Amino Acids (2012) 42:2353–2361
DOI 10.1007/s00726-011-0977-0

ORIGINAL ARTICLE

Statistical energy potential: reduced representation of Dehouck–Gilis–Rooman function by selecting against decoy datasets

Wen-Wei Lu · Ri-Bo Huang · Yu-Tuo Wei · Jian-Zong Meng · Li-Qin Du · Qi-Shi Du

Received: 29 May 2010 / Accepted: 6 July 2011 / Published online: 7 August 2011

Abstract Statistical effective energy function (SEEF) is derived from the statistical analysis of the database of known protein structures. Dehouck–Gilis–Rooman (DGR) group has recently created a new generation of SEEF in which the additivity of the energy terms was manifested by decomposing the total folding free energy into a sum of lower order terms. We have tried to optimize the potential function based on their work. By using decoy datasets as screening filter, and through modification of algorithms in calculation of accessible surface area and residue–residue interaction cutoff, four new combinations of the energy terms were found to be comparable to DGR potential in performance test. Most importantly, the term number was reduced from the original 30 terms to only 5 in our results, thereby substantially decreasing the computation time while the performance was not sacrificed. Our results further proved the additivity and manipulability of the DGR original energy function, and our new combination of the energy could be used in prediction of protein structures.

W.-W. Lu · R.-B. Huang · Y.-T. Wei · J.-Z. Meng · L.-Q. Du
College of Life Science and Biotechnology, Guangxi University, 100 University Road, 530004 Nanning, Guangxi, China

R.-B. Huang (✉) · Q.-S. Du
National Engineering Research Center for Non-Food Biorefinery, Guangxi Academy of Sciences, 98 Daling Road, 530007 Nanning, Guangxi, China
e-mail: rbhuang@gxas.ac.cn

R.-B. Huang
State Key Laboratory for Bioenergy and Enzyme Technology, Guangxi Academy of Sciences, 98 Daling Road, 530007 Nanning, Guangxi, China

Keywords Decoy dataset · Performance of energy function · PyTables · Protein structure prediction · Statistical effective energy function (SEEF)

Introduction

It has been nearly four decades now since Anfinsen (Anfinsen 1973) firmly established that the three-dimensional (3D) structures of proteins were determined by the sequences of their amino acids, i.e., the latter should contain all information encoded for the protein's tertiary structures that are extremely important for their biological functions. But up to now, it is still not possible for us to accurately predict the 3D structures from the amino acid sequences. The reason may be twofold, among others. Firstly, the building blocks (20 amino acids) of proteins are deemed to be too many, which often make any prediction algorithm far too complicated to be complemented. If the types of amino acids could be reduced to only two, i.e., hydrophilic and hydrophobic, the size of sequence pace could be dramatically decreased from, say 20^{100} to 2^{33} (Dill 1999). It is the same for the conformation space of proteins: for a protein consisting of 100 amino acids, if it folds into the 3D structures randomly, the possible conformers formed will be as high as 10^{100}, in which on average the number of independent conformations per amino acid residue is about 10 (Finkelstein 1997). Random sampling of all of these conformations to find the conformer with the lowest free energy (this conformer is usually called native state of protein) would take many million years even if every sampling step would take only 1 ns (Nolting 2006). Secondly, it has not been clear which factor or factors in the sequence information will play fundamentally important roles in determination of the 3D structure. We consider that

raising these kinds of questions has a very sound base: in reality, irrespective of the translation experiments, in vivo or in vitro, proteins only take about a few seconds to fold into their correct native state (Nolting 2006). The folding pathway for the proteins taking such a short period of time strongly implies that there may only be a few simple determinants that govern the folding phenomena. Finding this sort of factors from the sequence information has been a long-pursued goal among the protein science community.

In the three-dimensional structures of proteins, each atom in the molecule can feel forces exerted on it from other atoms, and these interactions can be bonded or nonbonded. These interactions are in fact related to energies that vary depending on the atom position or configuration. Description of this kind of energy is called potential energy function, which always plays a very important role in the study of protein structure, especially in protein structure prediction from amino acid sequences. There are two types of potential energy functions in use (Boas and Harbury 2007; Lazaridis and Karplus 2000). The first one is based on the true effective energy functions, which explicitly incorporate both bonded and nonbonded terms. These terms are derived from a fundamental analysis of the forces between the particles (atoms), i.e., from the quantum calculations and from the data of thermodynamics, crystallography and spectroscopy on a wide range of systems (Jorgensen and Tirado-Rives 2005; Mackerell 2004). These will, therefore, more likely reflect the true energies. It is called physical effective energy function (PEEF) (Lazaridis and Karplus 2000). The second one is called statistical effective energy function (SEEF), because it is derived from the statistical analysis of the database of known protein structures. SEEF is obtained from the following steps: first, the probabilities are calculated such that residues appear in specific configuration (such as rotamer conformations or buried versus surface environments) or that pairs of residues appear together in a defined relative geometry. These probabilities are converted into an effective potential energy by using the relationship between the probabilities and the energies that was established by Boltzmann (Boas and Harbury 2007):

$$\Delta G = -\mathrm{RT}\ln\left(P_{\mathrm{obs}}/P_{\mathrm{exp}}\right)$$

As Harbury and coworkers emphasized (Boas and Harbury 2007), the advantage of SEEF is that it can model any behavior one can observe in the protein structure database even without knowing the physical bases of the behavior. In fact, so far the most successful protein structure predictions mostly use SEEFs in their force fields (Skolnick et al. 2003). The disadvantage for SEEF, however, is probably in following two aspects. First, SEEF is phenomenological in nature and not able to predict any new behavior not existing in the structural database from which it is derived. Second, unlike the PEEFs that can be eventually merged into a single effective, standardized one (Boas and Harbury 2008), SEEFs are laboratory dependent worldwide and different SEEFs will need time to be standardized. For SEEFs, any improvement and further perfection would surely help to make progress toward this direction.

