Perceptron learning of pairwise contact energies for proteins incorporating the amino acid environment

Muyoung Heo, Suhkmann Kim, Eun-Joung Moon, Mookyung Cheon, Kwanghoon Chung, and Iksoo Chang
National Research Laboratory for Computational Proteomics and Biophysics, Department of Physics, Pusan National University, Busan 609-735, Korea

(Received 6 February 2004; revised manuscript received 10 May 2005; published 12 July 2005)

Although a coarse-grained description of proteins is a simple and convenient way to attack the protein folding problem, the construction of a global pairwise energy function which can simultaneously recognize the native folds of many proteins has resulted in partial success. We have sought the possibility of a systematic improvement of this pairwise-contact energy function as we extended the parameter space of amino acids, incorporating local environments of amino acids, beyond a 20×20 matrix. We have studied the pairwise contact energy functions of 20×20, 60×60, and 180×180 matrices depending on the extent of parameter space, and compared their effect on the learnability of energy parameters in the context of a gapless threading, bearing in mind that a 20×20 pairwise contact matrix has been shown to be too simple to recognize the native folds of many proteins. In this paper, we show that the construction of a global pairwise energy function was achieved using 1006 training proteins of a homology of less than 30%, which include all representatives of different protein classes. After parametrizing the local environments of the amino acids into nine categories depending on three secondary structures and three kinds of hydrophobicity (desolvation), the 16290 pairwise contact energies (scores) of the amino acids could be determined by perceptron learning and protein threading. These could simultaneously recognize all the native folds of the 1006 training proteins. When these energy parameters were tested on the 382 test proteins of a homology of less than 90%, 370 (96.9%) proteins could recognize their native folds. We set up a simple thermodynamic framework in the conformational space of decoys to calculate the unfolded fraction and the specific heat of real proteins. The different thermodynamic stabilities of *E.coli* ribonuclease H (RNase H) and its mutants were well described in our calculation, agreeing with the experiment.

DOI: 10.1103/PhysRevE.72.011906 PACS number(s): 87.14.Ee, 87.15.Aa, 87.15.Cc

I. INTRODUCTION

One of the most important problems in bioinformatics, biophysics, biology, and computer science is the protein folding problem. Three big issues are the prediction of protein structure, the design of amino acids sequence, and the understanding of the protein folding mechanism [1–12]. The main difficulty in solving the protein folding problem is the complicated nature of interaction energies between atoms in the protein. One can, in principle, develop an atomistic energy function for all atoms in the protein and look for the minimum energy conformations from the first principle. But, this approach has limited success only for the shorter proteins, and it requires precise energy functions for the longer proteins and huge computational resources. It is generally accepted that the amino acid sequence contains the essential features of proteins and that its native structure corresponds to that of the minimum free energy [1,9]. Therefore, it is practically important to develop a protein energy function which depends on the sequence, the character of amino acids, and can recognize the native structures of proteins.

The usual approach is to use a coarse-grained picture of amino acids after integrating out the details of proteins [1–17]. Each amino acid is considered as an isotropic sphere centered at the position of the C_α atom on the backbone of a protein, and one considers the interactions between these points instead of the interactions between all atoms. Since the details of a protein are coarse-grained, one needs the appropriate energy function which can recognize the native folds and capture the essential thermodynamic features of proteins. There have been many attempts [3,5–7,10–12] to design and construct such protein energy functions based on the known structures of proteins in the Protein Data Bank (PDB) (http://www.rcsb.org/pdb) [17]. The main aim has been to come up with a global protein energy function, based on a sequence of 20 amino acids, which can recognize the native folds of as many proteins as possible, so that one can proceed to use it for the thermodynamic description of proteins. The first effort was put forward by Miyazawa and Jernigan [10], who constructed a 20×20 matrix for the pairwise contact energies (scores) of the 20 amino acids using the quasichemical approximation. Although this pairwise energy function is simple, it could explain many of the characteristic properties of proteins and has had a great impact on describing the statistical properties of proteins. There have been several attempts to modify it to acquire better ability to recognize more proteins [11,12]. These approaches basically count the frequencies of pairwise contacts of two amino acids which are within the threshold distance in the protein structure.

Zhang and Kim [12], however, using the quasichemical approximation, expanded the parameter space of amino acids by considering the secondary structure in which each amino acid resides within the protein structure. Each amino acid can be in one of three secondary structures (α-helix, β-strand, others); thus the pairwise contact matrix becomes

60×60. Their contact matrix recognized the native folds of more than 97% of the 316 testing proteins whose length was less than 200 residues, but could not recognize all of them. Once the set of test proteins used in the statistical counting of pairwise contacts is given, the values of 1830 independent parameters are fixed and there is no room to improve them further to recognize all of the native folds of the test proteins.

In order to construct a protein energy function which can achieve the complete recognition of all native folds of test proteins, two optimization schemes for the protein energy parameters were employed. The basic idea was to optimize the energy parameters so that the energy of the amino acid sequence housed in the known native structure was always lower than in any decoy structures. The Z-score method maximizes the difference between the native energy and the average energy of the decoy structures, which may lead to the maximum stability of the native state against decoys [3,13]. However, there can be a few decoys whose energies are always lower than the native energy, even with the maximum Z-score. Therefore, the maximum Z-score does not always grant the complete recognition of all native folds simultaneously. An alternative approach is the perceptron learning method of the energy parameters, which optimizes them in order to achieve the complete recognition of the native folds of the training proteins simultaneously [5–7,16].

