NIH Public Access
Author Manuscript
IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2014 February 10.

Published in final edited form as:
IEEE/ACM Trans Comput Biol Bioinform. 2013 ; 10(5): 1176–1187. doi:10.1109/TCBB.2013.113.

Expanded explorations into the optimization of an energy function for protein design

Yao-ming Huang
University of California, San Francisco, Department of Bioengineering and Therapeutive Sciences. San Francsicso, CA 94143-2530.

Christopher Bystroff
Rensselaer Polytechnic Institute, Troy NY 12180.

Abstract

Nature possesses a secret formula for the energy as a function of the structure of a protein. In protein design, approximations are made to both the structural representation of the molecule and to the form of the energy equation, such that the existence of a general energy function for proteins is by no means guaranteed. Here we present new insights towards the application of machine learning to the problem of finding a general energy function for protein design. Machine learning requires the definition of an objective function, which carries with it the implied definition of success in protein design. We explored four functions, consisting of two functional forms, each with two criteria for success. Optimization was carried out by a Monte Carlo search through the space of all variable parameters. Cross-validation of the optimized energy function against a test set gave significantly different results depending on the choice of objective function, pointing to relative correctness of the built-in assumptions. Novel energy cross-terms correct for the observed non-additivity of energy terms and an imbalance in the distribution of predicted amino acids. This paper expands on the work presented at ACM-BCB, Orlando FL , October 2012.

Keywords

Biology and Genetics; Physics; Chemistry; Protein design; energy function; machine learning; correlation; rotamers; dead-end elimination

1 Introduction

The problem of finding Nature's secret formula for the energy of a protein was explored at ACM-BCB in Orlando FL, October 2012. This paper expands on the work presented at that conference, with new data, expanded methods, additional analysis, and with new insights gained through conversations at that conference.

Computational protein design solves the inverse protein folding problem - predicting the sequences of lowest energy given the protein tertiary structure [1],[2],[3]. The goal of computational sequence design is to search for protein sequences that hold desired structural and functional characteristics. Several studies have shown the success of computational protein design in different applications, such as increasing protein thermal stability [4], [5], [6], producing artificial enzymes [7], [8], [9], [10], introducing novel functions to existed

Yao-ming.Huang@ucsf.edu. bystrc@rpi.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

proteins [11], [12], [13], [14], improving binding affinity of proteins [15], [16], and even generating a novel protein topology [17].

The assumption that proteins fold to the global minimum energy conformation (GMEC) [18], provides the basis for an algorithm that optimizes energy as a function of the structure. However, due to computational limitations the search requires: 1) a coarse-graining of the structure as a sequence of discrete side chain rotamers [19], [20], [21], 2) an assumption of minimal backbone conformational flexibility, where a fixed backbone or a feasible set of backbones are used [22], [23], [24], 3) an approximation of the energy model as a pairwise decomposition), and 4) an efficient algorithm for searching the configurational space of rotamers and backbones, such as dead-end elimination [25], [26], [27], [28], [29], [30], Monte Carlo [31], [32], [33], genetic algorithm [5], [34] and branch-and-bound algorithm [35], [36], [37], and algorithms for sampling backbone conformations [38], [39], [40], [41]. These assumptions strike a compromise between speed and accuracy. The accuracy is adequate to the extent that the GMEC is still the native configuration.

The energy model generally includes van der Waals (VDW) interactions, electrostatic interactions, hydrogen bonding and solvation, each of which addresses a distinct behavior of the system. Although it is similar to the one used in molecular dynamics in many respects, the energy model used in computational protein design has unique features. The bonded energy terms often are ignored since fixed backbone atoms and rotamer libraries are used. The optimal relative contributions of individual energy terms may differ from those used in molecular dynamics because of the discretization of the side chains and the limited flexibility of the backbone in protein design. Successful discretization requires that the calculated energy of a rotameric state represent the lowest energy of all nearby, slightly rotated and shifted states. The most successful designs predict their corresponding crystal structure within 1.0 to 1.5Å RMSD [17], or approximately the length of one chemical bond.