Dehouck–Gilis–Rooman group has recently developed a novel SEEF in which the total folding free energy could be decomposed into a sum of lower order terms, i.e., the protein potentials could be decomposed into different coupling terms (42 terms in total), with each term being a function of a combination of sequence and structure descriptors, i.e., amino acid types (s_i), backbone conformations (t_i), solvent accessibility (a_i) and inter-residue distance (d_i) (Dehouck et al. 2006). This SEEF is of high accuracy in discriminating the native proteins from their decoys. It has outperformed all tested residue-based potentials, and even performed better than a couple of atom-based potentials (Dehouck et al. 2006). One of the most outstanding merits of Dehouck–Gilis–Rooman potentials is that it is formed mainly based on the interdependence of correlations among several different sequence and structure descriptors. It also allows one to evaluate the contribution of each descriptor and weight its importance after decomposition.

The purpose of this investigation is to find out if it is possible to screen some new combinations out of their original coupling terms to satisfy one's need, and if there exist some new combinations of coupling terms that will perform better than the original one. To this end, first, converting the original 42 flat files (corresponding to 42 coupling terms) in the Dehouck–Gilis–Rooman Web site into PyTables datasets needs to be done (Alted et al. 2002). These Pytables datasets turned out to be very convenient for future applications as well as current uses. Next, we also made certain modifications to the algorithms of the inter-residue distance and the solvent accessibility. After this, four initial combinations based on 17 coupling terms were obtained, and this new potential function was tested on three decoy datasets (Samudrala and Levitt 2000; Tsai et al. 2003; Das et al. 2007). These new combinations performed nearly as good as the Dehouck–Gilis–Rooman's potential. From this work, five important coupling terms were identified, as they may play dominant roles in the determination of protein structures.

Methods

Sequence and structure descriptors

The selection of protein sequence and structure descriptors was basically the same as in the previous work of the

Dehouck–Gilis–Rooman group (Dehouck et al. 2006); however, some modifications were made where necessary.

Amino acid type (s_i): 1 of 20 normal amino acids.

Backbone conformations (t_i): defined by the backbone torsion angle (ϕ, φ, ω), and grouped into seven domains called A, C, B, P, G, E and O (Rooman et al. 1991).

Solvent accessibility (a_i): defined as the ratio of the solvent accessible surface area of the considered residue to that of the extended tripetide Gly-X-Gly (Rose et al. 1985), and was grouped into five bins (Dehouck et al. 2006). The solvent accessible surface area was calculated using DSSP (Kabsch and Sander 1983). We used a polyhedron made of 1280 approximately equal triangles to approximate the sphere, the radius of which is the sum of that of the water molecular and the side-chain heavy atoms of the considered residue. When computing the accessible surface area, the integration method was used, with the centers of the triangles being the integration points and the triangle area being the weights. If a triangle is accessible to a solvent, its integrating value will be equal to 1, otherwise it will be equal to 0. We used the algorithm developed by Le Grand and Merz (Le Grand and Merz 2004) to determine if one triangle was solvent accessible or not. As a result, the surface area of this polyhedron was accurate within 0.5%.

Inter-residue spatial distance ($d_{i,j}$): the distances were computed between the side-chain centroids, normally noted C^{μ}, of two given residues and grouped into 27 bins (Dehouck et al. 2006). The C^{μ} corresponds to the geometric center of side-chain heavy atoms of a residue (Sun et al. 1992). In another original work, the Dehouck–Gilis–Rooman group used the geometric average of all side-chain heavy atoms of a given amino acid type as the side-chain center in a dataset of known structures (Kocher et al. 1994). During the selection of a new combination, we found that this kind of average distance not only increased the total free energy of both native protein and decoys, but also decreased the difference in total free energy between native protein and decoys, when compared with the direct distance calculation without such averaging.

These four descriptors ($\underline{s}_i$, t_i, a_i, $d_{i,j}$) can form different coupling terms by different combinations, with each term corresponding to one statistical potential. The Dehouck–Gilis–Rooman group created 42 coupling terms, also known as new generation of statistical potentials, and separated them into two categories: 28 local and 14 distance coupling terms (Dehouck et al. 2006).

