Soft Energy Function and Generic Evolutionary Method for Discriminating Native from Nonnative Protein Conformations

YI-YUAN CHIU,[1] JENN-KANG HWANG,[1–3] JINN-MOON YANG[1–3]
[1]*Institute of Bioinformatics, National Chiao Tung University, Hsinchu 30050, Taiwan*
[2]*Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 30050, Taiwan*
[3]*Core Facility for Structural Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan*

Received 28 June 2007; Revised 25 October 2007; Accepted 8 November 2007
DOI 10.1002/jcc.20897
Published online 7 January 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: We have developed a soft energy function, termed GEMSCORE, for the protein structure prediction, which is one of emergent issues in the computational biology. The GEMSORE consists of the van der Waals, the hydrogen-bonding potential and the solvent potential with 12 parameters which are optimized by using a generic evolutionary method. The GEMSCORE is able to successfully identify 86 native proteins among 96 target proteins on six decoy sets from more 70,000 near-native structures. For these six benchmark datasets, the predictive performance of the GEMSCORE, based on native structure ranking and *Z*-scores, was superior to eight other energy functions. Our method is based solely on a simple and linear function and thus is considerably faster than other methods that rely on the additional complex calculations. In addition, the GEMSCORE recognized 17 and 2 native structures as the first and the second rank, respectively, among 21 targets in CASP6 (Critical Assessment of Techniques for Protein Structure Prediction). These results suggest that the GEMSCORE is fast and performs well to discriminate between native and nonnative structures from thousands of protein structure candidates. We believe that GEMSCORE is robust and should be a useful energy function for the protein structure prediction.

 J Comput Chem 29: 1364–1373, 2008

Key words: energy function; protein structure prediction; structural bioinformatics; evolutionary computation

Introduction

The protein structure prediction (PSP) remains one of the fundamental unsolved problems in the field of computational biology.[1] A computational method for PSP involves two basic critical elements: an efficient method to search a large number of potential structure candidates and a reliable energy function.[1,2] The search methods can be roughly divided into two categories: template-based approaches (i.e. comparative modeling[3,4] and fold recognition[5,6]) and template-free approaches (e.g. *ab initio*[7,8]). A good energy function for PSP should screen a large number of potential solutions rapidly and simply while effectively discriminating native or near-native structures from thousands of protein structure candidates.[9,10] In general, the binding energy landscapes of these scoring functions are often complex and exhibit a rugged funnel shape. Therefore, an efficient search algorithm is required to find a global solution for various scoring functions.

Energy functions are generally rooted in the thermodynamic hypothesis-the native-state conformation, which occupies the lowest energetic state,[11] is the most stable among states. For a PSP method, an energy function, which can accurately depict the energy landscape of a protein conformation space, is an essential requirement to distinguish the native structures from lots of candidate conformations. Toward the aim of developing such an energy function, various energy functions have been developed for calculating the free energy, including knowledge-based,[10,12] empirical-based,[13,14] physics-based,[15–20] and solvent potentials.[21–23]

Knowledge-based energy functions[10,12] are generally derived from distributions of experiment structural data available from Protein Data Bank (PDB[24]). Reduced representation of protein structures was usually used in knowledge-based energy functions

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/0192-8651/suppmat

Correspondence to: J.-M. Yang; e-mail: moon@faculty.nctu.edu.tw

Contract/grant sponsor: National Science Council
Contract/grant sponsor: MOE

Table 1. The 12 Energy Terms Used in the GEMSCORE.

Energy name	Parameter name	Weight	Description
E_{elect}	w_1	3.83	Electrostatic energy
E_{vdW}	w_2	1.00	van der Waals potential
E_{bHB}	w_3	3.51	Hydrogen-bonding potential on backbone
$E_{SAS\text{-}bC}$	σ_1	0.94	Surface area of all C atoms on backbone
$E_{SAS\text{-}sC}$	σ_2	0.39	Surface area of all C atoms on sidechain
$E_{SAS\text{-}sS}$	σ_3	0.68	Surface area of all S atoms on sidechain
$E_{SAS\text{-}bO}$	σ_4	−0.54	Surface area of all O atoms on backbone
$E_{SAS\text{-}sO}$	σ_5	0.27	Surface area of all O atoms on sidechain
$E_{SAS\text{-}nO}$	σ_6	−0.58	Surface area of all negative charged O atoms on sidechain in residues Asp and Glu
$E_{SAS\text{-}bN}$	σ_7	−1.5	Surface area of all N atom on backbone
$E_{SAS\text{-}sN}$	σ_8	−0.34	Surface area of all N atom on sidechain
$E_{SAS\text{-}pN}$	σ_9	−0.55	Surface area of all positive charged N atoms on sidechain in residue His, Arg, and Lys

for simplification and reducing the computational time. To suit for the reduced representation, these energy functions may contain pseudo-potentials which are often lack of the physical meanings. Physics-based energy functions are based on physical mechanisms. They are often derived from *ab initio* quantum-mechanical calculations according to the principles of physics. One advantage of physics-based energy functions is the lucid physical meaning of each individual term. Despite their perceived advantages, physics-based energy functions have not been widely adopted mostly due to the high-computation cost.[15–20] In addition, to develop a physics-based energy function is often complicate to optimize many potential parameters of each energy terms influencing the performance of energy functions.

In this work, we developed a new energy function (GEMSCORE) that was modified from the energy function of GEMDOCK[25,26] and added solvent potentials.[21,22] The energy function, using the piecewise linear potential to soften the repulsive term of Lennard-Jones potential, of GEMDOCK has a good performance in flexible protein-ligand docking and drug screening.[27,28] The short range repulsive interactions (e.g. Lennard-Jones potential) tend to infinity at low interatomic separation leading to rough energy surfaces with high energy barriers. A soft scoring function has been applied for softening the repulsive intermolecular potential to decrease the strong sensitivity of interaction energies to local conformation changes.[29–31] Generally, a soft scoring function has the benefit of being computationally efficient, conversely, it may increase the number of false near-native solutions (structures). The tradeoff of its advantages and limitations can be optimized.

The GEMSCORE has simplified energy terms based on physical mechanisms, including electrostatic, the van der Waals, the hydrogen-bonding potential, and the solvent potential. To develop a simple and fast soft energy function for PSP, we adopted a reduced optimization scheme to reflect the contributions of the 12 energy terms, which are used in the GEMSCORE, for the near-native structures. A modified generic evolutionary method (GEM), which was successfully applied on some specific domains,[25,32–34] was adopted to optimize these term weights of the GEMSCORE.