Whether one can determine the pairwise energy parameters using the perceptron learning method, which achieves the complete recognition of the native folds of training proteins, is always an important and a difficult question. Domany *et al.* [4,5] showed that the possibility of successfully constructing such a pairwise contact energy function depends on the competing conformations against a native one and on the parametrization scheme of energy parameters for amino acids. As long as one adopts a 20×20 pairwise contact matrix in which the local environmental features of amino acids in protein structures are not considered, the answer depends on whether one obtains the competing decoys from a gapless threading or from competitive conformations by energy minimization. With decoys from a gapless threading, one can simultaneously recognize the native folds for a typical subset of roughly 100 proteins at most, above which there is no set of pairwise contact parameters. With decoys from competitive low-energy conformations, it was shown that it was not possible to construct a 20×20 pairwise contact matrix to recognize the native fold of a protein. This demonstrates that a simple parametrization for an energy function, such as a 20×20 pairwise contact matrix, does not recognize the native folds of proteins [5,6] since it is too crude a method to catch the structural and the energetic characters of amino acids in protein structures. However, they made an important suggestion that the inclusion of a hydrophobicity (desolvation) and a local structural feature can be a possible direction to explore in constructing the global protein energy function [5].

Within the context of employing decoys from a gapless threading, we have sought the possibility of a systematic improvement of the pairwise contact energy function as we extended the parameter space of amino acids, incorporating local environments of amino acids, beyond a 20×20 matrix. We have studied the pairwise contact energy functions of 20×20 (ε_{20}^{20}), 60×60 (ε_{60}^{60}), and 180×180 (ε_{180}^{180}) matrices depending on the extent of parameter space, and compared their effect on the learnability of energy parameters in the context of a gapless threading, bearing in mind that a 20×20 pairwise contact matrix was shown to be too simple to recognize the native folds of proteins. In this paper, we show that the construction of a global pairwise energy function was achieved using 1006 training proteins (P_{train}^{1006}) of a homology of less than 30% which include all the representatives of different protein classes. After parametrizing the local environments of amino acids into nine categories depending on three secondary structures and three kinds of hydrophobicity of amino acids, the ε_{180}^{180} matrix could be determined by perceptron learning and protein threading. These could simultaneously recognize all the native folds of P_{train}^{1006}. When these parameters were tested on a separate set of 382 test proteins (P_{test}^{382}) with high homology, 370 (96.9%) proteins could recognize their native folds. We set up a simple thermodynamic framework in the conformational space of decoys to calculate the unfolded fraction and the specific heat of real proteins. The different thermodynamic stabilities of *E.coli* RNase H and its mutants were well described in our calculation, agreeing well with the experiment [23].

In Sec. II, the energy function, the protein data set, and the classification of the local environments of amino acids are introduced. In Sec. III the perceptron learning of energy parameters, and in Sec. IV the threading test of our energy parameters and discussions on their quality are presented. In Sec. V, the thermodynamic stabilities of *E.coli* RNase H and its mutants are evaluated. Conclusions and the possible applications of this work are summarized in Sec. VI.

II. ENERGY FUNCTION, PROTEIN DATA SET, AND LOCAL ENVIRONMENTS OF AMINO ACIDS

Given a sequence of amino acids, we needed an energy function which could assess the fit of a sequence to the native structure or decoy structures. We employed the coarse-grained representation of amino acids and their local environmental information. One may construct a simple energy function according to the propensities of pairwise interactions of amino acids in the different local environments. The basic strategy is to optimize these propensities so that the energy of a sequence in the native structure is always lower than in the competing decoy structures. This criterion should also apply to the set of many proteins simultaneously. The energy function we used is the following:

$$H(s,\Gamma)=\sum_{i,j}\sum_{k,l}n(i,j;k,l)\varepsilon(i,j;k,l), \tag{1}$$

where H is the energy function which is a measure of how well a sequence s is housed in a structure Γ. The elements in the sum are the environment-dependent pairwise contact energies between two amino acids where $n(i,j;k,l)$ is the number of pairs of two amino acids of type i,j found in the local environment k,l, respectively, making the pairwise contacts within a threshold distance 6.5 Å, and $\varepsilon(i,j;k,l)$ is the energy associated with it. We considered the pairs of amino

acids only when two amino acids are separated by more than a three-peptide-bond unit along the backbone of a protein. Once the structures of proteins are given, $n(i,j;k,l)$ are determined from the PDBs. Our aim is to extract the energy parameters $\varepsilon(i,j;k,l)$ in the ε_{180}^{180} matrix to ensure the recognition of the native folds with the maximum stability among a set of decoy structures. Since this matrix is symmetric, the number of independent parameters is 16 290.

We chose P_{train}^{1006} from PDB select (http://www.cmbi.kun.nl/gv/pdbsel) and WhatIf (http://www.cmbi.kun.nl/whatif) [17]. In fact, there are 3032 representative proteins of a homology of less than 30% in PDB select and WhatIf, covering all of the different classes according to the structural classification of proteins (SCOP) [18]. The definition of "*X*% homology" for a set of proteins is that the proteins in such a set have a sequence similarity of less than *X*% between proteins. Therefore, *X*% is a cutoff (maximum) similarity between protein sequences. We used a representative list of PDB at the web site of PDB select and of WhatIf [17]. There, the sequence similarity was set up by an all-against-all Smith/Waterman sequence alignment algorithm between proteins. Local alignment searches for regions of local similarity between two sequences (of different length, for example) were conducted, and the entire length of the sequences did not need to be included. Therefore, local alignment methods were used to find matches between small regions of sequences between two proteins. Among these 3032 proteins, we selected those (1) whose structures were obtained by x-ray crystallography, (2) which do not have nonstandard amino acids, (3) which are not disconnected chains, and (4) which are not mutant structures. As a result, we had a training set of 1006 nonredundant proteins, whose length ranged from 53 to 994 amino acids.