The local potentials were as follows:

ΔW_{aa}, ΔW_{as}, ΔW_{at}, ΔW_{ts} and ΔW_{tt}; ΔW_{aaa}, ΔW_{aas}, ΔW_{aat}, ΔW_{ass}, ΔW_{ats}, ΔW_{att}, ΔW_{tss}, ΔW_{tts} and ΔW_{ttt}; ΔW_{aaaa}, ΔW_{aaas}, ΔW_{aaat}, ΔW_{aass}, ΔW_{aats}, ΔW_{aatt}, ΔW_{asss}, ΔW_{atss}, ΔW_{atts}, ΔW_{attt}, ΔW_{tsss}, ΔW_{ttss}, ΔW_{ttts} and ΔW_{tttt}.

The distance potentials were as follows:

ΔW_{ad}, ΔW_{sd} and ΔW_{td}; ΔW_{ada}, ΔW_{sds}, ΔW_{tdt}, ΔW_{asd}, ΔW_{atd} and ΔW_{tsd}; ΔW_{atsc}; $\Delta \hat{W}_{asdas}$, $\Delta \hat{W}_{atdat}$ and $\Delta \hat{W}_{tsdts}$; $\Delta \hat{W}_{atscats}$.

ΔW and $\Delta \hat{W}$ are statistical potentials of different combinations of descriptors; a_i, s_i, and t_i stand for solvent accessibility, amino acid type and back bone conformation, respectively, for amino acid i; d stands for inter-residue spatial distance. The "c" in ΔW_{atsc} and $\Delta \hat{W}_{atscats}$ is the contact potential. One can use any one or any combination of these potentials to compute protein free energy for any applications.

The PyTables dataset

PyTables (Alted et al. 2002) is a program package for managing hierarchical datasets and can efficiently and easily cope with extremely large amounts of data. It is built on top of the HDF5 library using the Python language and the NumPy package. It has a friendly object-oriented interface, with the performance-critical parts of the code generated by Pyrex compiled into efficient C language code. This feature makes it run fast. In addition, PyTables optimize memory and disk resources, so that data take much less space than other solutions such as relational or object-oriented databases.

The 42 statistical potentials are given in 42 flat text files (Dehouck et al. 2006). Some of them are too large in size: the biggest file is over 100 MByte and has at least 4,000,000 rows. This will be a performance bottleneck of data query and the main reason that we must turn all these files into a PyTables dataset using the Python language, with each file corresponding to one table in the dataset.

The decoy sets

We used three decoy sets to assess the ability of our new combinations to select the native proteins from decoy models. The first one is Decoys 'R' Us dataset (Samudrala and Levitt 2000), containing 25 native proteins, which can be separated into five classes depending on the methods used to generate them: 4state_reduced (Park and Levitt 1996): 1ctf, 1r69, 1sn3, 2cro, 4pti and 4rxn; fisa (Simons et al. 1997): 1fc2-c, 1hdd-c and 2cro; fisa_casp3 (Simons et al. 1997): 1bg8-a, 1bl0 and 1jwe; lattice-ssfit (Samudrala et al. 1999; Xia et al. 2000): 1ctf, 1dkt-a, 1fca, 1nlk, 1pgb and 1trl-a; lmds (Keasar and Levitt 2003): 1ctf, 1dtk, 1fc2-c, 1igd, 1shf-a, 2cro and 2ovo.

The second one is the Baker's group models (Tsai et al. 2003), also including 25 native proteins, each having $\sim$2,000 decoy structures generated by the ab initio protein Rosetta structure prediction method: 1a32, 1ail, 1am3,

1cc5, 1cei, 1hyp, 1flb, 1mzm, 1r69, 1utg, 1ctf, 1dol, 1orc, 1pgx, 1ptq, 1tif, 1vcc, 2fxb, 5icb, 1bq9, 1csp, 1msi, 1tuc, 1vif and 5pti.

The third one is called David Baker Models 2007, and was developed by Baker's group in 2007 for protein structure prediction for CASP7 (Das et al. 2007), including 59 native proteins, and each having 100 decoy structures generated by the Rosetta de novo structure prediction algorithm followed by all-atom refinement. Some of the decoys are "already" close to the native state: 1a19, 1a32, 1a68, 1acf, 1ail, 1aiu, 1b3a, 1bjf, 1bk2, 1bkr, 1bm8, 1bq9, 1c8c, 1c9o, 1cc8, 1cei, 1cg5, 1ctf, 1dhm, 1e6i, 1elw, 1enh, 1ew4, 1eyv, 1fkb, 1fna, 1gvp, 1hz6, 1ig5, 1iib, 1kpe, 1lis, 1lou, 1nps, 1opd, 1pgx, 1ptq, 1r69, 1rnb, 1scj, 1shf, 1ten, 1tig, 1tul, 1ubi, 1ugh, 1urn, 1utg, 1vcc, 1vie, 1vls, 1who, 2acy, 2chf, 2ci2, 2tif, 4ubp, 5cro and 256b.