Results and Discussion

Data Sets

A widely used approach to test energy functions for the PSP is to partially sample the conformational spaces utilizing constructed decoy sets. To optimize parameters of the GEMSCORE for discriminating the native structures and nonnative structures, we selected a decoy set which consists of 30 protein targets proposed by Tsai et al.[35] (Tables S1 and S2 in support material), as the training set. Each target consists of 1867 decoy structures which were based on Rosetta protocol and modified from Rosetta all-atom decoy set[36] and increasing the number of structures of near-native structures. We filtered out the targets with incomplete residues and 30 targets were included in the training set.

After the parameters optimization (Table 1), six widely used decoy sets are applied to evaluate the GENSCORE performance and to compare with other methods. These decoy sets includes 25 targets in the EMBL misfolded set,[37] seven targets in the 4state_reduced set (4state),[38] 10 targets in local-minima decoy set (lmds),[39] eight targets in the lattice_ssfit set (lattice),[40] four targets in the fisa decoy set,[41] and 42 targets in the Rosetta all-atom decoy set (RosettaAll).[36] We also applied the GEMSCORE on 21 CASP6 targets (Critical Assessment of Techniques for Protein Structure Prediction Six) which were directly obtained from CASP6 website. The training set and testing sets are collected and summarized in Table S1 (in support material).

Energy Terms in GEMSCORE

The weights and descriptions of the energy terms used in the GEMSCORE are listed in Table 1. The GEMSCORE was enhanced and modified from the soft scoring function of our previous works[25,26] for PSP by adding the solvation potential, which plays an important role in protein folding, and enhancing the hydrogen-bonding potential. The GEMSCORE is given as

$$E_{Total} = w_1 E_{elect} + w_2 E_{vdW} + w_3 E_{bHB} + E_{SAS} \quad (1)$$

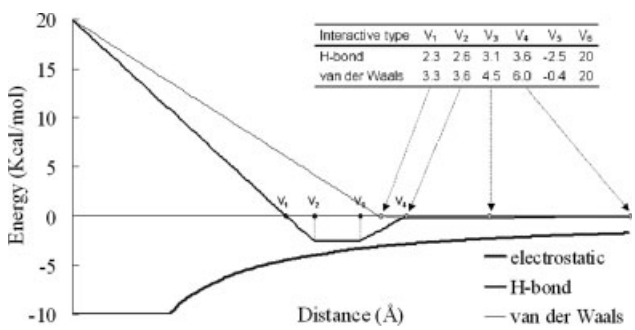

Figure 1. The linear energy functions of the pairwise atoms for the van der Waals interaction, hydrogen bond, and electrostatic potential in the GEMSCORE.

Table 2. The GEMSCORE Results with Different Combinations of Energy Terms.

	GEMSCORE			GEMSCORE without E_{SAS}			GEMSCORE without E_{bHB}			GEMSCORE without E_{bHB} and E_{SAS}		
	Z[a]	Z'[b]	Rank	Z	Z'	Rank	Z	Z'	Rank	Z	Z'	Rank
EMBL	N/A	N/A	25/25[c]	N/A	N/A	18/25	N/A	N/A	25/25	N/A	N/A	20/25
4state	−3.72	−0.93	6/7	−2.09	0.83	3/7	−3.21	−0.33	4/7	−1.43	1.48	0/7
fisa	−2.92	−0.32	3/4	−3.22	−0.33	2/4	−3.78	−1.31	3/4	−4.39	−1.59	2/4
lmds	−2.23	0.54	7/10	−4.37	−1.37	8/10	−2.56	0.20	7/10	−5.40	−2.34	8/10
lattice	−5.24	−2.15	8/8	−3.25	−0.84	7/8	−3.90	−1.40	8/8	−2.54	−0.50	6/8
RosettaAll	−5.19	−2.39	37/42	−4.69	−2.34	35/42	−4.96	−2.45	41/42	−4.81	−2.53	39/42

[a,b]The average Z score defined in the eqs. (5) and (6), respectively.
[c]The first number is the number of native structures ranking in the first rank; the second number is total number of target proteins in the decoy set.

where E_{elect} is the electrostatic energy, E_{vdW} is van der Waals potential, E_{bHB} is the hydrogen-bonding potential on the protein backbone, and E_{SAS} is the solvent potential. E_{vdW} and E_{bHB} are a simplified atomic pair-wise potential function (see Fig. 1). The values of the hydrogen-bonding potential on backbones should be larger than the ones of the van der Waals potential. The