Most search algorithms used in protein design require that the energy is calculated as a sum of pairwise terms, but this two-body decomposition hampers the accurate modeling of at least two important energetic components: hydrogen bonding and solvation energies, both better expressed as multi-body terms. Additional terms may be used to correct perceived weaknesses in the model. For example, void space is disfavored in protein structures [42] but resists modeling as a two-body equation. To penalize voids in the context of a fixed backbone, an energy term is added which is inversely proportional to the volume of the side chains in buried residues, forcing the selection of larger side chains, resulting in a tighter side chain packing in the core. Also, using the knowledge that H-bond donor or acceptor atoms are rarely unpaired in the buried state, we can penalize pairwise interactions of donor/acceptor atoms to non-donor/nonacceptor atoms. Pruning conformations with unpaired H-bond donor/acceptor atoms was previously found to be crucial for success in designing the binding specificity of periplasmic binding proteins [12].

Often the energy mode assumes that energy terms are independent, and the total energy is simply the linear sum of the individual terms, where weights are used to define the relative contribution of terms. However, it is likely that energy terms in the model may actually correlate with each other to some extent because they describe similar interactions, leading to over-counting or under-counting depending on whether the correlation is positive or negative. For example, clearly the volume-based term will be correlated to the VDW interactions, since both are roughly proportional to the atom density. Also, hydrogen bonding between donor and acceptor atoms occurs at VDW interaction distances, implying a correlation or anti-correlation between these two terms. A linear sum of weighted terms cannot capture these covariances, leading to an inaccurate energy [43].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Because the relative contributions statistical interactions of energy terms are not known *a priori*, they are determined by optimization of an objective function that is tied to a training set of experimental data. The validity of the optimized energy model is measured using statistical cross-validation by calculating the objective function on an independent test set of experimental data. Here the choice of objective function plays a critical role in the successful generalization of the model. If the assumptions built into the objective function are true and good, and the experimental data is representative, then the optimized energy model will be more general -- in our view, closer to the function used by Nature.

In this paper, we explore the validity of four objective functions and three linear and non-linear energy functions for protein design, using statistical cross-validation. The objective functions measure the success in prediction of either the amino acid sequence or the rotamer structure, using either the total log likelihood or the sum of probabilities. The results are a generally applicable energy function that accurately predicts sequence and structure in known proteins, and some insights into the best assumptions for constructing an objective function for energy function optimization.

2 Methods

2.1 The rotamer library

The Richardson backbone-independent rotamer library [21] is used in this study with some modifications. Polar hydrogen atoms of each rotamer are added for modeling electrostatic interactions. Two types of dummy atoms representing ideal positions for H-bond acceptor and donor atoms were added for modeling of hydrogen bonding interactions. The dummy atoms (DUM) for H-bond acceptors are positioned along the axis of the bond to a polar hydrogen, at a distance of 2.8Å from the donor atom. The dummy atoms (LP) for H-bond donors are placed at a distance of 2.8Å from the H-bond acceptor atom along the axis of the lone-pair orbitals, and separated by 120°, expect for ND1 on His and OH on Tyr which only have one lone pair dummy atom. Extra rotamers for LP and DUM atoms were generated by 1) assigning χ^2 of Ser, Thr and Cys, and χ^3 of Tyr to -60, 60 and 180 degrees, 2) defining three different protonated states of histidine, and 3) flipping χ^2 of Asn and His, and χ^3 of Gln by 180° to include missing rotamers in the original rotamer library [44]. The total number of rotamers is thereby increased to 296.

2.2 The training and testing protein sets

The training and testing proteins are chosen from a curated non-redundant data set of 100 high resolution protein structures [45]. Only proteins with single chains and with no missing sidechain/backbone atoms in the middle of chains are kept. A final set of 80 proteins was selected (Table 1) that covers a variety of tertiary structure types. The overall resolutions of selected proteins range from 0.8 to 1.7Å. 40 proteins of the list are used to derive the weights of the energy function, and the remaining 40 proteins are used for the evaluation test. Residues in N- or C-termini that do not have coordinates are ignored both for training and testing. When a protein structure has multiple models, only the first model was used. Only the first conformation is considered if there are alternative conformations. Polar hydrogen atoms and dummy atoms are added to all proteins as described above. In order to assign rotamer intrinsic energies to native sidechain conformations, we used the rotamer with the minimum RMSD to the native conformation.