The Performance measures

Three performance measures were used to evaluate this new combination. These measures were defined just as before (Dehouck et al. 2006):

1. The success rate (S_1): it is an important criterion in measuring if the energy function is good enough to distinguish between the true native proteins and their respective decoys. S_1 is defined as the percentage of proteins in which the native proteins can be correctly identified by the energy function, i.e., the free energy levels of these native proteins are the lowest among the tested dataset (the dataset is a mixture of native proteins and their decoys). For example, if we test 25 native proteins, with each protein having thousands of its own decoy structures, there are 8 native proteins that can be identified, i.e., the free energy levels of these 8 native proteins are the lowest by calculation using our energy function, then we could say that S_1 for this dataset is $8/25 = 32\%$.
2. The average Z-score (<Z>): for a native protein and its decoys in one decoy set, the Z-score is defined as:

$$Z_i = \frac{\langle \triangle W_{i\_\text{native}} \rangle - \langle \triangle W \rangle}{\sigma_W} \tag{1}$$

where $\langle \triangle W_{i\_\text{native}} \rangle$ is the free energy of native protein, and $\langle \triangle W \rangle$ is the average free energy of the decoys associated with this native protein, defined as:

$$\langle \triangle W \rangle = \frac{1}{N} \sum_{i=1}^{N} \triangle W_i \tag{2}$$

where N is the number of decoys, $\triangle W_i$ is the free energy of the decoy, and σ_W is the standard deviation of decoy free energies:

$$\sigma_W = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\triangle W_i - \langle \triangle W \rangle)^2} \tag{3}$$

Then, <Z> can be defined as:

$$\langle Z \rangle = \frac{1}{N_{\text{native}}} \sum_{i=1}^{N_{\text{native}}} Z_i \tag{4}$$

where N_{native} is the number of native proteins in a set of decoys.

3. S_{-1} is the percentage of native proteins with Z-scores lower than -1 over the total native proteins; this measure is used to evaluate the performance when the structures of native protein and decoys are very similar.

The selection of the new combination

When we tried to find a way to select a "good" combination of different statistical potential terms, the terms with closer correlation (in physical characteristics) among them was our favorite target since they seemed to perform better than the others (Dehouck et al. 2006). Fortunately, Dehouck–Gilis–Rooman (DGR) potential, featured in its additivity, allows researchers to do this kind of selection rather easily, which probably manifests the advantage in this aspect. With DGR potential, we could either add or delete the statistical potential terms freely as needed. We also noticed that the addition of even a single potential term would mean a substantial increase of the computation, and random deletion of the terms without sound reason could also lead to the loss of some important functionalities of the potentials. It is one of our main purposes in this study, therefore, to find a good or better (than the original one) combination of the terms, but also to reduce the number of terms to their minimum.

Selection of the initial combinations will normally be based on some performance measures. Here we used <Z> as the selection standard, and among potential terms in Dehouck–Gilis-Rooman's energy function, those with the lowest <Z> were selected.

If $C = \{p_1, p_2, \ldots, p_{42}\}$ represents a set of all potential terms, and p_i stands for potential terms, $i \in [1, 42]$; $C^* = \{p_1^*, p_2^*, \ldots, p_n^*\}$ represents the initial combination of the terms, and p_i^* for selected potential terms, $i \in [1, 42]$; $D = \{D_1, D_2, \ldots, D_n\}$ represents the protein decoy set for testing, and D_i stands for a native protein and its corresponding decoy set, i.e., $D_i = \{d_i^*, d_1, d_2, \ldots d_n\}$, here d_i^* is the native protein, d_k is one of its decoy, $i, k, n \in N$. Let Z represents the computed <Z> value and Z-min the current minimum of <Z>, then our algorithm for initial selection is as follows:

```
Let Z_min = 10000, C* = { }, Z = 0
For each p_i in C:
    Add p_i to C*
    Calculate Z for C* on D
    If Z < Z_min:
        Z_min = Z
    Else
        Delete p_i from C*
Output C*
```

Results

The selection of the initial combination

Based on our selection method, four initial combinations of potential terms were derived from three decoy datasets by using the algorithm described above.

1. The initial combination obtained by selecting against Decoys 'R'Us dataset is called C_RU, and the combination consists of the following ten potential terms:

   ΔW_{at}, ΔW_{ts}, ΔW_{aas}, ΔW_{ass}, ΔW_{tss}, ΔW_{tts}, ΔW_{ad}, ΔW_{sd}, ΔW_{sds}, ΔW_{atsc}

2. The initial combination obtained by selecting against David Baker Models dataset is called C_BAKER, and consisted of the following nine potential terms:

   ΔW_{at}, ΔW_{ts}, ΔW_{ass}, ΔW_{tss}, ΔW_{ad}, ΔW_{sd}, ΔW_{sds}, ΔW_{asd}, ΔW_{atsc}

3. The initial combination obtained by selecting against David Baker Models 2007 dataset is called C_BAKER2007, and consisted of the following six potential terms:

   ΔW_{at}, ΔW_{aas}, ΔW_{ass}, ΔW_{tss}, ΔW_{ad}, ΔW_{atsc}

4. The common terms in all three initial combinations above could be selected out and become another new combination called C_COMMON, which consisted of these five potential terms:

   ΔW_{at}, ΔW_{ass}, ΔW_{tss}, ΔW_{ad}, ΔW_{atsc}

In the DGR original potential, there were 42 potential terms, and in order to save time only the following 17 terms (on arbitrary basis) were used for our selection for the four different initial combinations above:

ΔW_{as}, ΔW_{at}, ΔW_{ts}, ΔW_{tt}, ΔW_{aas}, ΔW_{ass}, ΔW_{tss}, ΔW_{tts}, ΔW_{ad}, ΔW_{sd}, ΔW_{sds},
ΔW_{ada}, ΔW_{asd}, ΔW_{atd}, ΔW_{tsd}, ΔW_{atsc}, $\Delta \widehat{W}_{atdat}$

The two combinations of DGR original potential came from local and distance potentials, respectively, as follows (Dehouck et al. 2006):

$$\Delta W'_{\mathrm{LOC}} = \Delta W_{ts} + \Delta W_{tts} + \Delta W_{tss} + \Delta W_{ttts} + \Delta W_{as} + \Delta W_{aas} + \Delta W_{ass} + \Delta W_{aaas} + \Delta W_{ats} + 1/2(\Delta W_{at} + \Delta W_{aat} + \Delta W_{att} + \Delta W_{aaat} + \Delta W_{aatt} + \Delta W_{attt} + \Delta W_{tt} + \Delta W_{ttt} + \Delta W_{aaa}).$$

$$\Delta W'_{\mathrm{DIST}} = \Delta W_{sd} + \Delta W_{sds} + \Delta W_{td} + \Delta W_{tdt} + \Delta W_{ad}(\mathrm{SR}) + \Delta W_{tsd} + \Delta W_{asd}(\mathrm{SR}) + \Delta W_{atd} + \Delta W_{atsd} + \Delta \widehat{W}_{tsdts} + \Delta \widehat{W}_{asdas} + \Delta \widehat{W}_{atdat}.$$

We then tested our four initial combinations against different decoy datasets to evaluate their performances. The decoy dataset used were: Decoys 'R' Us dataset (Samudrala and Levitt 2000) and Baker's Models (Tsai et al. 2003), Baker's Models 2007 (Das et al. 2007). The principle of our test method was somewhat similar to the so-called Leave One Out Cross-Validation, or more commonly called Jackknife Test (Chou and Zhang 1995). The main idea of this kind of test is that the datasets used for training or/and selecting had to be left out, and could not be used in the testing. For instance, the initial combination C_RU was derived from Decoys 'R' Us dataset, so it could not be used to test its performance against the latter (or the test result could not be taken seriously because of bias).

First, let us evaluate the performances of the three initial combinations, i.e., C_BAKER, C_BAKER2007 and C_COMMON on the Decoys 'R'Us dataset. C_RU is not discussed here because it was derived from this dataset per se. These combinations showed nearly similar results among themselves on three performance measures [$\langle Z \rangle$, S_1 (%), S_{-1}(%)]. Compared with DGR original potential, our initial combinations showed better results in S_{-1}(%) value, but showed poorer performances in other two measures (Table 1). We reckon an overall better result may be obtained if all 42 terms in the DGR potential were used (instead of 17 terms used in this research).

Next, we will test the performances of initial combinations of C_RU, C_BAKER2007 and C_COMMON on David Baker Models dataset. C_BAKER was omitted because of the same reasons as above. There was little difference among these three initial combinations. They also showed quite similar results, compared with DGR potentials.

C_BAKER2007 was also left out when C_RU, C_BAKER, and C_COMMON were tested against David Baker Models 2007 dataset. To our surprise, these three combinations performed very well on this dataset. Since DGR original combinations were not tested against David Baker Models 2007, it is too early to judge the performances

Table 1 Performance of four initial combinations

Decoy set	Combination	<Z>	S_1 (%)	S_{-1} (%)
Decoys 'R'Us dataset	LOC	−4.16	76	92
	DIST	−4.65	80	88
	LOC + DIST	−5.25	84	88
	C_RU	−4.15	72	100
	C_BAKER	−3.54	68	96
	C_BAKER2007	−3.39	60	100
	C_COMMON	−3.29	64	100
David Baker Models	LOC	−2.06	20	88
	DIST	−2.32	28	88
	LOC + DIST	−2.65	36	92
	C_RU	−2.27	36	72
	C_BAKER	−2.69	40	88
	C_BAKER2007	−2.12	32	80
	C_COMMON	−2.23	36	80
David Baker Models 2007	C_RU	−2.78	73	95
	C_BAKER	−2.71	58	92
	C_BAKER2007	−2.78	64	92
	C_COMMON	−2.81	68	90

LOC and DIST are DGR original potential combinations (Dehouck et al. 2006); C_RU, C_BAKER, and C_BAKER2007 are our initial combinations derived from Decoys 'R'Us,David Baker Models, and David Baker Models 2007 datasets, respectively. C_COMMON is the combination obtained by selecting the common terms in three initial combinations above

of our combinations on the dataset without DGR results for comparison. It is worth emphasizing, however, that David Baker Models 2007 is a specially designed dataset, featuring highly similar structures between the native proteins and their decoys (Das et al. 2007). It may be one of the greatest challenges for any potential function to accurately distinguish between native and decoy proteins in this dataset. We were very satisfied with this result, since our initial combinations performed well on it and the results were far better than those obtained on the David Baker Models dataset (Table 1).

All combinations, including DGR original combinations, had quite poor performances on David Baker Models dataset, especially on $S1$ measure. It seems, however, that our combinations were a little better than DGR potential on $S1$ measure (Table 1).