We classified the local environments of the amino acids of a sequence in the protein structure into nine categories. Each amino acid can be found in one of three secondary structures (α-helix, β-strand, other). The PDB file for each protein in our study includes information about the assignment of the secondary structures of amino acids in the protein structure. We followed a HELIX/SHEET record inside a PDB file for amino acids that was assigned by the experimentalists. Unless otherwise assigned specifically by this list, the secondary structures of amino acids were assigned to "other," that is, to "loop" structure. The solvent-exposed ratio of amino acid was calculated using Richards' algorithm [19,20] as the ratio between the solvent accessible area of each amino acid, *X*, in its native structure and the corresponding area in the Gly-*X*-Gly extended structure. The values of the solvent exposed ratios $<10\%$, $10-50\ \%$, and $>50\%$, capturing the degree of hydrophobicity, were classified into the three classes of small, medium, and large exposure, respectively. Once this environmental classification of amino acids was completed, the 3D structural information of a sequence was transformed into the 1D string of local environmental parameters. Therefore, the energy function in Eq. (1) provides a quantitative measure of the propensities of the pairwise contacts of two amino acids within their corresponding local environments.

III. PERCEPTRON LEARNING OF ENERGY PARAMETERS

We first generated decoys of each protein by a gapless threading of P_{train}^{1006} on themselves. The sequence of each target protein was threaded on the structures (environments) Γ of all proteins of P_{train}^{1006} with a length equal to or longer than a target protein. The solvent-accessible area of amino acids mounted on a threaded fragment was approximated to be the same as that in the longer protein from which the fragment was taken. The total number of decoys for P_{train}^{1006} was about 78.2 million, and each decoy had to satisfy the following inequality to recognize the native folds of P_{train}^{1006} [5–7,16,21]:

$$\sum_{i,j=1}^{20}\sum_{k,l=1}^{9}[n(i,j;k,l)^D - n(i,j;k,l)]\varepsilon(i,j;k,l) > 0, \qquad (2)$$

where $n(i,j;k,l)^D$ and $n(i,j;k,l)$ are the occurrences of a pairwise contact $(i,j;k,l)$ in decoy D $(=1,2,\ldots,78.2$ million) and in its native structure, respectively. Our aim is to determine and to optimize the 16 290 parameters of $\varepsilon(i,j;k,l)$ to ensure that P_{train}^{1006} of known native structure have lower energies than when their sequences are housed in the decoy structures.

The general strategy to obtain the solution $\vec{\varepsilon} \equiv \{\varepsilon(i,j;k,l)\}$ is to determine the values of $\varepsilon(i,j;k,l)$ which satisfy Eq. (2) simultaneously for $D=1,2,\ldots,$ 78.2 million in the 16 290-dimensional space of parameters,

$$\sum_{i,j=1}^{20}\sum_{k,l=1}^{9}[n(i,j;k,l)^D - n(i,j;k,l)]\varepsilon(i,j;k,l) = \vec{n}^D \cdot \vec{\varepsilon} > 0. \qquad (3)$$

Here, $\vec{n}^D=[n(i,j;k,l)^D-n(i,j;k,l)]$ is fixed once the set of P_{train}^{1006} is known, and $\vec{\varepsilon}$ is the unknown vector to be determined. We started from an initial value of $\varepsilon_o(i,j;k,l)$ and calculated the scalar product $\vec{n}^D$ on $\vec{\varepsilon}$ for all 78.2 million inequalities. The vectors $\vec{n}^D$ whose gap $\vec{n}^D\cdot\vec{\varepsilon}$ is negative are the ones which do not satisfy the above inequality and the corresponding decoys are deemed to be the failed decoys. We selected the worst vector $\vec{n}^w$ among the failed decoys, which had the lowest value of energy gap, and updated $\vec{\varepsilon}_{t+1}=\vec{\varepsilon}_t+\alpha\vec{n}^w/|\vec{\varepsilon}_t+\alpha\vec{n}^w|$ $(0<\alpha<1)$ so that the energy gap for the worst decoy w increased. The 78.2 million scalar products were calculated again with the new $\vec{\varepsilon}_{t+1}$, and the set of failed decoys and the worst decoy were identified in order to update $\vec{\varepsilon}_{t+1}$ again. This procedure was iterated until the number of failed decoys out of 78.2 million decoys became zero. The main purpose of this update is to find $\vec{\varepsilon}$, which can stabilize the energies of the native states against the energies of the decoy structures so that the native states can be fully recognized. If a solution for Eq. (2) exists, namely that $\vec{\varepsilon}_{final}$ satisfies all 78.2 million inequalities, the vector $\vec{\varepsilon}_{final}$ converges to a region of points in the 16 290-dimensional space and the energy gap of the worst decoy $\vec{n}^w\cdot\vec{\varepsilon}_{final}$ becomes a positive finite within a finite number of iterations. If the iteration runs forever, neither providing a converging value of $\vec{\varepsilon}$ nor a positive finite value for the energy gap, the perceptron learning does not work, which means that the parametri-

TABLE I. The list of 12 failed proteins out of 382 test proteins. The number of failed decoys (N_{fd}) is within the lowest 0.1% of the total number of decoys (N_{td}) for each protein, showing that the native folds are almost recognized even for the failed proteins. N_{AA} is the length of each protein.