2.3 The energy function

A physical-based, pairwise energy function is developed to score and evaluate sidechain conformations, examine the effect of the objective functions in the parameter optimization, investigate the correlations of energy terms, and study the performance in reproducing the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

trend of experimental data. Atom radii, well-depths, surface tensions and atomic partial charges of GROMOS96 force field (ffG43a1) [46] are used in the energy model. Pairwise van der Waals radii between unbonded atoms are derived from 900 highest resolution crystal structures from PDBselect-25 April 2007 list [47] or simply adapted from the GROMOS force field. The energy function (1) is a linear combination of weighted non-bonded energy terms including:

1. van der Waals interactions (E_{Rep} and E_{Att})
2. electrostatic interactions (E_{Elect})
3. hydrogen bonding energies (E_{Hb} and E_{NonHb})
4. solvation energies (E_{Solv1} and E_{Solv2})
5. rotamer intrinsic energies (E_{Rot})
6. void space energies (E_{Vol})
7. amino acid reference energies (E_{Ref}).

Scale factors (w) are applied to each energy term, and these are determined through the parameter optimization.

$$E_{total}=w_{Rep}E_{Rep}+w_{Att}E_{Att}+w_{Elect}E_{Elect}+w_{Hb}E_{Hb}+w_{NonHb}E_{NonHb}+w_{Solv1}E_{Solv1}+w_{Solv2}E_{Solv2}+w_{Rot}E_{Rot}+w_{Vol}E_{Vol} \quad (1)$$

2.3.1 Van der Waals interaction—The van der Waals energy includes the repulsive interaction (E_{Rep}) and the attractive interaction (E_{Att}) of the 12-6 Lennard-Jones potential. Two models of van der Waals interaction are evaluated in this study. In the first model, pairwise van der Waals radii between unbonded atoms are derived from the 900 highest resolution crystal structures from PDB-select-25 April 2007 list. The energies are as follows:

$$E_{Rep}=D_0\left(\frac{\alpha R_0}{R_{ij}}\right)^{12} \quad (2)$$

$$E_{Att}=-D_0\left(\frac{\alpha R_0}{R_{ij}}\right)^{6} \quad (3)$$

where R_{ij} is the distance between atoms i and j, R_0 is the equilibrium pairwise van der Waals radii, and D_0 is the well-depth that describe the strength of interactions. The D_0 parameter is left out in the van der Waals energy calculation, as it is included in the weights. To compensate for the discrete nature of rotamers and the restrictive effect of fixed backbones, which can result in correct sequences being rejected due to minor collisions that could be relieved by small adjustments of rotamers and/or backbones, the van der Waal radii are scaled by α=0.9 [52]. This provides some extra space in which to move the sidechain.

In the second model, the 12th power van der Waals repulsion term is softened by a switching to a linear form at short distances where $R_{ij} < 0.9R_0$:

$$E_{Rep}=10.0\left(1-\frac{R_{ij}}{0.9R_0}\right)D_0 \quad (4)$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

To avoid the special cases of two-bond and three-bond interactions, E_{Rep} and E_{Att} are not evaluated between atoms in rotamer *r* at residue *s* or between atoms of rotamer *r* at residue *s* and the local backbone atoms of residue *s-1*, s and *s+1*. E_{Rep} and E_{Att} are set to zero if the distance R_{ij} is larger than a cut-off value of 8Å. Hydrogen and dummy atoms are excluded from the van der Waals energy calculation.

2.3.2 Electrostatic interaction—The electrostatic interactions between two atoms are calculated with a linear distance-dependent dielectric constant [53]:

$$E_{Elect}=f\left(\frac{Q_iQ_j}{dR_{ij}}\right) \quad (5)$$

where Q_i and Q_j are partial charges of an interaction pair of atom *i* and *j*, *f* is the well-depth defined in GROMOS96 force field, R_{ij} is the distance between two atoms, and $d=40R_{ij}$ is a distance-dependent dielectric constant. The simple implementation of the distance-dependent dielectric constant not only models environmental changes from the center of protein core to the surface, but also provides a convenient, inexpensive and fairly crude approximation to the solvent effect.