It is interesting to note that the combination C_COMMON performed quite well on all datasets tested. C_COMMON consisted of the common terms that appeared in C_RU, C_BAKER and C_BAKER2007. It only contained five terms, but still performed at least as good as the other three. This implies that some terms in the DGR original 42 terms may be far important than the others, and this kind of importance could be a generalized one since it can be proved on three different decoy datasets, as shown in Table 1. In other words, some terms in DGR's 42 terms may be decisive in describing the native protein structures, and the rest of the terms may be just auxiliary to those dominant terms. The function of these auxiliary terms, however, could not be justified when tested on David Baker Models 2007, since all the initial combinations gave nearly the same results (Table 1).

Weighted initial combinations

For the initial combinations described above, it could be regarded as that every term had been weighted using unity 1. Here, we tried to give a different weighting to every term to evaluate the impact of weighting on the performance.

The distribution of weighting was based on random method, i.e., random numbers were selected for 3,000 times in the integer interval [0, 10]. Finally, we obtained the weights that optimized the average Z-score on the training decoy set and they are listed as follows (in non-normalization form):

$$\begin{aligned}\text{C\_RU}: & 8\times\Delta\text{Wat}, 7\times\Delta\text{Wts}, 6\times\Delta\text{Waas}, 6\times\Delta\text{Wass},\\ & 2\times\Delta\text{Wtss}, 6\times\Delta\text{Wtts}, 1\times\Delta\text{Wad},\\ & 9\times\Delta\text{Wsd}, 7\times\Delta\text{Wsds}, 5\times\Delta\text{Watsc}\end{aligned}$$

$$\begin{aligned}\text{C\_BAKER}: & 1\times\Delta W_{\text{at}}, 6\times\Delta W_{\text{ts}}, 9\times\Delta W_{\text{ass}},\\ & 6\times\Delta W_{\text{tss}}, 2\times\Delta W_{\text{ad}}, 8\times\Delta W_{\text{sd}}, 5\times\Delta W_{\text{sds}},\\ & 5\times\Delta W_{\text{asd}}, 7\times\Delta W_{\text{atsc}}\end{aligned}$$

$$\begin{aligned}\text{C\_BAKER2007}: & 2\times\Delta W_{\text{at}}, 2\times\Delta W_{\text{aas}}, 9\times\Delta W_{\text{ass}},\\ & 7\times\Delta W_{\text{tss}}, 1\times\Delta W_{\text{ad}}, 5\times\Delta W_{\text{atsc}}\end{aligned}$$

The performance of the weighted initial combinations shows that the <Z> measures were improved in small amount for all three initial combinations: <Z> increased in a range of 0.2–0.3, and the overall increase was less than 10%. Weighting caused some sort of fluctuation on all three performance measures, but the impact seemed not so significant. This may imply that after initial selection against decoy datasets, the initial combinations could have already been the "good" one.

Discussion

In this report, decoy datasets were used as filter to select the "good" terms out of DGR original 42 terms (17 terms were tested in this work). Performance measure <Z> was used as the selection standard. As a matter of fact, other performance measures such as S_1 and S_{-1} could be used as selection standard as well. The selection could be regarded as an optimizing process, since only the terms that performed well on the decoy datasets should be selected. Our

Table 2 Overall performances of our combinations after weighting

Decoy set	Combination	<Z> (not weighted)	S_1 (%)	S_{-1} (%)	<Z> (weighted)	S_1 (%)	S_{-1} (%)
Decoys 'R'Us dataset	C_RU	−4.15	72	100	−4.45	64	96
	C_BAKER	−3.54	68	96	−3.50	56	92
	C_BAKER2007	−3.39	60	100	−3.71	64	96
David Baker Models	C_RU	−2.27	36	72	−2.14	28	72
	C_BAKER	−2.69	40	88	−3.00	32	100
	C_BAKER2007	−2.12	32	80	−2.52	24	88
David Baker Models 2007	C_RU	−2.78	73	95	−2.60	69	85
	C_BAKER	−2.71	58	92	−2.72	61	85
	C_BAKER2007	−2.78	64	92	−3.06	68	86

See Table 1 for the designation of different initial combinations

initial combinations performed roughly as good as the DGR original potential, but the number of potential terms was reduced dramatically, from the original 30 terms to as few as 5. To our surprise, our combinations performed very well on the latest David Baker Models 2007 dataset that is generally regarded as a great challenge for statistical energy function testing.

There are several interesting phenomena that could be observed in this work. First, from Tables 1 and 2, it can be seen that S_{-1} tends to change in parallel with <Z>, i.e., when <Z> becomes better, S_{-1} also performs better, but this is not true for S_1.

Second, adding the weights to the potential terms could change the values of the three performance measures, but the impact seemed to be trivial, and it even caused fluctuation of these measures. This could mean that the initial selection using decoy datasets as filter probably embodied some sort of optimization, and the weighting could not change that too much. To fully appreciate the impact of the weighting, the options could be: (1) adding some new potential terms from the 42 original DGR terms; (2). exploring the whole weighting space, which may be a difficult task for the computation. Another much easier way may be the development of a new optimization algorithm that can produce optimized results even in small weighting space.

