Quantitative Biology 2015, 3(4): 157–167
DOI 10.1007/s40484-015-0054-x

Review

On statistical energy functions for biomolecular modeling and design

Haiyan Liu*

School of Life Sciences, Hefei National Laboratory for Physical Sciences at the Microscales, and Collaborative Innovation Center of Chemistry for Life Sciences, University of Science and Technology of China, Hefei 230027, China; Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230027, China
* Correspondence: hyliu@ustc.edu.cn

Received August 12, 2015; Revised November 4, 2015; Accepted November 16, 2015

Statistical energy functions are general models about atomic or residue-level interactions in biomolecules, derived from existing experimental data. They provide quantitative foundations for structural modeling as well as for structure-based protein sequence design. Statistical energy functions can be derived computationally either based on statistical distributions or based on variational assumptions. We present overviews on the theoretical assumptions underlying the various types of approaches. Theoretical considerations underlying important pragmatic choices are discussed.

Keywords: potential of mean forces; statistical distribution; optimization; correlated variable; reference state

INTRODUCTION

In discovery as well as in engineering studies of biomolecular systems, structural modeling and structure-based design have been playing increasingly important roles [1–4]. At the foundations of these approaches are computational models developed to quantify the inter- and intra-molecular interactions between individual atoms or residues. These include physics-based models such as the various molecular mechanics force fields [5–7], and data-based models such as the so-called statistical or knowledge-based potential energy functions [8–16].

In principle, computational models of biomolecular interactions can be based exclusively on physical principles, for example, by using a full molecular mechanics model with explicit treatment of solvent. In practice, because of the high computational costs associated with such full molecular mechanics treatments, and also because of the compromised accuracy of most simplified models, using only physics-based models cannot yet allow the potential of computational approaches to be fully realized [17]. On the other hand, data-based models, namely, models derived from existing experimental data, can also provide foundations for accurate and efficient modeling and design. Some data-based models are system-specific, such as the structure-specific or protein family-specific sequence profiles. We are not to discuss this type of model here. Instead, we will focus on statistical energy functions (also called knowledge-based energy functions) that are meant to be generally applicable to different structural types or protein families. Such statistical energy functions can be used alone, or they can be used in combination with physics-based models of complementing strengths [18,19], to boost the accuracy and efficiency of both structural modeling and sequence design of biomolecules, especially proteins [16,20].

There exist extensive literature on statistical energy functions. Especially for structural modeling, there have been excellent reviews summarizing theoretical basis, implementation, benchmark as well as applications [21–28]. In the current paper, we try to analyze the topic with a compact and theoretically-oriented point of view. While some of the issues discussed here should have already been examined in existing literature, albeit to varied extents [16,29–33], previous discussions have been widely scattered at different places and often in disparate contexts. In this review, the various aspects are introduced and discussed in a related context. Besides statistical energy functions for structural modeling, we emphasize

models for sequence design, a topic less covered in the literature. The statistical energy functions for sequence design are closely related to those for structural modeling, but involve some distinct considerations. Resolving some of these issues has led us to develop the ABACUS energy function, which has been shown to complement and rival previous best methods in fixed-backbone *de novo* protein design [16], with experimentally verified design results. ABACUS is acronym for A Backbone-based Amino aCid Usage Survey. The program with examples can be downloaded from http://biocomp.ustc.edu.cn/Download.html.

STATISTICAL ENERGY FUNCTIONS FOR STRUCTURAL MODELING

The various protein structure prediction approaches (Figure 1) include constructing structural models using conformational sampling or optimization, or ranking different folds/structures contained in a set of candidate structures. They all rely on effective energy functions to model molecular interactions under given structures quantitatively.

Deriving models based on statistical distributions

The potential of mean forces picture

There are different ways to rationalize a data-based model of molecular interactions. One is through the potential-of-mean-forces concept in statistical mechanics [9,21]. Let us use the derivation of a distance-dependent, pair-wise interaction term between two types of protein atom, *a* and *b*, as an example. As training data, we have different proteins with known sequences and structures, each containing a number of atoms of types *a* and *b*. The direct interaction between *a* and *b* is supposed to contribute to the overall stability of different structural states of their containing proteins. On the other hand, the training proteins provide varied sequential and structural contexts for the interaction. For a training protein *c*, such context may be collectively noted by the coordinate $\mathbf{R}_c$. In given protein *c*, the potential of mean forces between *a* and *b* as a function of their inter-atomic distance *r* is related to the probability density of the distance in respective conformational ensemble, namely,

$$u_{ab}^{c}(r) = -k_B T \ln \rho_{ab}^{c}(r), \tag{1}$$

where k_B is the Boltzmann constant, and T is the temperature. The distance probability distribution $\rho_{ab}^{c}(r)$ is determined by the thermodynamics partition function of the molecular system,

$$\rho_{ab}^{c}(r) = \frac{1}{Z}\int \delta(r_{ab} - r)\exp^{-\frac{U(\mathbf{R_c},\ \mathbf{r_a},\ \mathbf{r_b})}{k_B T}} d\Gamma, \tag{2}$$

in which δ is the Dirac δ function, r_{ab} represents the distance variable between a and b, $U(\mathbf{R_c}, \mathbf{r_a}, \mathbf{r_b})$ represents the total energy of the system, $\int d\Gamma$ represents integration over the configurational space of the molecular system, and Z is the partition function defined by

$$Z = \int \exp^{-\beta U(\mathrm{R_c},\ \mathrm{r_a},\ \mathrm{r_b})} d\Gamma. \tag{3}$$

