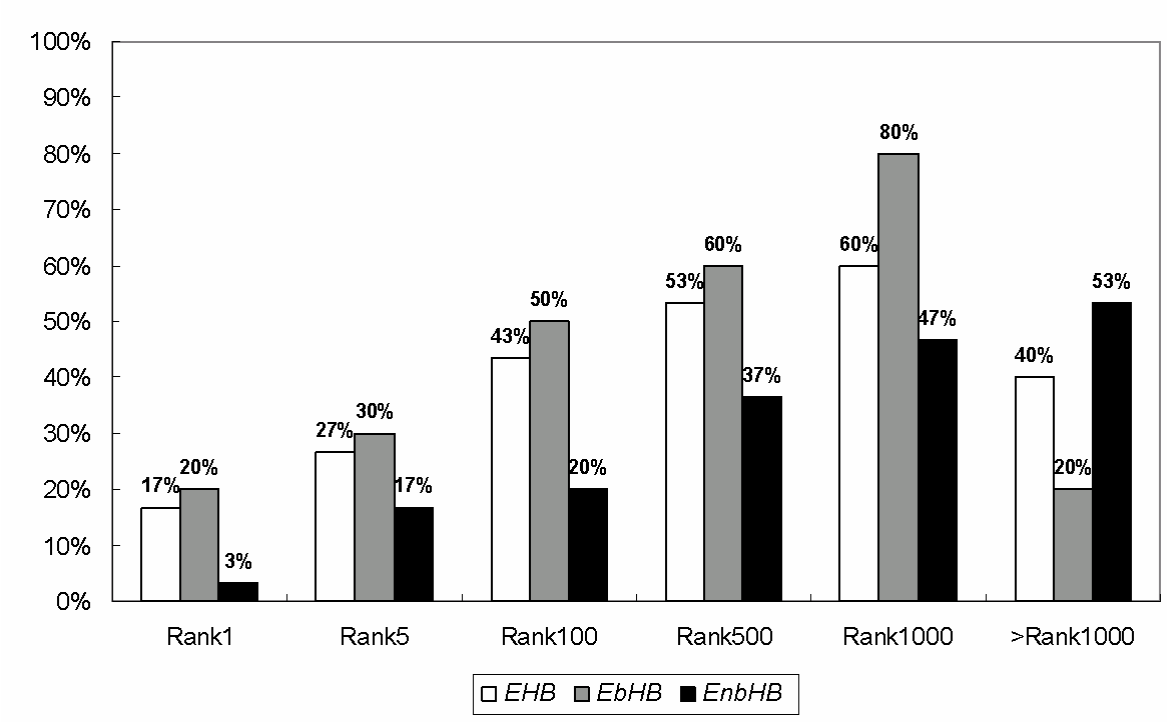


Figure 12. Comparison of different hydrogen-bonding potentials, E_{HB}, E_{bHB}, and E_{nbHB} on the training set. E_{HB} means the potential of all hydrogen-bonding interactions. E_{bHB}, and E_{nbHB} mean the potentials of all hydrogen-bonding interactions on backbone and hydrogen-bonding interactions which do not locate on backbone, respectively (i.e., $E_{HB}=E_{bHB}+E_{nbHB}$). Rank1, Rank5, Rank100, Rank500, Rank1000 and Other mean the percentages of native structures ranked within 1, 5, 100, 500, 1000 and larger 1000 in the training set with different hydrogen-bonding potentials, E_{HB}, E_{bHB}, and E_{nbHB}.

References

1. Doolittle, R.F., *Similar amino acid sequences: chance or common ancestry?* Science, 1981. **214**(4517): p. 149-59.

2. Greer, J., *Comparative modeling methods: application to the family of the mammalian serine proteases.* Proteins, 1990. **7**(4): p. 317-34.

3. Sander, C. and R. Schneider, *Database of homology-derived protein structures and the structural meaning of sequence alignment.* Proteins, 1991. **9**(1): p. 56-68.

4. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-70.

5. Jones, D.T., W.R. Taylor, and J.M. Thornton, *A new approach to protein fold recognition.* Nature, 1992. **358**(6381): p. 86-9.

6. Pedersen, J.T. and J. Moult, *Ab initio protein folding simulations with genetic algorithms: simulations on the complete sequence of small proteins.* Proteins, 1997. **Suppl 1**: p. 179-84.

7. Lomize, A.L., I.D. Pogozheva, and H.I. Mosberg, *Prediction of protein structure: the problem of fold multiplicity.* Proteins, 1999. **Suppl 3**: p. 199-203.

8. Samudrala, R., et al., *Ab initio protein structure prediction using a combined hierarchical approach.* Proteins, 1999. **Suppl 3**: p. 194-8.

9. Lee, J., et al., *Calculation of protein conformation by global optimization of a potential energy function.* Proteins, 1999. **Suppl 3**: p. 204-8.

10. Ortiz, A.R., et al., *Ab initio folding of proteins using restraints derived from evolutionary information.* Proteins, 1999. **Suppl 3**: p. 177-85.