Third, the C_COMMON, which was derived from the common terms that exist in other three initial combinations performed almost as well as the other three combinations, although it contained only five potential terms. This clearly indicates that some potential terms may have dominant roles in determining the three-dimensional structures of native proteins. All other potential terms may just have auxiliary roles. This observation confirmed the earlier projection (Dehouck et al. 2006) in which the contributions of individual potential terms could be separately evaluated. Finding the dominant potential terms will have great implication in the research of protein structure prediction.

Additivity of the energetic terms derived from the decomposition of a complex energy is extremely important for macromolecule systems, because this will permit scientists to weight the contributions of the individual components and identify the most critical factors in the determination of protein structure. Although it is still not possible for one to do an absolute partitioning of catalytic contribution into independent and energetically additive components for enzyme molecules (Kraut et al. 2003), manifestation of additivity of its composing terms in SEEF will be seen as a big progress in protein statistical potential study. Since Dehouck–Gilis–Rooman's SEEF has been successfully applied to the mutation prediction (Dehouck et al. 2009), the performance was better than all existing methods tested so far and has also been used in the identification of residue–residue interactions responsible for the thermostabilization of proteins at different temperatures (Folch et al. 2010), it can be a logical extension that our new combination of potentials may be used in the same applications.

There has been a large number of SEEFs developed so far (Shen and Sali 2006); some early SEEFs have been mainly based on one or two simple knowledge-based features such as pairwise interaction between amino acids Tanaka and Scheraga 1976) and were significantly extended by including the solvent terms (Miyazawa and Jernigan 1985). Sippl (1990) and others derived distance-dependent energy functions to incorporate both short-range and long-range pairwise interactions. Since then, more and more knowledge-based features were added to the basic pairwise interaction potential to improve the predictive power of the SEEF. For example, the pairwise terms were further augmented by incorporating dihedral angles (Kocher et al. 1994), solvent accessibility and hydrogen bonding

(Nishikawa and Matsuo 1993). Singh (Singh et al. 1996) was the first to derive the potential for four-body interactions, in which C_α from four nearest-neighbor residues were considered. These so-called multi-body statistical potentials were developed by several groups (Cohen et al. 2009; Krishnamoorthy and Tropsha 2003; Masso and Vaisman 2007; Ngan et al. 2006) and the performances were all good. The only problem with these multi-body statistical potentials was the high cost in the computation compared with the original pairwise interaction potential. While high-order interaction potentials are still being pursued, recent years have also witnessed another tendency in which more and more reduced representations are being used and are highly successful in the discrimination of native protein structures from the non-native proteins (decoys) (Kolinski 2004). The reduced representations can include only one of following structural features: alpha-carbon only representation, a side-chain rotamer center of mass, an alpha-carbon and side-chain rotamer center of mass representation or other centers of interactions, e.g., C_α atoms only (Fitzgerald et al. 2007; Rotkiewicz and Skolnick 2008). One striking and highly noteworthy tendency in recent years is to view as critically important the solvent access surface area (S_{ASA}) and its related features (hydrophobicity, atom distances from protein surface, etc.) in the determination of protein structures. One of the surprising results showed that only a single structure feature, atomic burial, which can be expressed by atom distances to the molecular geometrical center, was the only information needed to define native conformations of small globular proteins (Pereira de Araujo and Onuchic 2009). This is because the amount of information needed may be in fact so small that it can be encoded in the linear sequence of amino acids; other conformational properties, such as secondary structures or even pairwise contact interactions, would arise as a consequence of crucial but sequence-independent constraints (Pereira de Araujo and Onuchic 2009). Another interesting work also showed that a single solvent exposure measure, called half-sphere exposure (HSE), which separates a residue's sphere into two half spheres (HSE-up and HSE-down), could be potentially very useful to represent solvent exposure in protein structure prediction, design and simulation (Hamelryck 2005; Paluszewski et al. 2006; Song et al. 2008). Solvent exposure measures describe to what extent a residue in a protein interacts with its surrounding solvent molecules and hence can provide important information for understanding and predicting many aspects of protein structure and function (Hamelryck 2005; Song et al. 2008).

As pointed out by Harbury's group (Boas and Harbury 2008), if a standardized energy function could be established, it will facilitate the research on the relation between protein structure and function. Knowledge-based or empirical mean force potential is and will still be the dominant energy function in prediction and simulation of protein structures in the near future. This is because the SEEF was extracted from the experiment-resolved protein structures and, since the structure of folded proteins reflects the free energy of the interaction of all their components, including all enthalpic and entropic contributions, as well as solvent effects, such potentials will surely provide an excellent shortcut toward a powerful prediction function (Arab et al. 2010). The new potential energy established by Dehouck–Gilis–Rooman has provided the best performance in protein mutation analysis and structure prediction (Dehouck et al. 2009), and most importantly, their potential energy can be decomposed into the sum of different coupling terms consisting of different combinations of sequence and structure descriptors. This made it stand out as one of the most promising SEEFs. If potential energy function like the one from Dehouck–Gilis–Rooman could be further perfected, and finally standardized, it will certainly benefit the protein science community.

Acknowledgments This work was supported by the National Underpinning Technology Project of China 2007BAD75B05, "863" Project 2007AA02Z227 and "973" Project 2009CB724703. We also thank the Dehouck–Gilis–Rooman group who supplied us their potential data by setting up a public-accessible website.