We assume that the overall potential of mean forces $u_{ab}^{c}(r)$ comprises two parts: a context-dependent but atom type-independent part $\tilde{u}^{c}(r)$, and a context-independent but atom-type dependent part $u_{ab}(r)$, that is

$$u_{ab}^{c}(r) = \tilde{u}^{c}(r) + u_{ab}(r). \tag{4}$$

We emphasize that the $u_{ab}(r)$ includes not only the interactions that could be assigned directly to *a* and *b* by any intermolecular interaction theory, but also the thermodynamics contributions from an "averaged" environment. Because of the averaging, this term is supposed

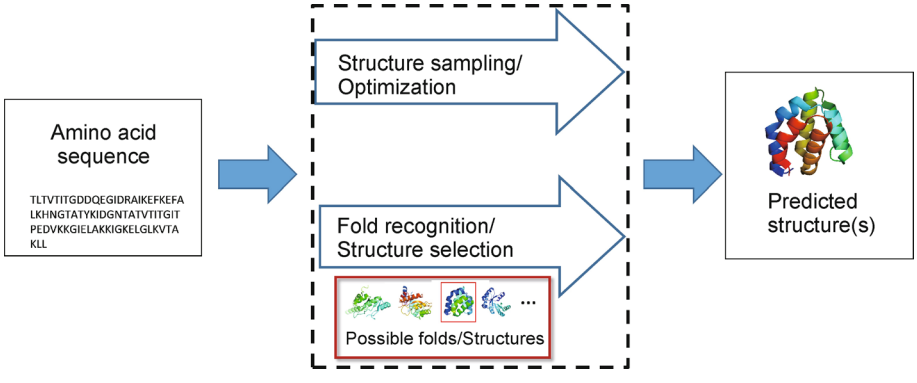

Figure 1. The basic processes of protein structure prediction. An effective (free) energy as a function of the structure is employed either to guide structure sampling, as objective for structure optimization, or to evaluate different structure candidates.

to be independent of the specific environment of a particular protein.

We further assume that the overall distribution $\rho_{ab}(r)$ as observed in all training proteins is an average (represented as a weighted summation below) over the distributions in individual proteins. This leads to

$$\rho_{ab}(r)=\sum_c w^c \rho_{ab}^c(r)=\left[\sum_c w'^c e^{-\frac{\tilde{u}^c(r)}{k_BT}}\right]e^{-\frac{u_{ab}(r)}{k_BT}}. \quad (5)$$

For simplicity, factors for weighting different training proteins and for probability normalization and so on have all been absorbed into the weights w^c for the probability distributions, or w'^c for the Boltzmann factors. Then the context-independent interaction may be formulated as

$$u_{ab}(r)=-k_BT\ln\frac{\rho_{ab}(r)}{\rho^{ref}(r)}, \quad (6)$$

with

$$\rho^{ref}(r)=-k_BT\ln\left[\sum_c w'^c e^{-\frac{\tilde{u}^c(r)}{k_BT}}\right]. \quad (7)$$

In Equations (6) and (7), the $\rho^{ref}(r)$ represents a "background" or "reference" distribution that involves an average of the context-dependent part over different training proteins. By definition, this reference distribution should be independent of the specific atom types *a* and *b*.

The above "derivation" of a context-independent, atom-type specific interaction $u_{ab}(r)$ implies a natural way of determining it from the training data (Figure 2): one needs to estimate the $\rho_{ab}(r)$ from the training data, which is straightforward, and to choose an appropriate $\rho^{ref}(r)$.

Choosing reference distributions

There have been different models for the reference distribution $\rho^{ref}(r)$. The simplest is to estimate it as the average of $\rho_{ab}(r)$ over different combinations of atom types, or formally

$$\rho^{ref}(r)=\sum_{ab} w_{ab}\rho_{ab}(r). \quad (8)$$

If the statistical interaction $u_{ab}(r)$ has been derived with a reference distribution according to Equation (8), the atom type-independent part of the interaction, i.e., $\sim u^c(r)$ in Equations (4) and (7), may not be all coming from the context of *a* and *b*. It may still contain strong residual context-independent interactions between *a* and *b*, which in Equation (8) do not necessarily averaged to zero over different atom types in the well-folded native protein structures from which the raw distributions $\rho_{ab}(r)$ have been built. The resulting overall statistical potential may suit for the comparisons between different native-like, well folded structures [11,18,34]. However, because the direct interactions between any pair *a* and *b* have not been completely included in $u_{ab}(r)$, the applicability of the resulting energy function in problems involving non-native like structures (e.g., structures encountered during *ab initio* folding) would be questionable [21]. For example, the energy function may be applied to evaluate the relative stability of different native-like structural states (e.g., those generated through threading), but it should not be used to evaluate the stability of any such native-like conformations relative to non-native like conformations (e.g., those generated with an unrestrained conformation sampling algorithm).

Given the limitation of Equation (8) which defines $\rho^{ref}(r)$ based on the training native proteins, it is desirable to define $\rho^{ref}(r)$ in other ways, so that the statistical potential can be applied to compare native like and non-native like conformations.

The distance-scaled, ideal-gas reference state (DFIRE) energy function is a model in which the definition of $\rho^{ref}(r)$ has been innovated [13]. It employs a $\rho^{ref}(r)$ that has been defined based on a reference state of non-interacting, finite-sized particles inside a spherical volume. With the physically sound definition of reference distributions, the DEFIRE energy function has found great successes in protein structure predictions.