PDB	N_{AA}	N_{td}	N_{fd}	PDB	N_{AA}	N_{td}	N_{fd}
1FYN	62	77160	5	1MHO	88	67522	2
1BWO	90	66795	3	1HRO	105	61461	75
1RDS	105	61461	2	1CO6	107	60764	2
1HE7	107	60764	4	1JSG	111	59402	10
1G96	111	59402	5	1H6W	151	47251	43
1IHK	157	45603	7	1MUP	157	45603	1

zation in the energy function [Eq. (1)] is not adequate. But it is also impractical to solve all 78.2 million inequalities simultaneously.

The basic ingredient for determining the optimal $\varepsilon(i,j;k,l)$, instead of solving all 78.2 million inequalities, is the following. Given that for each training protein there are many decoys generated from a protein threading, we imposed the condition that the native energy of a given protein must be lower than both (1) the average energy of a random sequence on its own native structure with the same composition of amino acids [21], and (2) the average energy of the sequence on the decoy structures. The former generated 1006 inequalities, and the latter, 1005. We first solved these 2011 inequalities by perceptron learning, the solution of which guides the approximate direction of the ultimate solution $\varepsilon(i,j;k,l)$ in 16 290-dimensional parameter space. Using these learned $\varepsilon(i,j;k,l)$, we performed a threading test to compare the energies of all 78.2 million decoys with their native state energies. The number of failed decoys whose energies are lower than their native state energies is 712 out of 78.2 million.

The inequalities from the failed decoys were added to the previous 2011 inequalities, and for all, perceptron learning, taking the (learned) $\varepsilon(i,j;k,l)$ as the initial condition, was performed again to find the new solution for $\varepsilon(i,j;k,l)$. We tried to achieve the maximum stability of native state against the competing decoys by maximizing the energy gap between the native energies of P_{train}^{1006} and their failed decoys. Then, a second threading test of P_{train}^{1006} with the new $\varepsilon(i,j;k,l)$ produced a new set of failed decoys to add to the previous set of inequalities. We iterated the procedures of (i) perceptron learning for updating energy parameters and (ii) protein threading, adding new inequalities until the number of failed decoys to add became zero [5–7,16,21]. When this was achieved, the total number of inequalities to solve was 2755.

Although the solution $\varepsilon(i,j;k,l)$ for solving the 2755 inequalities satisfied all of the 78.2 million inequalities, it was neither unique nor optimized. The optimization strategy is to push the energies of competing decoys as far away as possible from the native state energy so that the maximum stabilities of the native states of P_{train}^{1006} are achieved. For this purpose, we identified the competing decoys (among all 78.2 million decoys) whose energy gaps from their native state energy were smaller than the minimum gap of the 2755 decoys. Again we added the inequalities for these competing decoys to the previous 2755 inequalities, and thereby learned the optimized solution $\varepsilon(i,j;k,l)$ [5,16]. We also iterated the procedures of perceptron learning and protein threading until the number of competing decoys (among all 78.2 million decoys) whose energy gap was smaller than the minimum gap of the previous inequalities became zero, which resulted in solving just 3903 inequalities. We could optimize 16 290 pairwise contact energy parameters simultaneously, which recognized 100% of the native states of P_{train}^{1006} [22].

IV. THREADING TEST OF PAIRWISE ENERGY PARAMETERS AND THEIR QUALITY

A. How well does the energy function recognize native folds?

After we succeeded in learning the pairwise energy parameters, we checked the capability of our parameters to recognize the native folds of proteins that were not present in our learning set. We chose P_{test}^{382} that were distinct from P_{train}^{1006}. These 382 proteins have an average homology of 34% with a maximum of 90% among them. The reason we chose these highly homologous proteins for the threading test was to generate highly competitive decoys from threading P_{test}^{382} on themselves so that it would be hard to distinguish a native energy from the decoys' energies. We tried to impose difficult conditions on the threading test. Nevertheless, the threading test of P_{test}^{382} on themselves using the determined energy parameters showed that the native folds of 370 (96.9%) proteins could be recognized, and there were only 159 failed decoys out of the total of 12.1 million decoys. Table I lists 12 failed proteins, and the number of failed decoys for each protein was within the lowest 0.1% of the total number of decoys. For the further test of the capability for our energy parameters to recognize the native folds of test proteins, we also constructed decoys for test proteins by threading them not only within the set of test proteins but also onto the proteins in the training set or in the whole set since the more decoys, the better the threading test. Threading P_{test}^{382} onto P_{train}^{1006} (P_{whole}^{1388}) generated 28.5 (40.6) million decoys, and the native folds of 366 (95.8%) [365 (95.5%)] proteins could be recognized with 415 (574) failed decoys, respectively, using the determined energy parameters. In view of the fact that we chose P_{test}^{382} of a homology of less than 90% in order to perform a stringent threading test, the success ratio of more than 95% is a very good one. We

TABLE II. The success ratios for the proteins (from a set of 382 test proteins) belonging to α, β, α/β, and $\alpha+\beta$ classes when subject to the threading test. It shows a uniform success ratio of more than 90% for the different classes of proteins.

	α	β	α/β	$\alpha+\beta$	total
The number of proteins	87	101	122	72	382
The number of failed proteins	4	6	0	2	12
Success ratio (%)	95.4	94.1	100	97.2	96.9

classified the P_{test}^{382} according to their SCOP classification, as α, β, α/β, and $\alpha+\beta$ classes. We checked whether our pairwise energy parameters could provide a uniform success ratio of more than 90% for the different classes in the threading test. Table II shows such a success ratio for the proteins belonging to each class when they are subject to the threading test on P_{test}^{382}.