Conflit of interest We would like to declare that there is no conflict of interest among all parties involved in this study.

References

Alted F, Vilata I et al (2002–2009) PyTables: hierachical datasets in Python. http://www.pytables.org/

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230

Arab S, Sadeghi M, Eslahchi C, Pezeshk H, Sheari A (2010) A pairwise residue contact area-based mean force potential for discrimination of native protein structure. BMC Bioinformatics 11:16

Boas FE, Harbury PB (2007) Potential energy functions for protein design. Curr Opin Struct Biol 17:199–204

Boas FE, Harbury PB (2008) Design of protein–ligand binding based on the molecular-mechanics energy model. J Mol Biol 380:415–424

Chou KC, Zhang CT (1995) Review: prediction of protein structure classes. Crit Rev Biochem Mol Biol 30:275–349

Cohen M, Potapov V, Schreiber G (2009) Four distances between pairs of amino acids provide a precise description of their interaction. PLoS Comput Biol 5:e1000470

Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmstrom L, Wollacott AM, Wang C, Andre I, Baker D (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 69(Suppl 8): 118–128

Dehouck Y, Gilis D, Rooman M (2006) A new generation of statistical potentials for proteins. Biophys J 90:4010–4017

Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25:2537–2543

Dill KA (1999) Polymer principles and protein folding. Protein Sci 8:1166–1180

Finkelstein AV (1997) Protein structure: what is it possible to predict now? Curr Opin Struct Biol 7:60–71

Fitzgerald JE, Jha AK, Colubri A, Sosnick TR, Freed KF (2007) Reduced Cβ statistical potentials can outperform all-atom potentials in decoy identification. Protein Sci 16:2123

Folch B, Dehouck Y, Rooman M (2010) Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials. Biophys J 98:667–677

Hamelryck T (2005) An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins 59:38–48

Jorgensen WL, Tirado-Rives J (2005) Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. Proc Natl Acad Sci USA 102:6665–6670

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

Keasar C, Levitt M (2003) A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. J Mol Biol 329:159–174

Kocher JP, Rooman MJ, Wodak SJ (1994) Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol 235:1598–1613

Kolinski A (2004) Protein modeling and structure prediction with a reduced representation. Acta Biochim Pol 51:349–371

Kraut DA, Carroll KS, Herschlag D (2003) Challenges in enzyme mechanism and energetics. Annu Rev Biochem 72:517–571

Krishnamoorthy B, Tropsha A (2003) Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. Bioinformatics 19:1540–1548

Lazaridis T, Karplus M (2000) Effective energy functions for protein structure prediction. Curr Opin Struct Biol 10:139–145

Le Grand SM, Merz KM Jr (2004) Rapid approximation to molecular surface area via the use of Boolean logic and look-up tables. J Comput Chem 14:349–352

Mackerell AD (2004) Empirical force fields for biological macromolecules: overview and issues. J Comput Chem 25:1584–1604

Masso M, Vaisman II (2007) Accurate prediction of enzyme mutant activity based on a multibody statistical potential. Bioinformatics 23:3155–3161

Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–552

Ngan S-C, Inouye MT, Samudrala R (2006) A knowledge-based scoring function based on residue triplets for protein structure prediction. Protein Eng Des Sel 19:187–193

Nishikawa K, Matsuo Y (1993) Development of pseudoenergy potentials for assessing protein 3-D–1-D compatibility and detecting weak homologies. Protein Eng 6:811–820

Nolting B (2006) Protein folding kinetics. Biophysical methods. Springer, Berlin, pp 1–4

Paluszewski M, Hamelryck T, Winter P (2006) Reconstructing protein structure from solvent exposure using tabu search. Algorithms Mol Biol 1:20

Park B, Levitt M (1996) Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 258:367–392

Pereira de Araujo AF, Onuchic JN (2009) A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. Proc Natl Acad Sci USA 106:19001–19004

Rooman MJ, Kocher JP, Wodak SJ (1991) Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions. J Mol Biol 221:961–979

Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838

Rotkiewicz P, Skolnick J (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. J Comput Chem 29:1460–1465

Samudrala R, Levitt M (2000) Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 9:1399–1401

Samudrala R, Xia Y, Levitt M, Huang ES (1999) A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. Pac Symp Biocomput 505–516

Shen M-Y, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15:2507–2524

Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268:209–225

Singh RK, Tropsha A, Vaisman II (1996) Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol 3:213–221

Sippl MJ (1990) Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213:859–883

Skolnick J, Zhang Y, Arakaki AK et al (2003) TOUCHSTONE: a unified approach to protein structure prediction. Proteins 53(Suppl 6):469–479

Song J, Tan H, Takemoto K, Akutsu T (2008) HSEpred: predict half-sphere exposure from protein sequences. Bioinformatics 24:1489–1497

Sun S, Luo N, Ornstein RL, Rein R (1992) Protein structure prediction based on statistical potential. Biophys J 62:104–106

Tanaka S, Scheraga HA (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 9:945–950

Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D (2003) An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 53:76–87

Xia Y, Huang ES, Levitt M, Samudrala R (2000) Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 300:171–185