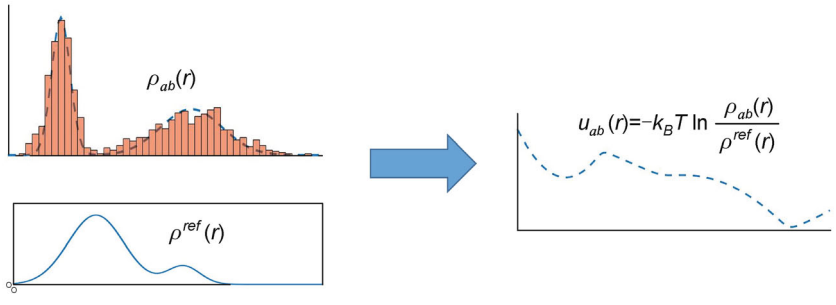

Figure 2. The distance-dependent interaction between two types of structural units (e.g., atoms or residues) *a* and *b* is determined by both a data-base-derived distribution and a unit-type independent reference distribution.

The parametric $\rho^{ref}(r)$ of DEFIRE is implicitly based on non-native like structural states. An alternative approach is to consider explicit non-native like structural states [26,32,35] when defining $\rho^{ref}(r)$. In such an approaches, ensembles of alternative conformations (called decoy conformations [35]) are usually generated with conformation sampling techniques, $\rho^{ref}(r)$ estimated from the set of sampled conformations. Based on Equations (4)–(6), the sampling should be carried out without considering the context-independent interaction $u_{ab}(r)$. On the other hand, other types of restraints (e.g., covalent chain structures and steric exclusions) must be applied during the sampling to reach any converged distribution. Thus the sampled conformation set depends on the particular restraints as well as on any other consequential set ups of the sampling procedure. So does the $\rho^{ref}(r)$ estimated from the sampled set of conformations. Then the resulting statistical interactions are justified only in scenarios in which conformations to be evaluated are generated by exactly the same sampling approach. This is in fact somewhat similar to the iterative development of coarse-grained potentials [30]. It also explains the inconsistent relative performances of using potentials constructed with different reference states to evaluate different sets of decoy protein [33].

Constructing interaction models based on optimization

To substitute the statistical-based definition of interactions in Equation (6) which needs somewhat unclearly defined reference distributions, optimization-based approaches have been proposed to derive knowledge-based interactions [29,36]. In this approach, a set of non-native training conformations are not used to estimate background distributions such as $\rho^{ref}(r)$. Instead, they are used as "decoy" structures to provide references which are supposed to be of higher energies relative to respective true native structures. The energy gap between the decoy states and respective native states are maximized with respect to parameters in the statistical energy. A common practice is to represent the energy gap by the Z-score of the native state energy with respect to the distribution of the decoy state energies, namely,

$$Z = \frac{E_{native} - \overline{E}_{decoy}}{\sigma_{E_{decoy}}}, \tag{9}$$

in which $\overline{E}_{decoy}$ and $\sigma_{E_{decoy}}$ denote the average and the root mean squared variation of the energies of the decoys, respectively. Usually a number of training proteins are considered, each associated with a native state and a number of decoy states. Then the Z-score averaged over the training proteins is optimized. For example, such a negative Z-score could be minimized with respect to the set of pair-wise statistical energy terms $\{u_{ab}^{opt}(r), a, b \in$ atom types$\}$ to obtain an optimum set of pair-wise terms, namely,

$$\{u_{ab}^{opt}(r), a, b \in \text{atom types}\} = \underset{\{u_{ab}(r),\, a,b \in \text{atom types}\}}{\operatorname{argmin}} Z[\{u_{ab}(r), a, b \in \text{atom types}\}] \tag{10}$$

Usually, stochastic optimization techniques such as Monte Carlo simulated annealing or genetic algorithm are used to solve the optimization problem.

The optimization approach toward constructing statistical energy function could also be thought of as using the following variational assumption as an explicit restraint on the energy function: the native conformational states should be of the lowest energy as compared with all non-native conformational states, and a good statistical energy function should reproduce this. Although in principle this restraint alone could be sufficient to uniquely determine the interactions, in practice such sufficiency is always compromised by approximations. The most consequential approximations include, but not limited to, the representation of the enormous number of non-native conformational states with a relatively small number of decoy conformations, and the restriction of the interaction to be of special forms so that efficient energy evaluation and parameter optimization can be executed. The eventual choices of how to implement these approximations are of critical importance for the accuracy and efficiency of the final model. With such factors generally recognized, there have been continuous and diverse research efforts investigating different ways to produce decoy structures and various functional forms with parametrization schemes for the statistical energy function [32,33,37].

Miscellaneous issues

Correlated coordinates

For transferability and for ease of estimation, a statistical energy is often defined as a summation of individual energy terms that are of simple forms. For example, the most common energy terms are one dimensional functions, each using an inter-particle distance as the variable. With such simple choice of coordinate variables, the variables for different energy terms may be significantly correlated. Simply add these energy terms together to compose a total energy function is in principle not justified. For example, because of the covalent connections between the atoms contained in one residue, the

different interatomic distances between atoms in two residues are correlated and their distributions are not independent from each other. One resulting caveat may be the following: as the number of inter-atomic distances between two residues is equal to the multiplication between the numbers of atoms contained respectively in the two residues, adding all the inter-atomic terms together would lead to artificially strong interactions between residues with large sidechains.

One possible way to resolve the inter-coordinate correlation issue is to choose new, often collective (i.e., coarse-grained) coordinates as variables for the energy function. For example, the problem of over counting correlated inter-atomic distances may be mended by using a single inter-residue statistical interaction term, although information such as relative orientation between side chains would then be left out if a simple inter-residue distance is used as the variable. An alternative is to define multi-dimensional energy terms that each depends on a number of strongly correlated coordinate variables. For example, the inter-residue interaction term can depend on both distances and orientations [32,38–40], or simultaneously on multiple inter-atomic distances. The downside is that the estimation of multi-dimensional distributions can be much more complicated and less accurate than the estimation of one-dimensional distributions.