11. Standley, D.M., et al., *Protein structure prediction using a combination of sequence-based alignment, constrained energy minimization, and structural alignment.* Proteins, 2001. **Suppl 5**: p. 133-9.

12. Bradley, P., et al., *Rosetta predictions in CASP5: successes, failures, and prospects for complete automation.* Proteins, 2003. **53 Suppl 6**: p. 457-68.

13. Jones, D.T. and L.J. McGuffin, *Assembling novel protein folds from super-secondary structural fragments.* Proteins, 2003. **53 Suppl 6**: p. 480-5.

14. Fang, Q. and D. Shortle, *Prediction of protein structure by emphasizing local side-chain/backbone interactions in ensembles of turn fragments.* Proteins, 2003. **53 Suppl 6**: p. 486-90.

15. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96): p. 223-30.

16. Russ, W.P. and R. Ranganathan, *Knowledge-based potential functions in protein design.* Curr Opin Struct Biol, 2002. **12**(4): p. 447-52.

17. Sippl, M.J., *Knowledge-based potentials for proteins.* Curr Opin Struct Biol, 1995. **5**(2): p. 229-35.

18. Moult, J., *Comparison of database potentials and molecular mechanics force fields.* Curr Opin Struct Biol, 1997. **7**(2): p. 194-9.

19. Hao, M.H. and H.A. Scheraga, *Designing potential energy functions for protein folding.* Curr Opin Struct Biol, 1999. **9**(2): p. 184-8.

20. Vajda, S., M. Sippl, and J. Novotny, *Empirical potentials and functions for protein folding and binding.* Curr Opin Struct Biol, 1997. **7**(2): p. 222-8.

21. Sippl, M.J., *Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.* J Mol Biol, 1990. **213**(4): p. 859-83.

22. Park, B. and M. Levitt, *Energy functions that discriminate X-ray and near native folds from well-constructed decoys.* J Mol Biol, 1996. **258**(2): p. 367-92.

23. Park, B.H., E.S. Huang, and M. Levitt, *Factors affecting the ability of energy functions to discriminate correct from incorrect folds.* J Mol Biol, 1997. **266**(4): p. 831-46.

24. Tobi, D., et al., *On the design and analysis of protein folding potentials.* Proteins, 2000. **40**(1): p. 71-85.

25. Dominy, B.N. and C.L. Brooks, *Identifying native-like protein structures using physics-based potentials.* J Comput Chem, 2002. **23**(1): p. 147-60.

26. Felts, A.K., et al., *Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model.* Proteins, 2002. **48**(2): p. 404-22.

27. Zhou, H. and Y. Zhou, *Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction.* Protein Sci, 2002. **11**(11): p. 2714-26.

28. Zhu, J., et al., *How well can we predict native contacts in proteins based on decoy structures and their energies?* Proteins, 2003. **52**(4): p. 598-608.

29. Fujitsuka, Y., et al., *Optimizing physical energy functions for protein folding.* Proteins, 2004. **54**(1): p. 88-103.

30. Lee, M.C. and Y. Duan, *Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model.* Proteins, 2004. **55**(3): p. 620-34.

31. Hsieh, M.J. and R. Luo, *Physical scoring function based on AMBER force field and Poisson-Boltzmann implicit solvent for protein structure prediction.* Proteins, 2004. **56**(3): p. 475-86.

32. Zhang, C., et al., *An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state.* Protein Sci, 2004. **13**(2): p. 400-11.

33. Price, D.J. and C.L. Brooks, 3rd, *Modern protein force fields behave comparably in molecular dynamics simulations.* J Comput Chem, 2002. **23**(11): p. 1045-57.

34. Novotny, J., R. Bruccoleri, and M. Karplus, *An analysis of incorrectly folded protein models. Implications for structure predictions.* J Mol Biol, 1984. **177**(4): p. 787-818.

35. Samudrala, R. and M. Levitt, *Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction.* Protein Sci, 2000. **9**(7): p. 1399-401.

36. Simons, K.T., et al., *Ab initio protein structure prediction of CASP III targets using ROSETTA.* Proteins, 1999. **Suppl 3**: p. 171-6.

37. Kihara, D., et al., *TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints.* Proc Natl Acad Sci U S A, 2001. **98**(18): p. 10125-30.

38. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*. 1989, MA.USA: Addison-Wesley Publishing Company Inc. Reading.

39. Back, T., *Evolutionary Algorithms in Theory and Practice*. 1996, New York: Oxford University. 3-17.

40. Fogel, D.B., *Evolutionary Computation: Toward a New Philosophy of Machine Intelligent*. 1995, New York: IEEE Press.

41. Muhlenbein, H. and D. Schlierkamp-Voosen, *Predictive models for the breeder genetic algorithm: I. Continuous parameter optimization.* Evolutionary Compuation, 1993. **1**(1): p. 25-49.