2.3.3 Hydrogen bonding—A simple energy function of hydrogen bonding is used in this study to avoid unfavorable hydrogen bonding geometry in the structure and to prevent the burial of H-bond donors. Two kinds of hydrogen bonding energy are included in this model to 1) favor the formation of H-bonds (E_{Hb}) and to 2) penalize unsatisfied H-bond donors and acceptors (E_{NonHb}). First, the ideal H-bond geometric templates for H-bond donor atoms of backbones and rotamers are created to place dummy atoms, which represent the spacial positions that must be occupied by the H-bond acceptor and donor atoms. Dummy atoms are placed for all polar hydrogen atoms and H-bond acceptor atoms, which mimic the optimal locations of H-bond acceptor and donor atoms respectively. E_{Hb} and E_{NonHb} can then be modeled for two atoms *i* and *j* as a normal distribution in distance:

$$E_{Hb}=-\kappa\exp\left(-\left(\frac{R_{ij}}{\rho}\right)^2\right);E_{NonHb}=0, \quad (6)$$

if atom *i* is a dummy atom and atom *j* is a H-bond acceptor or donor atom, or vice versa,

$$E_{Hb}=0;E_{NonHb}=\kappa'\exp\left(-\left(\frac{R_{ij}}{\rho'}\right)^2\right), \quad (7)$$

if atom *i* is a dummy atom and atom *j* is not a H-bond acceptor or donor atom, or vice versa, and

$$E_{Hb}=0;E_{NonHb}=0, \quad (8)$$

if both atoms *i* and *j* are dummy atoms, where κ is the estimated optimal energy of forming a H-bond, ρ and ρ' are the Gaussian decay parameters, predetermined as 0.1 and 0.25 respectively, and R_{ij} is the distance between two atoms. Atoms that are covalently bonded to H-bond acceptor atoms are excluded in the E_{NonHb} calculation to avoid over-penalizing unsatisfied H-bond pairs resulting from the formation of H-bonds. The parameters κ and κ' are included in weights and determined by the parameter search.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

2.3.4 Solvation energy—The desolvation of hydrophobic sidechains is the main driving force of protein folding, and implicit solvent models are often used in computational protein design to address the effect. The simplest implicit solvent model is used in this study to achieve a fast access of the solvation energy [54], [55], where it is expressed as the free energy of proteins in solvent (ΔG) assuming the energy to be proportional to buried solvent accessible surface areas (so-called SAS) of atoms in the protein structures [56]:

$$\Delta G = \sum_{all\ residues} \sum_{\substack{i\ \in\ all\ atoms \\ of\ a\ residue}} A_i \sigma_i \quad (9)$$

where A_i is the solvent accessible surface area of atom *i* and σ_i is the atomic surface tension derived from the force field. To remain consistent with the search algorithms used in computational design, a pairwise decomposition of the solvation potential [54], [57] is implemented with two components, E_{solv1} for rotamer/backbone interactions and E_{solv2} for rotamer/rotamer interactions.

$$E_{Salv1} = \sum_i -(\Delta G_{i,local} - \Delta G_{i,bb}) \quad (10)$$

$$E_{Solv2} = s\sum_{i<j} -(\Delta G_{i,bb} + \Delta G_{j,bb} + \Delta G_{ij,bb}), \quad (11)$$

where $\Delta G_{i,local}$ is the solvation free energy derived from the SAS of the residue *i* in the presence of the local tripeptide (residue *i-1*, *i* , *i+1*), $\Delta G_{i,bb}$ and $\Delta G_{j,bb}$ are the solvation free energies derived respectively from the SAS of the residue *i* and *j* in the presence of the backbone, $\Delta G_{i,local}$-$\Delta G_{i,bb}$ is the solvation free energy derived by the SAS of the residue *i* buried by parts of the backbone excluding the local tripeptide, and $\Delta G_{ij,bb}$ is the solvation free energy derived from SAS where residue *i* and *j* are coexisted in the presence of the backbone. The over-counting of the surface area resulting from pairwise calculations is solved by a predetermined scale factor of *s* = 0.67 [54]. Solvent accessible surface areas are obtained by scaling up contact surface areas calculated from MASKER [58] and the scale factor is included in weights and are determined by the parameter search. This model reflects the dependence of exposed residues/atoms on the hydrophobicity of proteins.