Combining the statistical distribution-based and the optimization-based approach can somewhat ease the burdens on coordinate selection. For example, the total energy may be treated as a weighted summation of energy terms each based on a statistical distribution [18]. The weights, which are subjected to optimization, may account for some of the (unknown) extents of over counting due to correlations.

Finally, correlations between local coordinate variables (e.g., interatomic distances) with more global ones (e.g., radius of gyration) may be implicitly handled by considering explicitly sampled reference sates. For example, if a potential of mean force of the radius of gyration is to be applied in conjunction with a set of local interaction terms, either the reference distribution for the local interaction terms should be sampled with a given potential of mean force for the radius of gyration, or the reference distribution for the radius of gyration should be estimated with conformations sampled with given local interaction terms [31].

Optimizing an energy funnel

Besides the energy gap between native and decoy states, other objectives have been proposed for deriving statistical energy functions variationally [29]. In a number of studies, the supposedly funnel-shaped energy landscapes around the native states are considered as the explicit objectives for optimization [41]. To define such an objective, a structural metric is proposed to quantify the deviation of decoy states from respective native states. Then the correlation between this structural deviation and the energy difference between respective decoy and native states is maximized with respect to the energy function parameters. The structural metric is usually defined in a coarse grained manner, for example, as the fraction of native inter-residue contacts in a decoy structure. For the purpose of representing the conformational variations near the respective native states, the decoy structures may be generated through deliberate distortion of respective native structures. This can be especially effective for the development of models for protein-protein docking [42].

A statistical energy that produces a smooth funnel-shaped energy landscape is advantageous in structural modeling applications, as it may suffer less from the local minimum problem. However, the underlying hypothesis is somewhat less solid and the conformational ensemble to represent the funnel should be considered carefully to avoid introducing artifacts.

STATISTICAL ENERGY FUNCTIONS FOR STRUCTURE-BASED SEQUENCE DESIGN

In protein design (Figure 3), design goals such as folding into specific target structures or binding to specific partner proteins are often encoded into effective energies as functions of the amino acid sequence. Sequences that minimize respective energy functions are identified and proposed as design results to fulfill corresponding design goals.

Deriving models based on statistical distributions

Analogizing sequence distributions to thermodynamics distributions

While statistical interaction models for structural modeling may find, at least partially, theoretical foundations in statistical mechanics or in the thermodynamics hypothesis about protein native structures, it can be more difficult to do so for statistical models for sequence design. Here we invoke an alternative rational that is based on conditional probability distributions.

We consider the problem of finding a sequence that uniquely folds into an intended structure (we assume that the intended structure is designable, namely, there exists sequences that we are looking for). We transform the problem into one about probabilities: based on the observed sequences and structures of native proteins, can we estimate a probability function that gives the distribution of sequences associated with (or conditioned

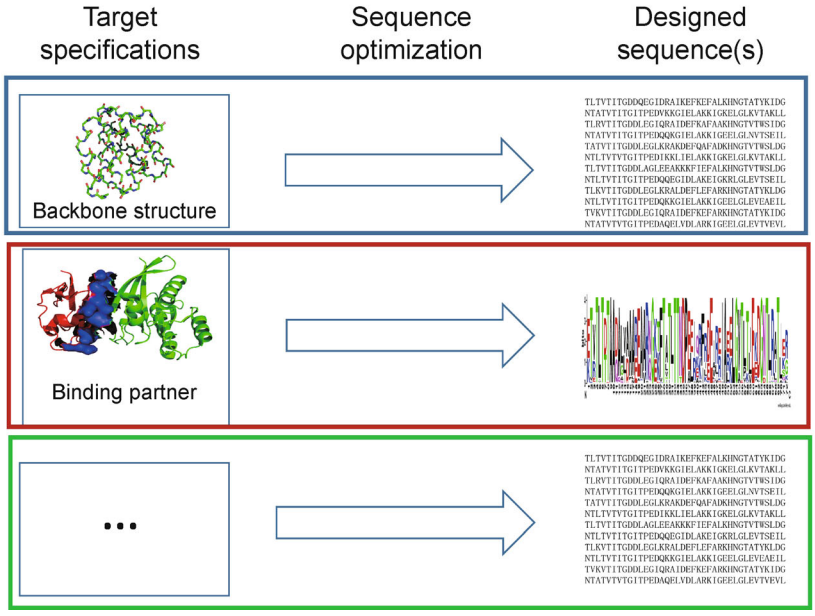

Figure 3. The basic process of protein sequence design. An effective energy as a function of the sequence needs to be optimized with respect to sequence to achieve specific design goals, such as folding into a given three-dimensional structure, or binding to a given target protein.

on) the intended structure? If we can do this, writing the respective function as $P(sequence|structure)$, we can try to find the most probable sequence by maximizing this conditional probability, namely,

$$sequence^{opt} = \underset{sequence}{\operatorname{argmax}} P(sequence|structure). \quad (11)$$