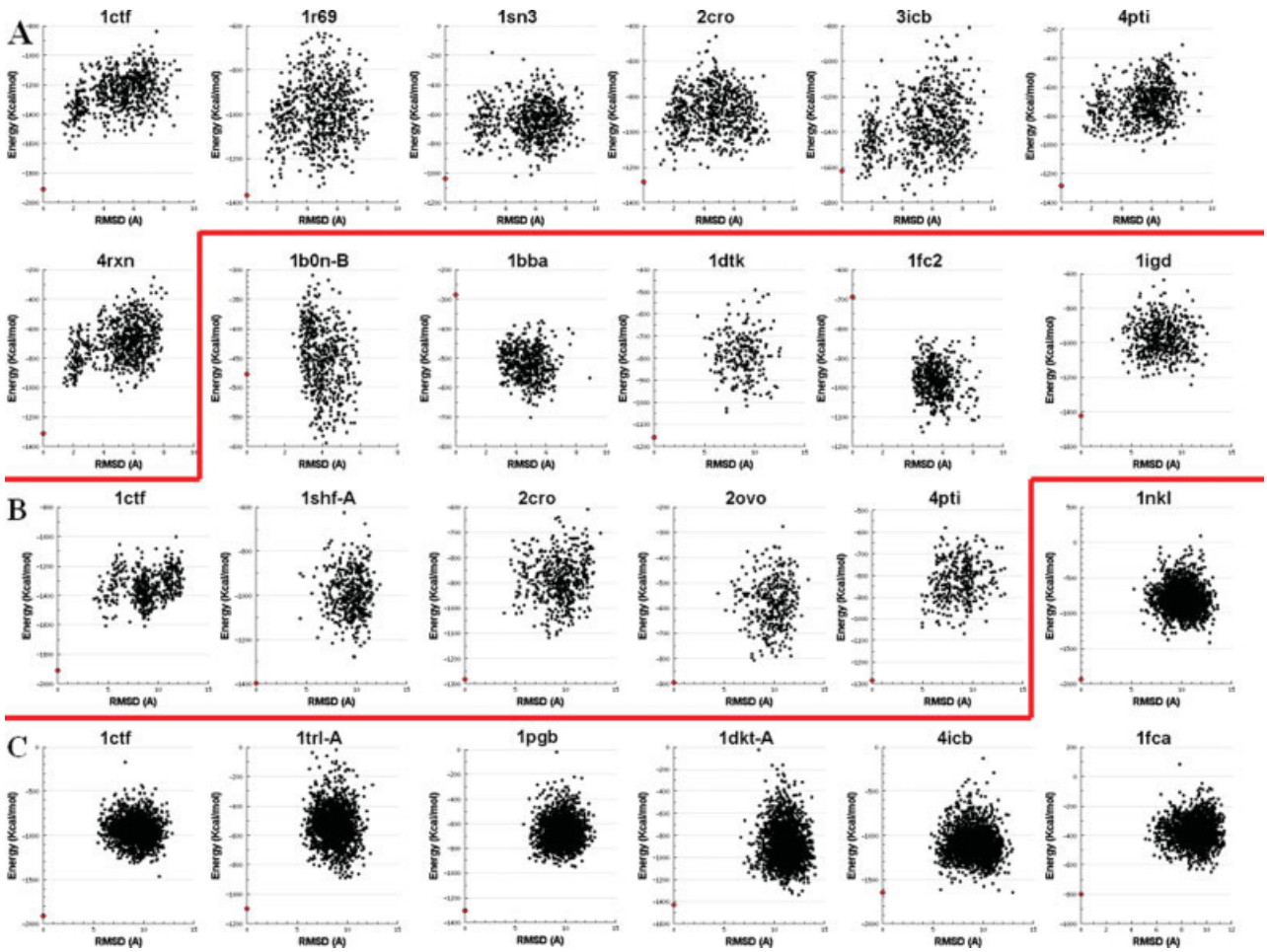

Figure 2. The correlations between the GEMSCORE potentials and RMSD values of 24 target proteins on three data sets: (A) 4state, (B) lmds, and (C) lattice sets. The RMSD of a native structure is zero and it is indicated as a red dot.

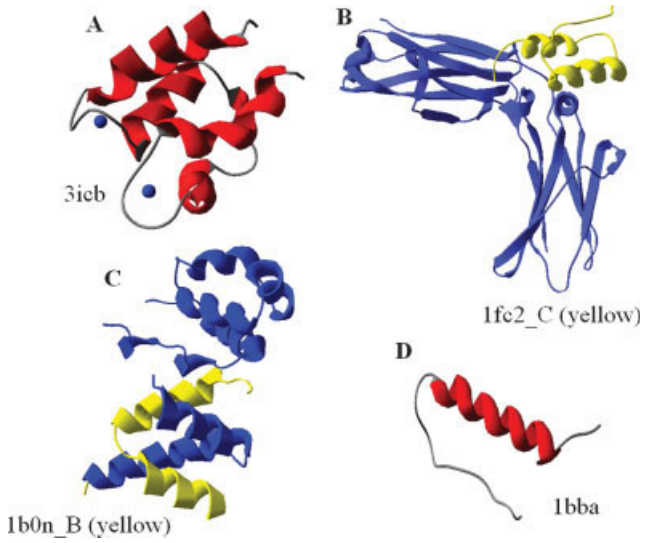

Figure 3. The GEMSCORE for four bad target proteins: (A) 3icb, (B) 1fc2_C, (C) 1b0n_B, and (D) 1bba. The protein 3icb contains two calcium ions; the protein 1fc2_C or the protein 1b0n_B is a part (i.e. a chain or a domain) of a protein; and the protein 1bba is an NMR structure.

GEMSCORE considers the E_{vdW} when the distance of a pair atom is less than 6 Å. As shown in Table 1, the weight values of E_{elect} (3.83) and E_{bHB} (3.51) are much larger than E_{vdW} (1.0). In general, the total E_{vdW} value is typical about 10–15 times of the total E_{bHB} value in the GEMSCORE and the ratio is reduced to ~3 if the weight values are considered.

The GEMSCORE uses the atomic solvation parameter (ASP) proposed by Wesson and Eisenberg[21,22] to approximate to the solvation energy. The ASP is defined as

$$E_{SAS} = \sum_{i=1}^{N} \sigma_i A_i, \quad (2)$$

where σ_i is the ASP for atom type i, and A_i the solvent-accessible surface area of the atom type i. A probe radius of 1.4 Å is used to calculate an atomic solvent-accessible surface area. Wesson and Eisenberg[22] classified atoms into five types including C, uncharged O or N, S, O^-, and N^+. According to an atom on the sidechain or the backbone, we further divided atoms into nine atom types: atom C on backbone or sidechain, S, atom O on backbone or sidechain, O^-, atom N on backbone or side-chain, and N^+ (Table 1).

The GEM method was applied to optimize the parameters of these 12 energy terms in the GEMSCORE based on 30 target proteins in the training set (Table S2 in support material). We set the GEM parameters, including the initial step sizes (size σ = 0.8 and ψ = 0.2.), the family competition length (L = 2), the population size (N = 200), and the recombination probability (p_c = 0.3) according to the experiments of various parameters (see Methods). The GEM optimization stops when either the convergence below certain threshold value or the iterations exceed a maximal preset value, which was set to 200. Therefore,

Table 3. Comparisons the GEMSCORE with Other Works on Five Testing Data Sets.

Decoy set	GEMSCORE	Fujitsuka et al.[18]	Krishnamoorthy[44]	Hu[45]		Zhang[43]		Zhu et al.[17]	Zhou[42]	Lee and Duan[19]	Bhattacharyay et al.[23]
				KFF	KDF	DFIRE-SCM	DFIRE-allatom				
EMBL misfold	25/25[a] (N/A[b])	N/A[c]	N/A	N/A	N/A	N/A	N/A	N/A	25/25 (N/A)	24/25 (N/A)	N/A
4state	6/7 (−3.58[d])	5/7 (−2.36)	3/7 (2.61)	6/7 (3.19)	5/7 (2.69)	6/7 (3.94)	6/7 (3.49)	N/A (−3.96)	6/7 (3.49)	N/A (−4.95)	4/7 (−2.89)
lmds	7/10 (−2.95)	4/6 (−3.85)	2/10 (1.09)	4/10 (1.29)	3/10 (1.14)	3/10 (2.56)	7/10 (0.90)	N/A (−1.75)	7/10 (0.67)	N/A (−4.49)	4/8 (−2.75)
lattice	8/8 (−6.54)	N/A	5/8 (3.54)	4/8 (3.01)	5/8 (3.76)	8/8 (6.19)	8/8 (9.47)	N/A (−4.08)	8/8 (8.94)	N/A (−6.75)	6/8 (−4.06)
fisa	3/4 (−2.92)	N/A	N/A	N/A	N/A	3/4 (4.70)	3/4 (4.80)	N/A (−3.09)	3/4 (4.49)	N/A(−2.09)	N/A

[a]The first number is the number of native structures ranking in the first rank; the second number is total number of target proteins in the decoy set.
[b]The average Z-score is not available because of only one decoy structure in the EMBL misfolded decoy set.
[c]The data is not available in the original paper.
[d]The value in parentheses is the average Z score of a decoy set.