Taking the energy parameters at the stage immediately after applying the inequalities of (1) native energy lower than average energy for random sequence and (2) native energy lower than average decoy energy to P_{train}^{1006} as mentioned in Sec. III, and threading P_{train}^{1006} on themselves, only 712 out of 78.2 million decoys failed. This is lower than the ratio of failed decoys in the test set (159 out of 12.1 million). Therefore, one might wonder whether the introduction of the inequalities for individual decoys actually improve the results for the test set at all, or if this is just an overfitting. Threading P_{test}^{382} on themselves using the same intermediate energy parameters, the number of failed decoys was 3403 out of 12.1 million and 352 (92.1%) native folds of the 382 proteins could be recognized. However, when we included the inequalities for individual decoys for learning the energy parameters after procedures (1) and (2), the learned 16 290 parameters could recognize all the native folds of P_{train}^{1006} and 370 (96.9%) native folds of P_{test}^{382}, and the number of failed decoys was 159 out of 12.1 million decoys. Thus, the introduction of inequalities for individual decoys indeed improved the results of native folds recognition for the test set and decreased the number of failed decoys.

Since we took an approximation to take the solvent accessible area of amino acid mounted on a threaded fragment to be the same as that in the longer protein, one might worry that the decoy structures were less compact than the native structures, which might have resulted in the easy determination of our energy function. In order to show that our decoy set is sufficient and illustrate the quality of our energy function for the recognition of native folds, we tested our energy function against the nativelike decoys. Therefore, we also tested the quality of our 16 290 pairwise-contact parameters against Levitt's decoy database at http://dd.stanford.edu. Table III shows the results of the threading test for recognizing 15 proteins against Levitt's decoys generated by (a) a lattice-ssfit algorithm and (b) a four-state-reduced algorithm. When we used our energy parameters extracted from P_{train}^{1006}, Table III(a) shows the number of failed decoys whose energies were less than the energy of a native state. Eight proteins out of 15 could recognize their native folds, and the number of failed decoys for the remaining seven proteins were, at most, within the lowest 3.5% of the corresponding total numbers of decoys. We noticed that 1BEO and 1R69 were already included in our set of P_{train}^{1006}. However, our energy parameters, constructed using decoys from the gapless threading of P_{train}^{1006}, could successfully recognize the native folds of 1BEO and 1R69 against the nativelike decoys generated by Levitt. It was useful to construct again 16 290 new energy parameters from the gapless threading of P_{train}^{1006} together with 15 Levitt's proteins, and to check how well the native folds of the 15 proteins were recognized against Levitt's nativelike decoys. Table III(b) shows that new energy parameters could recognize the native folds of 12 proteins completely and of three proteins almost completely. Our energy function could also recognize 22 native folds out of 24 single-chain target proteins from CASP4 and CASP5 competition (Critical Assessment of Techniques for Protein Structure Prediction) (http://predictioncenter.llnl.gov/). For each target protein, we chose the 20 best predicted structures provided by the participants of CASP4 and CASP5, among which the one with the lowest energy as determined by our energy function could pass a threading test on P_{train}^{1006}. These tests convincingly illustrated the quality of our energy function for native folds recognition.

B. Comparision to other energy functions and approaches

We also constructed two kinds of ε_{60}^{60} pairwise contact matrices considering either the secondary structure or the hydrophobicity of amino acids using perceptron learning and protein threading. Both of these could simultaneously recognize the native folds of all the native states of P_{train}^{1006}. When these two energy functions were subject to a threading test on P_{test}^{382}, the energy function with the secondary structure information could recognize the native folds of 300 (78.5%) proteins out of P_{train}^{1006}, whereas 367 (96%) proteins were recognized by the energy function with the hydrophobicity information. This illustrated the better role played by the hydrophobicity of amino acids in recognizing the native structures of proteins. Both the hydrophobicity and the secondary structure are important in assessing the protein structure, and the impact of hydrophobicity is elucidated in this calculation through the process of designing global pairwise contact energies for proteins. We repeated the same calculation of perceptron learning for the ε_{20}^{20} matrix, and we were not able to determine a function which could recognize the native folds of P_{train}^{1006} simultaneously. This is in accord with the previous finding [5,6] that it is not possible to parametrize a simple function, such as ε_{20}^{20} matrix, to recognize the native folds of proteins even with a gapless threading. Our

TABLE III. The results of the threading test for 15 proteins against Levitt's decoys at http://dd.stanford.edu. N_{AA} is the number of amino acids, N_d is the number of decoys, and N_{fd} is the number of failed decoys for each protein from a threading test, against Levitt's difficult decoys, using the energy parameters obtained by perceptron learning and the gapless threading of (a) 1006 proteins and (b) 1006 proteins together with Levitt's 15 proteins.

Algorithm for generating decoys	PDB	N_{AA}	N_d	N_{fd}	
				(a)	(b)
Lattice_ssfit algorithm	1BEO	98	2000	0	0
	1CTF	68	2000	0	0
	1FCA	55	2000	33	0
	1NKL	78	2000	0	0
	1PGB	56	2000	0	0
	1TRL	62	2000	2	0
	1DKT	72	2000	0	0
	4ICB	76	2000	0	0
Four-state_reduced algorithm	1CTF	68	631	1	0
	1R69	63	676	0	0
	1SN3	65	661	5	2
	2CRO	65	675	24	4
	3ICB	75	654	4	3
	4PTI	58	688	0	0
	4RXN	54	678	1	0

work, however, illustrated that ε_{60}^{60} and ε_{180}^{180} matrices incorporating the local environments of amino acids are learnable and can recognize the native folds of P_{train}^{1006} simultaneously. And it showed the important role played by the secondary structures and the hydrophobicity of amino acids in recognizing the native folds of proteins even in the context of employing competing conformations by a gapless threading. It would be worthwhile to construct the 60×60 or the 180×180 pairwise contact energy parameters, especially with respect to the nativelike decoys of true low energies, recognizing the native folds of the 1006 training proteins simultaneously, although this was not feasible for the 20×20 pairwise contact energy parameters [5,6].