The sought probability distribution is of many variables or high dimensional: the amino acid residue type at every variable sequence position is a variable. To make the problem tractable, simplifications, which are essentially assumptions about the forms and parameters of this probability function, need to be made. We may analogize this distribution to the thermodynamics equilibrium distribution of a hypothetical physical system of a finite number of particles, each in discretely allowed states: every variable position of the protein corresponds to a particle in the analogous system, while the amino acid type at the position corresponds to the state of the particle. In addition, we may consider the different structural arrangements of the amino acid positions to be correspond to different configurations of the imaginary discrete-state particles, the energies and interactions of the particles (and thus the total energy of the system) being dependent not only on their states (amino acid types), but also on their configurations (structure). With this analogy, we may assume that the sequence distribution we are interested in is of the same form as the equilibrium state distribution of the hypothetical system, which is the following Boltzmann distribution

$$P(sequence|structure) \equiv P(a_1, a_2, ..., a_L|structure) = \frac{1}{Z} e^{-\beta E(a_1, a_2, ..., a_L|structure)}, \quad (12)$$

where L represents the total number of amino acid positions (or the length of the peptide chain), a_i with $i \in \{1, 2, ..., L\}$ represents the amino acid type at position i, β represents "temperature", E represents "energy", and the partition function

$$Z = \sum_{a_1} \sum_{a_2} \cdots \sum_{a_L} e^{-\beta E(a_1, a_2, ..., a_L|structure)}, \quad (13)$$

in which each summation is over the 20 types of amino acid residue. With the above analogy, we may formulate our assumptions on the basis of $E(a_1, a_2, ..., a_L|structure)$ instead of $P(sequence|structure)$.

Estimating the one-residue and two-residue energies

Next, we assume that the total energy can be broken into single body or one-residue terms, or the sum of single-body and two-body or two-residue terms, ignoring the remaining higher order interactions [10,16,43]. This leads

to

$$E(a_1,\ a_2,\ ...,\ a_L|structure) \approx \sum_{i=1}^{L} e_i(a_i|structure) + \sum_{i=1}^{L-1}\sum_{j=i+1}^{L} e_{ij}(a_i,\ a_j|structure). \quad (14)$$

The second, double summation term should be omitted if only the single-body is to be considered.

Now we are at a position to consider how the individual single-body and two-body energies are dependent on the structure (or configuration of the imaginary particles). It is reasonable to consider that these energies are only dependent on the local structural environments and the relative arrangements of the amino acid positions involved, i.e.,

$$e_i(a|structure) = \tilde{e}_i(a), \quad (15)$$

and

$$e_{ij}(a,\ a'|structure) = \tilde{e}_{ij}(a,\ a'). \quad (16)$$

We use $\tilde{e}_i$ to represent the one-position energy that depends on the local structural environment of position i, and $\tilde{e}_{ij}$ to represent two-position energy that depends on the local structural environments of both i and j, as well as on the relative arrangements of the two positions. We have replaced the variable notations a_i,a_j by a,a', respectively, as these variables no longer need to be differentiated by sequence position indices.

It is somewhat straightforward to derive $\tilde{e}_i$ and $\tilde{e}_{ij}$ from a set of training protein structures and sequences, except that one need to define certain representations of the "local structural environment", as well as of the "relative arrangements". Such representations are used to identify positions in the training proteins of structural arrangements (configurations) identical to position i or positions i and j. Based on the previously introduced potential of mean forces concept, the amino acid type-dependent energy terms $\tilde{e}_i$ and $\tilde{e}_{ij}$ are determined from the amino acid type distributions $\tilde{\rho}_i(a)$ and $\tilde{\rho}_{ij}(a,a')$ at these positions. Similar to Equation (6), the amino acid type variables substituting the distance and variable, we have

$$\tilde{e}_i(a) = -\frac{1}{\beta}\ln\tilde{\rho}_i(a), \quad (17)$$

and

$$\tilde{e}_{ij}(a,a') = -\frac{1}{\beta}\ln\frac{\tilde{\rho}_{ij}(a,\ a')}{\rho_{ij}^{ref}(a,\ a')}. \quad (18)$$

In Equation (18), we use $\rho_{ij}^{ref}(a,\ a') = \tilde{\rho}_i(a)\tilde{\rho}_j(a')$ to remove the contribution of one residue, local environment dependent energies in the two residue term.

Deriving models based on optimization

Just as in structural modeling where the energy gap between native and non-native structures can be used as an objective to optimize the potential, in sequence design, the energy gap between native sequences and alternative ones can also be used to optimize the potential. This approach was first adopted by RosettaDesign [20,24]. In general, we can write a sequence design energy function with a certain set of undetermined parameters $\mathbf{\Theta}$, as $E(sequence|structure, \mathbf{\Theta})$ (for example, in RosettaDesign, the $\mathbf{\Theta}$ included the relative weights of different energy terms, as well as the reference energies of different residue types). To determine $\mathbf{\Theta}$, an objective related to the energy differences between native and non-native sequences of training proteins is chosen. The energy differences should be calculated with $E(sequence|structure, \mathbf{\Theta})$. For non-native sequences, the most commonly considered are single residue changes. For each position i in a set of training protein structures, we may calculate the energies associated with each of the 20 amino acid types at that position, $\varepsilon_i(a|\mathbf{\Theta})$, and look at the probability of the native amino acid type a_i^0 according to a Boltzmann distribution based on the energies,

$$p_i(a_i^0|\mathbf{\Theta}) = \left[\sum_a e^{-\beta\varepsilon_i(a|\mathbf{\Theta})}\right]^{-1} e^{-\beta\varepsilon_i(a_i^0|\mathbf{\Theta})}. \quad (19)$$

Then the objective for optimization may be formed as

$$\Omega(\mathbf{\Theta}) = \ln L(\mathbf{\Theta}) = \ln\prod_i p_i(a_i^0|\mathbf{\Theta}) = \sum_i \ln p_i(a_i^0|\mathbf{\Theta}). \quad (20)$$

The optimization problem

$$\Theta^{opt} = \underset{\mathbf{\Theta}}{\operatorname{argmax}}\,\Omega(\mathbf{\Theta}) \quad (21)$$

may be solved with stochastic optimization techniques such as simulated annealing or genetic algorithms.