Table 4. Comparisons the GSMSCORE with Related Works on the 4state, lmds, and Lattice Sets.

				Hu[45]					
PDB ID	GEMSCORE	Fujitsuka et al.[18]	Krishnamoorthy[44]	KFF	KDF	Zhu[17]	Zhou[42]	Lee[19]	Bhattacharyay[23]
4state									
1ctf	1/−5.41[a]	1/–2.50	7/2.62	1/3.64	1/3.14	–/–3.33	1/3.86	–/–4.26	1/–3.53
1r69	1/−3.14	1/–2.50	3/2.90	1/3.77	1/3.79	–/–3.63	1/4.23	–/–5.35	1/–3.68
1sn3	1/−2.95	1/–3.20	113/1.04	1/2.15	27/1.79	–/–5.70	1/3.79	–/–6.33	1/–2.52
2cro	1/−3.42	1/–2.30	1/3.04	3/2.57	1/2.66	–/–3.55	1/3.29	–/–5.11	2/–3.01
3icb	39/–1.40 1/–3.21[b]	3/−1.60	1/2.90	1/2.56	1/2.68	–/–1.97	4/2.28	–/–2.86	1/–2.26
4pti	1/−4.24	1/−2.70	1/3.18	1/4.17	1/2.79	–/–5.09	1/3.62	–/–5.35	5/–2.51
4rxn	1/−4.56	12/−1.70	5/–2.58	1/3.45	1/1.99	–/–4.43	1/3.33	–/–5.36	4/–2.71
lmds									
1b0n-B	110/−0.82	N/A	28/1.48	406/–0.94	19/2.05	–/–2.55	430/–1.17	–/–0.61	439/1.18
1bba	498/2.91	N/A	488/–1.93	500/–3.58	487/–1.83	N/A	501/–16.3	–/4.99	N/A
1ctf	1/−6.25	1/−4.70	205/0.20	1/3.62	1/3.31	–/–4.38	1/3.54	–/–5.12	1/–3.42
1dtk	1/−2.61	2/−2.30	1/2.63	59/0.64	185/–1.11	–/–3.51	1/2.62	–/–6.10	N/A
1fc2	494/2.16	N/A	4/2.06	501/–3.08	486/–1.87	–/–0.22	501/–5.72	–/–3.38	409/0.91
1igd	1/−5.89	1/−6.20	372/–0.71	1/5.18	1/3.93	–/–5.80	1/5.16	–/–6.16	1/–2.87
1shf-A	1/−5.22	N/A	189/0.32	5/2.14	12/1.82	–/–7.53	1/6.68	–/–8.26	1/–2.90
2cro	1/−5.49	1/−4.00	1/3.88	2/2.65	1/3.24	–/–5.97	1/4.70	–/–8.03	1/–3.42
2ovo	1/−3.20	1/−4.10	46/0.99	1/3.11	38/1.21	–/–4.51	1/3.21	–/–6.00	16/–1.67
4pti	1/−5.08	17/−1.80	7/1.98	1/3.14	108/0.62	–/–7.04	1/3.96	–/–6.24	6/–2.24
lattice									
1beo	1/−8.44	N/A	1/5.35	15/2.45	1/3.94	–/–4.86	1/12.09	–/–7.95	1/–3.67
1ctf	1/−8.38	N/A	1/4.18	1/3.76	1/5.35	–/–3.22	1/10.05	–/–6.98	1/–5.04
1dkt-A	1/−3.31	N/A	89/1.67	17/2.42	8/2.64	–/–5.89	1/6.87	–/–6.40	8/–2.73
1fca	1/−5.61	N/A	1/4.91	56/2.00	98/1.76	–/–5.89	1/7.18	–/–8.30	1/–7.38
1nkl	1/−8.35	N/A	1/4.38	1/3.60	1/3.51	–/–3.97	1/9.29	–/–2.60	1/–4.54
1pgb	1/−8.94	N/A	14/2.58	1/3.95	1/4.91	–/–2.66	1/11.87	–/–9.55	1/–4.01
1trl-A	1/−3.91	N/A	1179/−0.23	56/1.97	18/2.67	–/–4.27	1/12.09	–/–5.49	101/–1.61
4icb	1/−5.41	N/A	1/5.47	1/3.92	1/5.31	–/–1.85	1/10.05	N/A	1/–3.50

[a]The first number is the number of native structures ranking in the first rank; the second number is total number of target proteins in the decoy set. The larger of the absolute value of a *Z*–score, the better of this method is.
[b]The result of adding two calcium ions into the Protein 3icb.

GEM generated 800 solutions in one generation and terminated after it exhausted 160,000 solutions in the worse case. These parameters were decided after experiments conducted to recognize complexes of test decoy systems with various values.

ASPs of nine atom types are shown in Table 1. The parameteric values of an atom on the backbone and sidechain are different for all atom types, including N, O, and C. The parameter values of N (−1.5) and C (0.94) on the backbone are the most negative and the most positive, respectively. The values of two charged atoms, N^+ (−0.55) and O^- (−0.58), are similar. In addition, the atom accessible surface areas were calculated from 30 native structures in the training set. The accessible surface areas, on average, of N^+ (in Arg and Lys) and O^- (in Asp and Glu) are ∼55 and ∼36 Å^2, respectively. The ASPs and accessible surface areas are similar to the results proposed by Wesson and Eisenberg.[22]

GEMSCORE Results on Six Benchmarks

Table 2 and Figure 2 show the results of the GEMSCORE on 96 targets in six test decoy sets (Table S1 in support material). The GEMSCORE successfully identified 86 native proteins as the first rank among these 96 proteins. The GEMSCORE without E_{SAS} yielded the worst performance (73 native proteins as the first rank) and the GEMSCORE without E_{bHB} obtained the best performance (89 native proteins as the first rank). The original soft scoring function of the GEMDOCK gets 75 native proteins as the first rank. The GEMSCORE using all energy terms yielded the best performance for the 4state data set among the methods with different combinations of energy terms. The performance of the GEMSCORE was decreased for the EMBL, 4state, and RosettaAll data sets if the GEMSCORE discarded atomic solvation potentials (Table 2). We found that the GEMSCORE using nine atom types for atomic solvation potentials outperformed it using five atom types.[22] For the 4state data set, the original soft scoring function of the GEMDOCK is unable to recognize the native structure as the first rank but the GEMSCORE with E_{SAS} and with E_{bHB} yields 3 and 4, respectively, native structures as the first rank. Figure 2 shows that many structures in the 4state data set are near the native structures (<4 Å), in contrast, only few structures in both lmds and lattice date sets are less than 4 Å. These results imply that the hydro-

Table 5. The GEMSCORE Results on 21 Targets in CASP6.