Bowie *et al.* [13] constructed a scoring profile by a statistical counting of amino acids found in different local environments for converting the three-dimensional structural information of proteins to a one-dimensional string of local environments of amino acids. They classified the local environments of amino acids according to 18 classes with the total $20\times18=360$ parameters. Their aim was to determine the favorable alignment of a protein sequence to the environmental string of a protein whose 3D structure was already known. They demonstrated that a 3D-1D structural profile could detect amino acid sequences compatible with a known 3D structure of a protein, using the Z-score method, for four families of proteins: the globins, cyclic AMP receptor proteins, ribose binding proteins, and the actins. Wilmanns and Eisenberg [14] presented other modified pair-preference 3D profiles (type II: 210×8 matrix; type III: 80×9 matrix; type IV: 80×9 matrix) that characterize the local environments according to the statistical preferences of the profiled residue for neighbors of specific residue types, main-chain conformations, or secondary structure. They combined the original and three pair-preference 3D-profile methods for the identification of a sequence of β/α barrel proteins and showed that these combined profiles enhance the assignment of sequences to known 3D structures. However, they used too few proteins to construct good scoring profiles, since the quality of these statistical approaches depends on a 3D-structure database. In this paper, we introduced nine local environments of amino acids leading to $9\times20=180$ possible residue/environment assignments, but then constructed an ε_{180}^{180} energy function, which was similar to the scoring profiles mentioned above. The purpose of our energy parameters, constructed from 1006 training proteins, was the recognition of the native structure of a protein against conformational decoys by protein threading for a given sequence of amino acids, whereas that of the score profiles by the Eisenberg group was to detect amino acid sequences compatible with a known 3D structure of a protein. Therefore, the performances of these two approaches are not subject to a direct comparison. Within the context of our work, we could, however, construct the scoring profiles for the 180 one-body parameters and the 16 290 pairwise contact parameters adopting the same definition used by the Eisenberg group [13], that of a statistical counting of the occurrence of amino acids being in different local environments. And then, we compared the success ratios of recognizing native folds of P_{train}^{1006} in the threading test when using these two kinds of energy parameters. The 627 (62.4%) [968 (96.3%)] native folds out of the 1006 training proteins were recognized by the 180 one-body (16 290 pairwise contact) energy parameters. Although the ratio between the numbers of energy parameters does not directly reflect better or poor performance of recog-

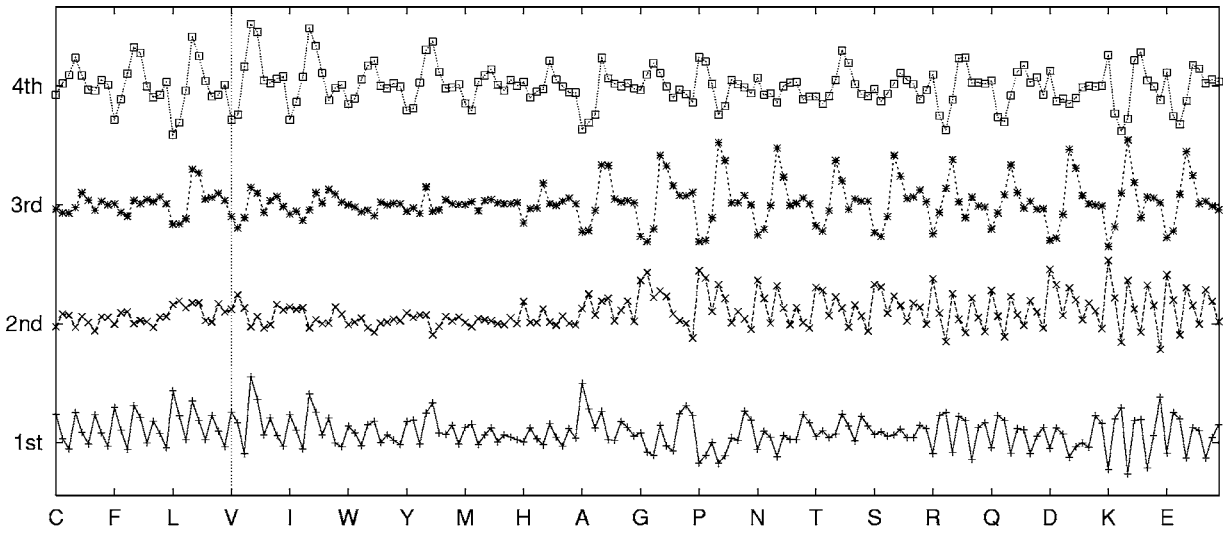

FIG. 1. Relative amplitudes of eigenvector components for the four largest eigenvalues of ε_{180}^{180} matrix. "1st" represents the largest eigenvalue, "2nd" the second largest value, and so on.

nizing native folds, our test showed a systematic improvement of the performance of the energy function as we extended the parameter space of energy parameters from the 20×9 one-body matrix to the 180×180 pairwise contact matrix.