Objectives altered from Equations (19)–(20) can be used. For example, instead of the Boltzmann probability, the rank of the native residue type in the 20 residue types according to energy may be considered. Either the averaged rank of all positions or the fraction of positions at which the native type is ranked as the lowest can be considered as objectives for optimization. As these objectives are strongly correlated, optimization of any them may do the job equally well, and their resulting differences in a final energy model may no longer be critical as compared with effects of other approximations of the model.

By limiting the non-native sequences for consideration to single residue substitutions, we are assuming that the

basic form of $E(sequence|structure, \mathbf{\Theta})$ can already capture the defects of non-native sequences that deviate significantly from the native ones, and the refinement of $\mathbf{\Theta}$ is needed for the fine tuning of accuracy. In the context of sequence design, this fine tuning can be critical for accuracy because sequence design is prone to small sequence errors and there are enormous number of possibilities for such small errors.

Miscellaneous issues

Extracting structure features

In Equations (11)–(16), structural features most informative [44] about compatible sequences should be extracted from the overall structures to condition the amino acid type distributions. For conditions of one residue distributions, commonly employed local structure features include secondary structure type, solvent exposure, backbone conformation, and so on. For two-residue distributions, previously the most commonly used feature is inter-residue distance.

One issue that has often been overlooked is the correlated dependence of amino-acid preferences on different structural features. For example, given the amino acid type distribution conditioned on the secondary structure type and the distribution conditioned on the solvent exposure, their product (normalization assumed) does not necessary produce the distribution conditioned on both features. Thus a proper way is to consider the distribution jointly conditioned on both features. The joint dependence of residue pair preference on both local structural features and relative positioning in three dimensional space was the reason for us in the ABACUS model to define a two-residue term Equations (16) and (18) that depends on not only the relative positioning of the two residues, but the respective local structural environments as well [16].

Class-based and neighbor-based statistics

In many studies, statistical analysis was performed by classifying the training data according to a selected structural feature. We may call this as class-based statistics [44–46]. For example, residue positions might be classified according to their solvent exposure into exposed or buried classes (often more classes were defined). For each class of position the respective amino-acid distribution could be determined. During design a position in the intended structure would be classified in the same manner and the corresponding distribution applied. With this approach, positions with intermediate properties would be described poorly. An alternative is the neighbor-based statistics [16,47], in which neighbors of respective design targets are searched in the structural feature space. The target is always located at the center of the utilized training data. Figure 4 illustrates differences between the two approaches to deriving single residue statistics that depends on a one-dimensional structural coordinate.

The neighbor-based statistics is especially suitable for considering joint dependences on different structural

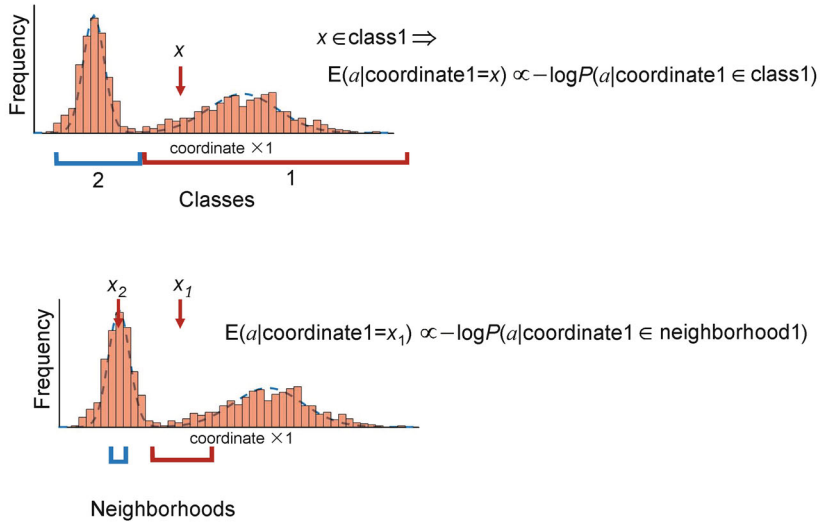

Figure 4. Example structure-dependent amino acid energies derived with class-based (top) and neighbor-based approach (bottom). This energy term as a function of residue type is assumed to be dependent on only one structural coordinate. It is difficult (easy) to extend the cluster-based (neighbor-based) approach to include simultaneous or joint dependences on multiple structural coordinates.

features [16]: defining a similarity metric in a higher-dimensional space is much easier and probably much more robust than defining classes within a higher-dimensional distribution. In addition, the similarity cutoff can be chosen adaptively based on accumulated size of the training neighbors, balancing between the size and the relevance of training samples [16]. In ABACUS, we have applied neighbor-based statistics with adaptive similarity cutoffs to find one-residue dependences on solvent exposure as well as to two-residue dependences on spatial arrangements. In the later the root mean square deviations of all backbone atoms is used as the similarity metric, treating relative orientations in a more or less natural and unbiased manner.