Target Id	PDB code	No. of predicted structures	No. of residues	RMSD			Z-score of the native structure	Rank[a]
				Min[b]	Max[c]	Average[d]		
T0240	1u07	44	90	5.38	23.68	15.87	−2.12	1
T0266	1wdv	64	152	1.60	16.89	2.58	−1.84	1
T0271	1vgg	34	161	2.65	79.31	10.51	−1.17	1
T0274	1wgb	47	159	3.15	29.71	4.75	−1.88	1
T0275	1wjg	74	137	2.59	16.74	4.67	−1.59	2
T0277	1wj8	71	119	1.53	61.88	4.89	−1.82	2
T0282	1xfk	35	332	4.31	22.00	8.54	−2.21	1
T0200	1t70	49	255	6.98	21.59	13.67	−2.86	1
T0267	1wk4	40	175	2.80	18.01	5.71	−1.62	1
T0263	1wd6	52	101	3.47	39.27	8.61	−2.01	1
T0212	1tza	43	126	5.07	23.66	14.3	−3.51	1
T0239	1rki	47	98	10.68	23.35	14.16	−3.73	1
T0281	1whz	57	70	1.59	27.79	9.98	−1.77	4
T0201	1s12	47	94	4.91	18.84	10.92	−1.86	1
T0242	2blk	46	116	11.55	18.73	14.99	−2.91	1
T0273	1wdj	32	187	11.87	57.65	18.31	−4.05	1
T0211	1xpw	47	144	3.89	28.04	8.53	−2.16	1
T0213	1te7	54	103	4.84	19.42	10.11	−1.16	6
T0214	1s04	48	110	2.11	55.78	13.83	−2.67	1
T0224	1rhx	56	87	3.85	18.73	6.92	−8.93	1
T0230	1wcj	52	104	7.94	24.14	12.4	−2.35	1

[a]The rank of the native structure.
[d]Minimum RMSD, maximum RMSD, and average RMSD between the native structure and predicted structures.

gen-bonding potential of the backbone and atomic solvent potentials are important to discriminate the native structures from the near-native structures.

Hydrogen-bonding interactions are important for the protein folding, protein–protein interactions, and many other biological functions. Here, the GEMSCORE discarded the hydrogen-bonding potentials of sidechain–sidechain and sidechain–backbone interactions according to the following observations: (1) The performance of the GEMSCORE was decreased when we considered all kinds of hydrogen-bonding potentials (Table S3 in supporting material) on the 4state set; (2) The protein structures mainly were decided by the backbone conformations which are often determined by the secondary structures, such as α-helixs and β-strands. The hydrogen bond on the backbone is the essential force for the secondary structures; (3) The sidechain conformation is generally more flexible than the one of on the backbone for protein structures. As shown in Table S3, the GEMSCORE yielded the worst performance when it considered only the hydrogen-bonding potentials of sidechain–sidechain interactions.

Figure 2 shows the correlations between the GEMSCORE potentials and the root mean square deviation (RMSD) between the native structure and decoy structures in three data sets: 4state, lmds, and lattice sets. The native structures are indicated as the red dots and the respective RMSD values are zero. Figure 2 shows that our energy function is able to identify native and near-native structures from lots of decoy structures.

The factors causing the GEMSCORE to misidentify for 10 targets among these 96 test targets can be roughly divided into four categories. In the first category, the protein structure contains metal irons which were removed from the structure, such as the target 3icb (Fig. 3A). Protein 3icb in the 4state set is a vitamin D-dependent calcium-binding protein and contains two calcium ions, which locate in the loop section of native structure. Lack of Ca^{2+} at loop could cause unstable loop and fold into misfolded structure. The GEMSCORE is able to discriminate the native structure from nonnative structures when these two Ca^{2+} atoms were added in the structure. In the second category, the target structure is a part (i.e. a chain or a domain) of a protein structure (complex), such as target 1fc2 in the fisa set and the target 1b0n-B in the lmds set (Figs. 3B and 3C). In the third category, the target structure (e.g. 1b0n-B) misses the coordinates of some residues in crystal structures. In the final category, the target protein is an NMR structure. The GEMSCORE obtained the successful percentages on NMR structures and X-ray structures are 70.4% (19/27 structures) and 94.2% (65/69 structures), respectively. The GEMSCORE identified 37 native structures from 42 native protein targets in the RosettaAll set. All of those misidentified targets are NMR structures. The protein 1bba, which was misidentified by the GEMSCORE in the lmds set, is also an NMR structure (Fig. 3D).

Compare GEMSCORE and Other Approaches

In general, it is neither straightforward nor completely fair to compare the results of different scoring methods given that each employs different accuracy measures, optimization methods, and test complexes. Tables 3 and 4 show the results of the GEM-