C. Characterization of energy function

In order to study the physiochemical characteristics of the amino acids extracted from the 16 290 energy parameters, we performed an eigenvector analysis of an ε_{180}^{180} symmetric matrix constructed of 16 290 parameters and presented the relative magnitudes of the eigenvector components of the four highest eigenvalues in Fig. 1. The horizontal axis represents amino acids and their nine local environments in order of α-buried, α-medium, α-exposed, β-buried, β-medium, β-exposed, other-buried, other-medium, and other-exposed. The first (largest) eigenvalue shows a behavior of period three for (C, F, L, V, I) and (E, K, Q, R); namely, the first eigenvector mostly represents the hydrophobicity or the hydrophilicity of the amino acids. The first eigenvector also gave some specific features of the amino acids, such that Valine (V) has a strong propensity for buried β-strand, Alanine (A) has a strong propensity for α-helix, and Glycine (G) and Proline (P) have a strong propensity for other structures. Also the same behavior of period 3 is seen for the hydrophilic amino acids (N, T, S, R, Q, D, K, E) in the eigenvector of the second largest eigenvalue. However, the eigenvectors of the third and the fourth largest eigenvalues show a behavior of period nine for hydrophilic and hydrophobic amino acids, respectively, which means that, in the third and the fourth eigenvectors, the information for the secondary structures is inherited. A singular value decomposition analysis of this ε_{180}^{180} matrix yielded information similar to that observed here. The detaild biological interpretations for these 16 290 parameters by both the self-organizing map (SOM) method and singular value decomposition (SVD) analysis will be presented elsewhere soon.

V. THERMODYNAMIC STABILITIES OF MESOPHILIC AND THERMOPHILIC *E.coli* RNase H PROTEINS

We applied our pairwise energy parameters to evaluate the changes in the thermodynamic stabilities of several *E.coli* RNase H proteins, and compared the results with those of the experiment [23]. RNase H (1RDD) consists of 155 amino acids and has five helices and five strands. It is known as a mesophilic protein, but the point mutations H62P(1RBR), V74L(1LAV), K95G(1RBT), V74I(1LAW), and K95N(1RBU) can convert it to a thermophilic protein, which is more stable than the wild type. Since our energy parameters could stabilize P_{train}^{1006} completely and could recognize more than 96% of P_{test}^{382}, they should also be able to pick up the thermodynamic changes of RNase H protein due to mutations. For each RNase H mutant, we treated the set of decoys given by threading the sequence onto P_{train}^{1006} as the ensemble of excited states required to establish an approximate partition function,

$$Z(s) = \sum_{\{\Gamma\}} e^{-H(s,\Gamma)/T}, \tag{4}$$

where the sum represents the conformational space $\{\Gamma\}$ of decoys and T is a temperature with an arbitrary unit. The probability to find a sequence s in a structure Γ is given by

$$P_{\Gamma}(s) = \exp[-H(s,\Gamma)/T] \Big/ \sum_{\{\Gamma'\}} \exp[-H(s,\Gamma')/T] \tag{5}$$

and the unfolded fraction becomes

$$P_{\mathrm{UF}} = 1 - \exp[-H(s,\bar{\Gamma})/T] \Big/ \sum_{\{\Gamma'\}} \exp[-H(s,\Gamma')/T], \tag{6}$$

where $\bar{\Gamma}$ is the native-state structure for a sequence s. One can evaluate the fluctuation in the energy of a protein as T varies, namely the specific heat

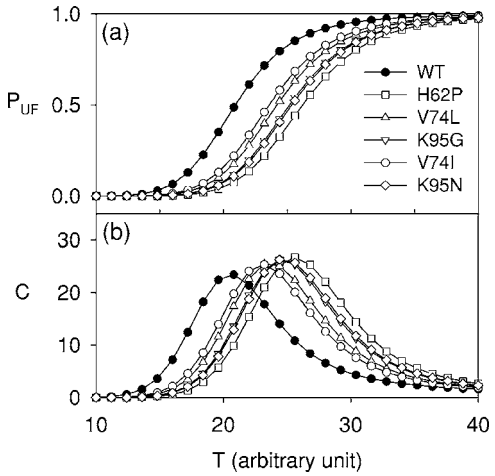

FIG. 2. The unfolded fraction (a) and the specific heat (b) of RNase H proteins as a function of temperature. It shows that the folding transition temperature of a wild RNase H is lower than those of five mutants of RNase H, agreeing with the experiment [23] that a wild RNase H is converted to the thermophilic protein by the point mutations.

$$C = \frac{1}{T^2}[\langle H^2\rangle - \langle H\rangle^2], \tag{7}$$

where $\langle\cdots\rangle$ denotes the average over the conformational space of decoys.

We calculated P_{UF} and C for wild and mutant RNase H using our global pairwise energy function. Figures 2(a) and 2(b) show P_{UF} and C from our calculation. The temperature at which C is the maximum, which is a probe for the folding transition temperature, is lower for wild RNase H than for mutant RNase H, which shows that the thermostability of mutants is enhanced, agreeing well with the experiment [23]. Our approach provides a simple and fast way to probe the thermodynamic change due to mutations in proteins and can be applied to prescreen many mutants for drug targets.

We also employed other energy functions in calculating the specific heat in order to check the effect of mutations on the changes in thermodynamic stabilities. The first energy function we used was Go-potential [24] and the second energy function was ε_{60}^{60} we calculated by the perceptron learning employing the hydrophobicity of amino acids. It showed that the folding transition temperatures of two mutants (V74I, K95N) when using Go-potential [24] and of a mutant (K95G) when using the second function were lower than that of the wild type, which did not agree with the experimental results. This test demonstrated that the results achieved from using our energy parameters, possessing the essential environmental information of amino acids, described better the thermodynamic behavior of RNase H than what can be obtained from other energy functions.