Positive design and negative design

If we assume the energy function for sequence design represents a true physical energy (in many current sequence design programs, at least some of the terms in the overall model are rooted in true physical energies), sequence design by minimizing this energy function is a form of "positive design". Physically, the requirements on a protein molecule to uniquely fold into a particular three-dimensional structure is that the intended structural state should be of the lowest energy as compared with any other possible structural states, or formally,

$$E(structure_{\text{intended}}|sequence) \leqslant E(structure_{any}|sequence) \tag{22}$$

The goal of sequence design is to find sequences that satisfy Equation (22). For this purpose, to consider only the intended structural state and the associated $E(structure_{\text{intended}}|sequence)$ would not be enough. We also need to know, for a candidate sequence, if there exist any other structural states that are of lower energy than the intended structural state. Thus, both positive design (namely, stabilizing the intended structural state) and negative design [47] (that is, destabilizing alternative, unintended structural states) should be considered. Most current sequence design researches has focused on positive design, how to realize negative design generally remaining elusive. Some kind of "average" negative design may have been implicated by trying to solve the following optimization problem instead of trying to find sequences satisfying Equation (22),

$$sequence^{opt} = \underset{sequence}{\operatorname{argmin}} [E(structure_{\text{intended}}|sequence) - E^{ref}(sequence)]. \tag{23}$$

Here $E^{ref}(sequence)$ depends only on the sequence and is usually a summation over the energies of individual residue, which depend solely on residue type. It could be argued that this term in Equation (23) can substitute $E(structure_{any}|sequence)$ in Equation (22) on average. In practice, the reference energy is assumed to be a sum of residue-type specific constant energies over all residues, the constant energies determined through parameter optimization.

TESTING AND VERIFICATION

There have been many established benchmark sets to assess and compare different statistical models for structural modeling. A wide range of structural modeling tasks are covered, including comparative modeling, fold recognition, *de novo* folding, as well as complex structure prediction. In such cases, comparing the computationally modeled structures with existing experimentally-determined structures would usually constitute a stringent test if the respective experimental information has been unknown or unconsidered by the modeler.

The test or assessment of sequence design models is quite different. There usually exist an enormous number of possible solutions that satisfy the design specifications (for example, there are many protein sequences that fold into an intended structure). Thus comparing designed sequences with existing ones does not tell much. Any stringent test must comprise new experiments to verify the specific design results. In other words, theoretical tests are always built on hypotheses or assumptions with known flaws. For example, sequence recovery rate is often considered as a criterion in sequence redesign tests, although we know that high recovery rate does not equal to good design. *De novo* structural modeling on designed sequences has also been used as an assessment tool [16,48]. However, current *de novo* modeling aimed at finding the right fold for an input foldable sequence, not at discriminating between foldable and unfoldable sequences. Despite these and possibly other caveats, theoretical tests can be carried out in large scales, and they are economic for comparing different models. The results of theoretical test can be meaningful if looked at statistically but not individually [16]. Thus systematic theoretical assessments should play important roles for the progressive development of biomolecular design methods.

CONCLUDING REMARKS

In last decades, enormous amount of biomolecular sequence and structure data have been accumulated through the extensive efforts of molecular biologists and structural biologists from around the globe. It has been more than 20 years since approaches to extract systematic and quantitative models about sequence-structure relationship were proposed. Data-based structural modeling

approaches have become more-or-less mature and found wide applications. On the other hand, we believe that data-based biomolecular design is still in its infancy. Compared with 20 years ago, the amount of available data have increased by at least two orders of magnitudes. In the meantime, the speed for data-accumulation still keeps increasing rapidly. Besides these, new experimental techniques can be adopted to support model development with more efficient and extensive feedbacks. For example, in ref. [16], we have used a general approach to assess and improve the foldability of designed proteins without tedious and expensive protein expression and purification. With these continuing improvements, we expect to see new developments of protein design to address many currently difficult problems, such as the design of protein fold or of novel molecular recognition interface.

ACKNOWLEDGEMENTS

This work has been supported by National Natural Science Foundation of China (Grant Nos. 31370755 and 21173203) and the Chinese Ministry of Science and Technology (Grant No. 2012AA02A704).

COMPLIANCE WITH ETHICS GUIDELINES

The author Haiyan Liu declare that he has no conflict of interests.

This article does not contain any studies with human or animal subjects performed by the author.

REFERENCES

1. Jacobson, M. and Sali, A. (2004) Comparative protein structure modeling and its applications to drug discovery. Annu. Rep. Med. Chem., 39, 259–276.
2. Skolnick, J., Zhou, H. and Gao, M. (2013) Are predicted protein structures of any value for binding site prediction and virtual ligand screening? Curr. Opin. Struct. Biol., 23, 191–197
3. DiMaio, F., Echols, N., Headd, J. J., Terwilliger, T. C., Adams, P. D. and Baker, D. (2013) Improved low-resolution crystallographic refinement with Phenix and Rosetta. Nat. Methods,10, 1102–1104
4. Koellhoffer, J. F., Higgins, C. D. and Lai, J. R. (2014) Protein engineering strategies for the development of viral vaccines and immunotherapeutics. FEBS Lett., 588, 298–307
5. Brooks, B. R., Brooks, C. L., Mackerell, A. D. Jr, Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., et al. (2009) CHARMM: the biomolecular simulation program. J. Comput. Chem., 30, 1545–1614
6. Christen, M., Hünenberger, P. H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. A., Kräutler, V., Oostenbrink, C., et al. (2005) The GROMOS software for biomolecular simulation: GROMOS05. J. Comput. Chem., 26, 1719–1751
7. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. and Kollman, P. A. (1995) A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. J. Am. Chem. Soc., 117, 5179–5197.
8. Sippl, M. J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol., 213, 859–883
9. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. and Sippl, M. J. (1990) Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J. Mol. Biol., 216, 167–180
10. Bowie, J. U., Lüthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253, 164–170
11. Lu, H. and Skolnick, J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins, 44, 223–232
12. Jiang, L., Gao, Y., Mao, F., Liu, Z. and Lai, L. (2002) Potential of mean force for protein-protein interaction studies. Proteins, 46, 190–196
13. Zhang, C., Liu, S., Zhou, H. and Zhou, Y. (2004) An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci., 13, 400–411
14. Russ, W. P. and Ranganathan, R. (2002) Knowledge-based potential functions in protein design. Curr. Opin. Struct. Biol., 12, 447–452
15. Li, Z., Yang, Y., Zhan, J., Dai, L. and Zhou, Y. (2013) Energy functions in *de novo* protein design: current challenges and future prospects. Annu. Rev. Biophys., 42, 315–335
16. Xiong, P., Wang, M., Zhou, X., Zhang, T., Zhang, J., Chen, Q. and Liu, H. (2014) Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. Nat. Commun., 5, 5330
17. Lazaridis, T. and Karplus, M. (2000) Effective energy functions for protein structure prediction. Curr. Opin. Struct. Biol., 10, 139–145
18. Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. and Baker, D. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins, 34, 82–95
19. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. and Baker, D. (2004) Protein structure prediction using Rosetta. Methods Enzymol., 383, 66–93
20. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science, 302, 1364–1368
21. Zhou, Y., Zhou, H., Zhang, C. and Liu, S. (2006) What is a desirable statistical energy function for proteins and how can it be obtained? Cell Biochem. Biophys., 46, 165–174
22. Boas, F. E. and Harbury, P. B. (2007) Potential energy functions for protein design. Curr. Opin. Struct. Biol., 17, 199–204
23. Chen, T. S. and Keating, A. E. (2012) Designing specific protein-protein interactions using computation, experimental library screening, or integrated methods. Protein Sci., 21, 949–963
24. Das, R. and Baker, D. (2008) Macromolecular modeling with rosetta. Annu. Rev. Biochem., 77, 363–382
25. Fan, H., Schneidman-Duhovny, D., Irwin, J. J., Dong, G., Shoichet, B. K. and Sali, A. (2011) Statistical potential for modeling and ranking of protein-ligand interactions. J. Chem. Inf. Model., 51, 3078–3092
26. Shen, M. Y. and Sali, A. (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci., 15, 2507–2524
27. Floudas, C. A., Fung, H. K., McAllister, S. R., Monnigmann, M. and