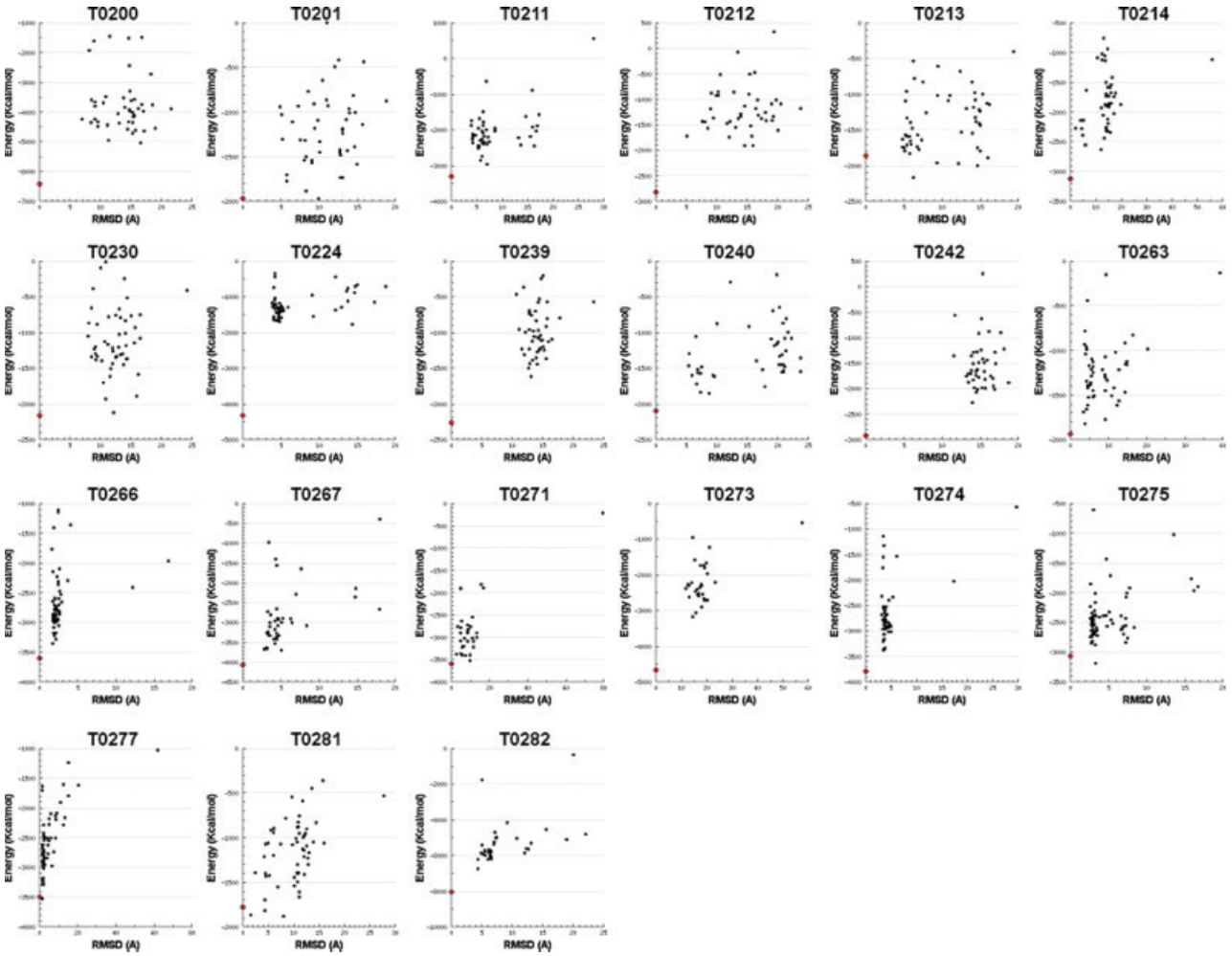

Figure 4. The correlations between the GEMSCORE potentials and RMSD values of 21 targets in the CASP6. The RMSD of a native structure is zero and it is indicated as a dot. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

SCORE and eight comparative methods on five data sets, including the EMBL misfolded set (25 proteins), 4state set (7 proteins), lmds set (10 proteins), lattice set (6 proteins), and fisa set (4 proteins). These eight related works consist of knowledge-based and empirical energy functions,[42–44] solvation potentials,[23] and physics energy functions.[17–19,45] We compared these methods based on two wide-used performance factors which are the ranks and *Z*-scores (see material and method) of the native structures. The GEMSCORE achieved equal performance with the best of eight comparative energy functions and was better than other methods.

Most of these scoring functions are unable to identify the native structures from decoy structures for three targets [e.g. 1bon-B, 1bba, and 1fc2 (see Fig. 3)] which are also misidentified by the GEMSCORE (Table 4). The top best three methods (the GEMSCORE and two related works[42,43]) identify the wrong native structure for the target 3icb in the 4state data set. For this target, the GEMSCORE can identify the native structure if two Ca^{2+} atoms were added into the structure. These results suggest that most of scoring functions are unable to recognize the native structures with following properties: the structure is a part of a protein and the structure missing resides or important atoms (e.g. metal irons).

GEMSCORE Results on CASP6 Targets

Despite the success of the GEMSCORE on the above test sets, the discrimination of the native structure against the decoy sets is the first test towards PSP. For the real-world PSP, the recognition of the near-native structures is the more important property of an energy function for *de novo* structure prediction. Here, the GEMSCORE was evaluated for 21 targets in CASP6 (Critical Assessment of Techniques for Protein Structure Prediction) and it can distinguish the 17 native structures from 21 protein targets (Table 5). Figure 4 shows the correlations between the GEMSCORE potentials and RMSD values between native structure and predicted structures for 21 these targets. For each target, the native structure was obtained from PDB and the predicted struc-

tures were collected from the website http://www.predictioncenter.org/casp6/tar/predictions/. The domain targets and incomplete crystal structures were discarded and 21 targets were selected. On average, ~50 predicted structures were collected for each target and the minimum and maximum RMSD between the predicted structures and the native structure were also provided. The GEMSCORE recognized 17 native structures as the first rank and two native structures as the second rank (Table 5). In the future, we will test the GEMSCORE using a diverse set of real-world PSP targets to systematically improve prediction accuracy.

Conclusion

We have developed a simple and efficient energy function, GEMSCORE, for the PSP. The GEMSCORE is able to identify 86 native proteins that are the first rank among 96 target proteins from more 70,000 structures on six well-known benchmarks. For these six benchmark datasets, the predictive performance of the GEMSCORE, based on native structure ranking and *Z*-scores, yields very comparable accuracies with the best of eight energy functions and is better than other energy functions. Experiments show that the GEMSCORE is efficient to discriminate between native and nonnative structures from thousands of protein structure candidates. We believe that the GEMSCORE is robust and can be a useful energy function for the PSP.

Methods

Figure 5 shows the flowchart of developing the GEMSCORE for PSP. We selected 30 proteins (56,010 structures, Table S2 in supporting material) from the Rosetta data as the training set for parameter optimized by using the GEM method. After the weights of energy potentials were optimized, other six data sets were used to evaluate the GEMSCORE performance and to compare with previous works.