42. Morris, G.M., et al., *Automated docking using a lamarckian genetic algorithm and empirical binding free energy function.* J Comp Chem, 1998. **19**: p. 1639-1662.

43. Gehlhaar, D.K., et al., *Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming.* Chem Biol, 1995. **2**(5): p. 317-24.

44. Hart, W.E., *Comparing evolutionary programs and evolutionary pattern search algorithms: A drug docking application.* Proc. Genetic and Evolutionary Computation Conf., 1999: p. 855-862.

45. Yang, J.M., C.Y. Kao, and J.T. Horng, *A continuous genetic algorithm for global optimization.* Proc. of the Seventh Int. Conf. on Genetic Algorithms, 1997: p.

230-237.

46. Yang, J.M. and C.Y. Kao, *A family competition evolutionary algorithm for automated docking of flexible ligands to proteins.* IEEE Trans Inf Technol Biomed, 2000. **4**(3): p. 225-37.

47. Yang, J.M., J.T. Horng, and C.Y. Kao, *A genetic algorithm with adaptive mutations and family competition for training neural networks.* International Journal of Neural Systems, 2000. **10**: p. 333-352.

48. Yang, J.M. and C.C. Chen, *GEMDOCK: a generic evolutionary method for molecular docking.* Proteins, 2004. **55**(2): p. 288-304.

49. Yang, J.M., *Development and evaluation of a generic evolutionary method for protein-ligand docking.* J Comput Chem, 2004. **25**(6): p. 843-57.

50. Eisenberg, D. and A.D. McLachlan, *Solvation energy in protein folding and binding.* Nature, 1986. **319**(6050): p. 199-203.

51. Wesson, L. and D. Eisenberg, *Atomic solvation parameters applied to molecular dynamics of proteins in solution.* Protein Sci, 1992. **1**(2): p. 227-35.

52. Still, W.C., et al., *Semianalytical treatment of solvation for molecular mechanics and dynamics.* J Am Chem Soc, 1990. **112**: p. 6127-6129.

53. Yang, J.-M., et al., *An evolutionary approach with pharmacophore-based scoring functions for virtual database screening.* Lecture Notes in Computer Science, 2004. **3102**: p. 481-492.

54. Yang, J.M. and T.W. Shen, *A pharmacophore-based evolutionary approach for screening selective estrogen receptor modulators.* Proteins, 2005. **59**(2): p. 205-20.

55. Baker, E.N. and R.E. Hubbard, *Hydrogen bonding in globular proteins.* Prog Biophys Mol Biol, 1984. **44**(2): p. 97-179.

56. Koretke, K.K., Z. Luthey-Schulten, and P.G. Wolynes, *Self-consistently optimized energy functions for protein structure prediction by molecular dynamics.* Proc Natl Acad Sci U S A, 1998. **95**(6): p. 2932-7.

57. Mirny, L.A. and E.I. Shakhnovich, *How to derive a protein folding potential? A new approach to an old problem.* J Mol Biol, 1996. **264**(5): p. 1164-79.

58. Klepeis, J.L. and C.A. Floudas, *Cmparative study of global minimum energy conformations of hydrated peptides.* J Comp Chem, 1999. **20**(6): p. 636-654.

59. Tsai, J., et al., *An improved protein decoy set for testing energy functions for protein structure prediction.* Proteins, 2003. **53**(1): p. 76-87.

60. Simons, K.T., et al., *Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins.*

Proteins, 1999. **34**(1): p. 82-95.

61. Simons, K.T., et al., *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* J Mol Biol, 1997. **268**(1): p. 209-25.

62. Bonneau, R., et al., *Rosetta in CASP4: progress in ab initio protein structure prediction.* Proteins, 2001. **Suppl 5**: p. 119-26.

63. Holm, L. and C. Sander, *Evaluation of protein models by atomic solvation preference.* J Mol Biol, 1992. **225**(1): p. 93-105.

64. Samudrala, R., et al., *A combined approach for ab initio construction of low resolution protein tertiary structures from sequence.* Pac Symp Biocomput, 1999: p. 505-16.

65. Ferrara, P., J. Apostolakis, and A. Caflisch, *Evaluation of a fast implicit solvent model for molecular dynamics simulations.* Proteins, 2002. **46**(1): p. 24-33.

66. Levitt, M. and S. Lifson, *Refinement of protein conformations using a macromolecular energy minimization procedure.* J Mol Biol, 1969. **46**(2): p. 269-79.

67. Levitt, M., *Energy refinement of hen egg-white lysozyme.* J Mol Biol, 1974. **82**(3): p. 393-420.

68. Levitt, M., *Molecular dynamics of native protein. II. Analysis and nature of motion.* J Mol Biol, 1983. **168**(3): p. 621-57.

69. Levitt, M., et al., *Potential energy function and parameters for simulations of the molecular dynamicsof proteins and nucleic acids in solution.* Comput Phys Commun, 1995. **91**: p. 215-231.

70. Samudrala, R. and J. Moult, *An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction.* J Mol Biol, 1998. **275**(5): p. 895-916.