Rajgaria, R. (2006) Advances in protein structure prediction and *de novo* protein design: A review. Chem. Eng. Sci., 61, 966–988.
28. Moal, I. H., Moretti, R., Baker, D. and Fernández-Recio, J. (2013) Scoring functions for protein-protein interactions. Curr. Opin. Struct. Biol., 23, 862–867
29. Koretke, K. K., Luthey-Schulten, Z. and Wolynes, P. G. (1998) Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. Proc. Natl. Acad. Sci. USA, 95, 2932–2937
30. Noid, W. G. (2013) Perspective: Coarse-grained models for biomolecular systems. J. Chem. Phys., 139, 090901
31. Hamelryck, T., Borg, M., Paluszewski, M., Paulsen, J., Frellsen, J., Andreetta, C., Boomsma, W., Bottaro, S. and Ferkinghoff-Borg, J. (2010) Potentials of mean force for protein structure prediction vindicated, formalized and generalized. PLoS One, 5, e13714
32. Zhang, J. and Zhang, Y. (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS One, 5, e15386
33. Deng, H., Jia, Y., Wei, Y. and Zhang, Y. (2012) What is the best reference state for designing statistical atomic potentials in protein structure prediction? Proteins, 80, 2311–2322
34. Zhang, Y. and Skolnick, J. (2004) SPICKER: a clustering approach to identify near-native protein folds. J. Comput. Chem., 25, 865–871
35. Chuang, G. Y., Kozakov, D., Brenke, R., Comeau, S. R. and Vajda, S. (2008) DARS (Decoys As the Reference State) potentials for protein-protein docking. Biophys. J., 95, 4217–4227
36. Bastolla, U., Farwer, J., Knapp, E. W. and Vendruscolo, M. (2001) How to guarantee optimal stability for most representative structures in the Protein Data Bank. Proteins, 44, 79–96
37. Chae, M. H., Krull, F. and Knapp, E. W. (2015) Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction. Proteins, 83, 881–890
38. Wu, Y., Lu, M., Chen, M., Li, J. and Ma, J. (2007) OPUS-Ca: a knowledge-based potential function requiring only Calpha positions. Protein Sci., 16, 1449–1463
39. Kortemme, T., Morozov, A. V. and Baker, D. (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J. Mol. Biol., 326, 1239–1259
40. Zhou, H. and Skolnick, J. (2011) GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys. J., 101, 2043–2052
41. Carlsen, M., Koehl, P. and Røgen, P. (2014) On the importance of the distance measures used to train and test knowledge-based potentials for proteins. PLoS One, 9, e109335
42. Kozakov, D., Brenke, R., Landon, M. R., Comeau, S. R. and Vajda, S. (2007) Development of dars (decoys as the reference state) potentials for docking and scoring. Abstr. Pap. Am. Chem. Soc., 233, 239–239.
43. Miyazawa, S. and Jernigan, R. L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J. Mol. Biol., 256, 623–644
44. Karchin, R., Cline, M. and Karplus, K. (2004) Evaluation of local structure alphabets based on residue burial. Proteins, 55, 508–518
45. de Brevern, A. G., Valadié, H., Hazout, S. and Etchebest, C. (2002) Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Protein Sci., 11, 2871–2886
46. Li, Q., Zhou, C. and Liu, H. (2009) Fragment-based local statistical potentials derived by combining an alphabet of protein local structures with secondary structures and solvent accessibilities. Proteins, 74, 820–836
47. DeBartolo, J., Dutta, S., Reich, L. and Keating, A. E. (2012) Predictive Bcl-2 family binding models rooted in experiment or structure. J. Mol. Biol., 422, 124–144
48. Bazzoli, A., Tettamanzi, A. G. B. and Zhang, Y. (2011) Computational protein design and large-scale assessment by I-TASSER structure assembly simulations. J. Mol. Biol., 407, 764–776