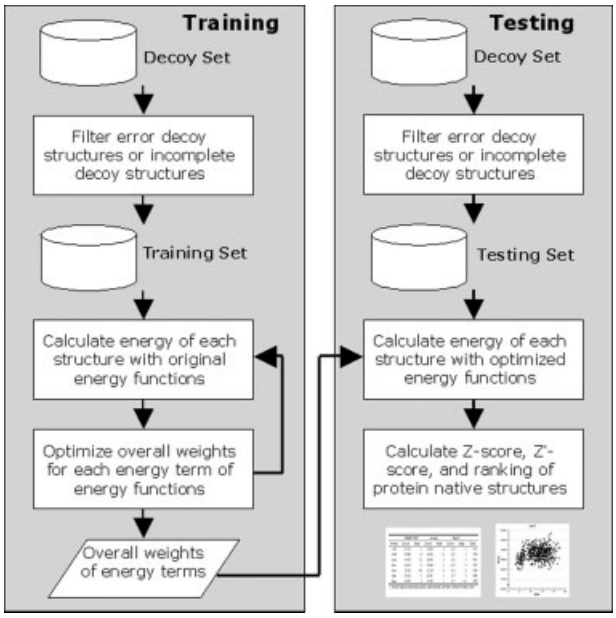

Figure 5. Overview of the GEMSCORE for protein structure prediction.

Energy Terms in GEMSCORE

The GEMSCORE was enhanced and modified from the soft scoring function of the GEMDOCK for PSP by adding solvation potentials and enhancing hydrogen-bonding potentials [eq. (1)]. The GEMSCORE consists of is the electrostatic energy (E_{elect}), van der Waals potential (E_{vdW}), hydrogen-bonding potential on the protein backbone (E_{bHB}), and the solvent potential [E_{SAS}, eq. (2)]. The electrostatic energy (E_{elect}) is defined as $E_{\text{elect}} = \sum_{i=1}^{N}\sum_{j=1}^{N} 332\frac{q_i q_j}{4r_{ij}^2}$, where r_{ij} is the distance between the atoms i and j, q_i and q_j are the formal charges of atoms i and j, and 332 is a factor that converts the electrostatic energy into kilocalories per mole. To increase the tolerance of our new energy function for near-native structures, we set the low bound of the electrostatic energy to -10 (see Fig. 1).

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

$$F(r_{ij}^{B_{ij}}) = \begin{cases} V_6 - \dfrac{V_6 r_{ij}^{B_{ij}}}{V_1} & \text{if } r_{ij}^{B_{ij}} \le V_1 \\ \dfrac{V_5(r_{ij}^{B_{ij}} - V_1)}{V_2 - V_1} & \text{if } V_1 < r_{ij}^{B_{ij}} \le V_2, \\ V_5 & \text{if } V_2 < r_{ij}^{B_{ij}} \le V_3 \\ V_5 - \dfrac{V_5(r_{ij}^{B_{ij}} - V_3)}{V_4 - V_3} & \text{if } V_3 < r_{ij}^{B_{ij}} \le V_4 \\ 0 & \text{if } r_{ij}^{B_{ij}} > V_4 \end{cases} \quad (3)$$

$r_{ij}^{B_{ij}}$ 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 the backbone. The parameters of these different potentials, $V_1,\ldots, V_6$, are given in Figure 1.

Objective Function of GEM

The GEMSCORE has four energy potentials (i.e. E_{elect}, E_{vdW}, E_{bHB}, and E_{SAS}) with 12 parameters (Table 1) defined in eqs. (1) and (2). The GEM was used to find the most suitable energy weights of these 12 parameters by minimizing the objective function given as

$$S = \sum_{i}^{M} (f(z_i) + f(z_i')) \quad (4)$$

where i is the i-th protein; M is the number of proteins in a training set (M is 30 in this paper, Table S2 in supporting mate-

rial); $f(Z_i)$ is the normalized Z-score of the energy value, which is calculated by the eq. (1), of the protein i; $f(Z')$ is a related Z'-score of the protein i.

A good energy function is able to correctly distinguish native structures from non-native structures and to be judged by the size of energy gap between native and nonnative structures. A mostly used measure for assessing this quality is the Z-score defined as $Z = (E_{\text{native}} - \langle E \rangle)/\sigma$, where E_{native} is the energy value of a native structure of a protein; $\langle \text{E} \rangle$ and σ are the mean and standard deviation of energy values of all nonnative structures in a decoy set, respectively. While we seek the weights of an energy function, we makes Z-scores of all proteins in the training set simultaneously low enough. We normalize the Z-score of the protein i by

$$f(Z_i) = \frac{1}{1 + \exp^{-Z_i}}, \tag{5}$$

where $f(Z_i)$ maps the Z-score of the protein i into a value ranging from 0 to 1.

To consider the energy value of a native structure, the lowest among structures in a decoy set, we added a related measure Z'-score given as $Z' = (\text{E}_{\text{native}} - \text{E}_{\text{lowest}})/\sigma$ where E_{lowest} is the lowest energy value among the nonnative structures in the decoy set and σ is the standard deviation of energy value of all nonnative structures. The Z'-score gives a quantitative measure of how well separated the native structure from its lowest energy neighbor in the decoy set. The Z'-score of the protein i is normalized by

$$f(Z_i') = \frac{1}{1 + \exp^{-Z_i'}} \tag{6}$$

The Z-scores and Z'-scores of 30 target proteins in the training set are shown in Table S2 (in supporting material).

Generic Evolutionary Method

The GEM method was used to minimize the objective function to find out optimal weights for 12 energy terms of GEMSCORE. The core idea of the GEM method was to design multiple operators that cooperate using the family competition model, which is similar to a local search procedure. This approach has been successfully applied for some specific problems, such as protein-ligand docking, drug screening, and protein side-chain prediction.[25,32–34] GEM is a multioperator approach that combines three mutation operators: decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. It incorporates family competition 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.

The GEM minimizing the objective function is briefly described as follows: It randomly generates a starting population with N solutions with 12 parameter values of energy terms used in the GEMSCORE (Table 1). 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 $i = 1,\ldots,N$, where N is the population size. n is 12 in this work. The vector x is the adjustable variables representing a particular weights of 12 energy terms to be optimized. σ and ψ are the step-size vectors of decreasing-based Gaussian mutation and self-adaptive Cauchy mutation. For each generated solution x, we can calculate the fitness value of this solution by using the objective function defined in eq. (4).

After the GEM method initializes the solutions, it enters the main evolutionary loop which consists of two stages. 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 two parameters, namely, the working population (P, with N solutions) and mutation operator, 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_j^s is used to denote the j th adjustable optimization variable of a solution s, $\forall\, j \in \{1,\ldots,12\}$.

Recombination Operators

GEM 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:

$$x_i^c = \begin{cases} x_i^a & \text{with probability 0.8} \\ x_i^b & \text{with probability 0.2} \end{cases}$$

The generated child inherits genes from the "family father" with a higher probability 0.8. Intermediate recombination works as: $w_i^c = w_i^a + \beta(w_i^b - w_i^a)/2$, 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 energy term vectors (x).