VI. SUMMARY

Within the context of employing decoys from a gapless threading in this study, we have sought the systematic improvement of a global pairwise contact energy function as we extended the parameter space of amino acids, incorporating local environments of amino acids, beyond a 20×20 matrix. We have studied the pairwise contact energy functions of ε_{20}^{20}, ε_{60}^{60}, and ε_{180}^{180} matrices according to the extent of parameter space, and compared their effect on the learnability of energy parameters in the context of a gapless threading, bearing in mind that a 20×20 pairwise contact matrix has been shown to be too simple to recognize the native folds of proteins. We showed that the construction of a global pairwise energy function was achieved using P_{train}^{1006} of a homology of less than 30%, which included all representatives of different protein classes. After parametrizing the local environments of the amino acids into nine categories depending on three secondary structures and three kinds of hydrophobicity (desolvation), the 16 290 pairwise contact energies (scores) of the amino acids could be determined by perceptron learning and protein threading. These could simultaneously recognize all the native folds of P_{train}^{1006}. When subject to a stringent threading test for P_{test}^{382}, more than 96% of these could recognize their native folds. We showed a systematic improvement of protein energy functions as we extended the parameter space of energy parameters from a 20×9 one-body matrix to 20×20, 60×60, and 180×180 pairwise contact matrices. Our work can be regarded as an intermediate step towards constructing better protein potentials starting from a simple contact potential. We hope that this work will stimulate the construction of the 60×60 or the 180×180 pairwise contact energy parameters, especially with respect to the nativelike decoys of true low energies, which then could render conclusive evidence for the success of recognizing the native folds of the 1006 training proteins simultaneously, although this was not feasible for the 20×20 pairwise contact energy parameters. We set up a simple thermodynamic framework in the conformational space of decoys to calculate the unfolded fraction and the specific heat of real proteins. The different thermodynamic stabilities of *E.coli* RNase H and its mutants were well described in our calculation, agreeing with the experiment.

ACKNOWLEDGMENTS

We acknowledge helpful discussions with Jayanth R. Banavar. This work is supported by the National Research Laboratory program of the Ministry of Science and Technology, Korea under Grant No. M1-0203-00-0029. We also thank the Proteome Supercomputing Center of Pusan National University for allowing us to use the 282 CPU pc-cluster.

[1] C. B. Anfinsen, Science **181**, 223 (1973); P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, *ibid.* **267**, 1619 (1995); D. Baker, Nature (London) **405**, 39 (2000).

[2] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Freeman, New York, 1999).

[3] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **89**, 9029 (1992); M. S. Friedrichs and P. G. Wolynes, Science **246**, 371 (1989).

[4] L. Mirny and E. Domany, Proteins: Struct., Funct., Genet. **26**, 391 (1996); M. Vendruscolo, E. Kussel, and E. Domany, Folding Des. **2**, 295 (1997); M. Vendruscolo and E. Domany, *ibid.* **3**, 329 (1998).

[5] M. Vendruscolo and E. Domany, J. Chem. Phys. **109**, 11101 (1998); M. Vendruscolo, R. Najmanovich, and E. Domany, Phys. Rev. Lett. **82**, 656 (1999); M. Vendruscolo, R. Najmanovich, and E. Domany, Proteins: Struct., Funct., Genet. **38**, 134 (2000); K. Park, M. Vendruscolo, and E. Domany, *ibid.* **40**, 237 (2000).

[6] G. Salvi and P. DeLosRios, Phys. Rev. Lett. **91**, 258102 (2003).

[7] R. I. Dima, J. R. Banavar, and A. Maritan, Protein Sci. **9**, 812 (2000); R. I. Dima *et al.*, J. Chem. Phys. **112**, 9151 (2000).

[8] U. Mayor *et al.*, Nature (London) **421**, 863 (2003).

[9] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **227**, 876 (1992).

[10] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985); J. Mol. Biol. **256**, 623 (1996).

[11] A. Godzik, A. Kolinski, and J. Skolnick, J. Mol. Biol. **227**, 227 (1992); A. Kolinski and J. Skolnick, Proteins: Struct., Funct., Genet. **18**, 338 (1994).

[12] C. Zhang and S. H. Kim, Proc. Natl. Acad. Sci. U.S.A. **97**, 2550 (2000).

[13] J. U. Bowie, R. Luthy, and D. Eisenberg, Science **253**, 164 (1991).

[14] M. Wilmanns and D. Eisenberg, Proc. Natl. Acad. Sci. U.S.A. **90**, 1379 (1993).

[15] D. Fisher, D. Rice, J. U. Bowie, and D. Eisenberg, FASEB J. **10**, 126 (1996).

[16] W. Krauth and M. Mezard, J. Phys. A **20**, L745 (1987); I. Chang, M. Cieplak, R. I. Dima, A. Maritan, and J. R. Banavar, Proc. Natl. Acad. Sci. U.S.A. **98**, 14350 (2001).

[17] U. Hobohm, M. Scharf, R. Schneider, and C. Sander, Protein Sci. **1**, 409 (1992); U. Hobohm and C. Sander, *ibid.* **3**, 522 (1994).

[18] L. Holm and C. Sander, Science **273**, 595 (1996).

[19] B. Lee and F. M. Richards, J. Mol. Biol. **55**, 379 (1971).

[20] N. Pattabiraman, K. B. Ward, and P. J. Fleming, J. Mol. Recognit. **8**, 334 (1995).

[21] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno, Phys. Rev. Lett. **80**, 5683 (1998); F. Seno, M. Vendruscolo, A. Maritan, and J. R. Banavar, *ibid.* **77**, 1901 (1996).

[22] The list of 1006, 382 proteins and the values of 16 290 energy parameters are available at http://protein.phys.pusan.ac.kr. The details can be provided upon request.

[23] K. Ishikawa, S. Kanaya, K. Morikawa, and H. Nakamura, Protein Eng. **6**, 85 (1993).

[24] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).