Mutation Operators

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

$$w_i' = w_i A(\cdot)$$
$$x_i' = x_i + w_i' D(\cdot)$$

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. If the mutation is a self-adaptive mutation, $A(\cdot)$ is evaluated as $\exp[\tau' N(0,1) + \tau N_i(0,1)]$ 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 mutation or Cauchy mutation. For example, the self-adaptive Cauchy mutation is defined as

$$\psi_i^c = \psi_i^a \exp[\tau' N(0,1) + \tau N_i(0,1)]$$
$$x_i^c = x_i^a + \psi_i^c C_i(t).$$

We set τ and τ' to $(\sqrt{2n})^{-1}$ and $(\sqrt{2\sqrt{n}})^{-1}$, respectively, according to the suggestion of evolution strategies.[46] A random variable is said to have the Cauchy distribution ($C(t)$) if it has the density function: $f(y;t) = \dfrac{1/\pi}{t^2 + y^2}, -\infty < y < \infty$. In this work, 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

$$\sigma^c = \gamma \sigma^a,$$
$$x_i^c = x_i^a + \sigma^c N_i(0,1).$$

Acknowledgments

Authors are grateful to both the hardware and software supports of the Structural Bioinformatics Core Facility at National Chiao Tung University.

References

1. Aloy, P.; Pichaud, M.; Russell, R. B. Curr Opin Struct Biol 2005, 15, 15.
2. Hao, M. H.; Scheraga, H. A. Curr Opin Struct Biol 1999, 9, 184.
3. Chen, C. C.; Hwang, J. K.; Yang, J. M. Nucleic Acids Res 2006, 34, W152.
4. Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. Nucleic Acids Res 2003, 31, 3381.
5. Wallner, B.; Fang, H.; Elofsson, A. Proteins 2003, 53, 534.
6. Jones, D. T. J Mol Biol 1999, 287, 797.
7. Bonneau, R.; Tsai, J.; Ruczinski, I.; Chivian, D.; Rohl, C.; Strauss, C. E.; Baker, D. Proteins 2001, Suppl 5, 119.
8. Kihara, D.; Lu, H.; Kolinski, A.; Skolnick, J. Proc Natl Acad Sci USA 2001, 98, 10125.
9. Vajda, S.; Sippl, M.; Novotny, J. Curr Opin Struct Biol 1997, 7, 222.
10. Russ, W. P.; Ranganathan, R. Curr Opin Struct Biol 2002, 12, 447.
11. Anfinsen, C. B. Science 1973, 181, 223.
12. Sippl, M. J. Curr Opin Struct Biol 1995, 5, 229.
13. Gatchell, D. W.; Dennis, S.; Vajda, S. Proteins 2000, 41, 518.
14. Tsuchiya, Y.; Kinoshita, K.; Nakamura, H. Proteins 2004, 55, 885.
15. Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. Proteins 2002, 48, 404.
16. Price, D. J.; Brooks, C. L., III. J Comput Chem 2002, 23, 1045.
17. Zhu, J.; Zhu, Q.; Shi, Y.; Liu, H. Proteins 2003, 52, 598.
18. Fujitsuka, Y.; Takada, S.; Luthey-Schulten, Z. A.; Wolynes, P. G. Proteins 2004, 54, 88.
19. Lee, M. C.; Duan, Y. Proteins 2004, 55, 620.
20. Hsieh, M. J.; Luo, R. Proteins 2004, 56, 475.
21. Eisenberg, D.; McLachlan, A. D. Nature 1986, 319, 199.
22. Wesson, L.; Eisenberg, D. Protein Sci 1992, 1, 227.
23. Bhattacharyay, A.; Trovato, A.; Seno, F. Proteins 2007, 67, 531.
24. Deshpande, N.; Addess, K. J.; Bluhm, W. F.; Merino-Ott, J. C.; Townsend-Merino, W.; Zhang, Q.; Knezevich, C.; Xie, L.; Chen, L.; Feng, Z.; Green, R. K.; Flippen-Anderson, J. L.; Westbrook, J.; Berman, H. M.; Bourne, P. E. Nucleic Acids Res 2005, 33, D233.
25. Yang, J.-M.; Chen, C.-C. Proteins 2004, 55, 288.
26. Yang, J.-M. J Comput Chem 2004, 25, 843.
27. Yang, J.-M.; Shen, T.-W. Proteins 2005, 59, 205.
28. Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B. S.; Hsu, D. F. J Chem Inform Model 2005, 45, 1134.
29. Palma, P. N.; Krippahl, L.; Wampler, J. E.; Moura, J. J. G. Proteins 2000, 39, 372.
30. Fernandez-Recio, J.; Totrov, M.; Abagyan, R. Protein Sci 2002, 11, 280.
31. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. J Comput Chem 2003, 24, 1637.
32. Yang, J. M.; Tsai, C. H.; Hwang, M. J.; Tsai, H. K.; Hwang, J. K.; Kao, C. Y. Protein Sci 2002, 11, 1897.
33. Yang, J. M.; Kao, C. Y. IEEE/OSA J Lightwave Technol 2001, 19, 559.
34. Yang, J. M.; Kao, C. Y. Neural Comput Appl 2001, 10, 214.
35. Tsai, J.; Bonneau, R.; Morozov, A. V.; Kuhlman, B.; Rohl, C. A.; Baker, D. Proteins 2003, 53, 76.
36. Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B. A.; Bystroff, C.; Baker, D. Proteins 1999, 34, 82.
37. Holm, L.; Sander, C. J Mol Biol 1992, 225, 93.
38. Park, B.; Levitt, M. J Mol Biol 1996, 258, 367.
39. Samudrala, R.; Levitt, M. Protein Sci 2000, 9, 1399.
40. Samudrala, R.; Xia, Y.; Levitt, M.; Huang, E. S. In Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, Honolulu, 1999; pp 505.
41. Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. J Mol Biol 1997, 268, 209.
42. Zhou, H.; Zhou, Y. Protein Sci 2002, 11, 2714.
43. Zhang, C.; Liu, S.; Zhou, H.; Zhou, Y. Protein Sci 2004, 13, 400.
44. Krishnamoorthy, B.; Tropsha, A. Bioinformatics 2003, 19, 1540.
45. Hu, C.; Li, X.; Liang, J. Bioinformatics 2004, 20, 3080.
46. Back, T. Evolutionary Algorithms in Theory and Practice; Oxford University: New York, 1996.