2.3.5 Rotamer intrinsic energy—The intrinsic energies of rotamers given an amino acid (also called rotamer self-energy) are represented by

$$E_{rot} = -\ln(f_{observed,r}/f_{\exp ected,r}), \quad (12)$$

where $f_{observed,r}$ is the frequency of a rotamer *r* given a particular amino acid, which is observed from the statistical analysis of the protein rotamer preference and is calculated as the number of a particular rotamer divided by the number of total rotamers from the amino acid. $f_{expected,r}$ is the expected frequency of a rotamer *r* given a particular amino acid and is calculated as one divided by the number of rotamers given the amino acid.

2.3.6 Void space energy—The void space energies are used to model the effectiveness of sidechain packing and to favor the conformations that have tighter packing. The energies are reversely proportional to the volume of rotamers and only applied for buried sidechains:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$E_{Vol} = -V_r \quad (13)$$

where V_r is the volume of the buried rotamer r. Rotamer volumes were precalculated.

2.3.7 Amino acid reference energy—The amino acid reference energies are sequence specific energies that implicitly model changes in the stability of the unfolded state of the protein with mutation:

$$E_{Ref} = \sum_i ref_{aa} \times N_{aa} \quad (14)$$

where ref_{aa} is an amino acid specific energy and N_{aa} is the count of the amino acid.

2.4 Parameterization of the energy function

The weights addressing relative contributions of energy terms are optimized using defined objective functions. Several objective functions have been examined in this study, where the weights are optimized to maximize the Boltzmann probabilities of the native residues (*sequence recovery*) or the native side chain conformations (*structure recovery*) at each residue of the training proteins. This data-driven optimization of parameters assumes that the native selection (either the native side chain or the native side chain conformation) should be the lowest energy selection.

Any side chain that has a B factor larger than 60 is excluded in the training process, leaving 7191 residues (or 6523 residues when not including Gly) for training. The weights w of the energy function (1) are determined by maximizing the objective functions (15) and (16) over all residues in the 40 training set proteins as a function of w. The objective functions sum the probabilities, or log-probabilities, of the native residues, or the native rotamer, for each residue in the protein training set, and are formulated as

$$\sum_{t \in proteins} \sum_{i \in residues} \frac{S}{\sum_{r \in rotamers} \exp\left(-E^r_{t,i}\right)}, \text{or} \quad (15)$$

$$\sum_{i \in proteins} \sum_{i \in residues} \ln\left(\frac{S}{\sum_{r \in rotamers} \exp\left(-E^r_{t,i}\right)}\right), \quad (16)$$

where the denominator of the distribution functions is the sum of the exponential to the power of negative total energy E for all rotamers r over all 20 amino acids at residue i of protein t, and the numerator S (*structure oriented*) is

$$\exp\left(-E^R_{t,i}\right), \quad (17)$$

where R is the native sidechain conformation at residue i of protein t in structure recovery. The numerator S (*sequence oriented*) is

$$\sum_{R'} \exp\left(-E^{R'}_{t,i}\right), \quad (18)$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where *R*' are all rotamers of the native residue at residue *i* of protein *t* in sequence recovery [18]. Two objective functions have different operations: one (15) is the sum of the probabilities and the other one (16) is the sum of the log of the probabilities. In the training process, only one rotamer of a residue is changed at a time and the total energy is calculated in the context of the structure where all other residues are kept in their native conformations. Only the neighboring sidechains need to be considered in energy calculations. Any sidechain that has one sidechain atom located within 5Å radius of any sidechain atom of the residue *i* and has a B factor larger than 60 is defined as a neighboring sidechain of *i*.

The parameter search is performed using the replica-exchange Monte Carlo simulation [48]. Briefly, it assumes a system consisting of multiple replicas (R_1, R_2,…, R_x) of a potential solution to the optimized weights. Each replica has an associated and unique temperature, and replicas are arranged in the order of temperatures from high to low ($T_1 > T_2 >\ldots> T_x$). Starting from an initial set of random weights, each replica then in parallel performs a simple Monte Carlo search of 1,000 substitution steps. A substitution is performed by randomly selecting one weight and randomly perturbing it. The probability of accepting the perturbed weight is governed by the Metropolis criterion, based on the change in the objective function. After 1,000 steps, current values of the objective function are compared between replica R_x and its neighboring replica R_{x-1} that has a higher temperature, and an exchange of weights is made if the higher temperature replica has the higher objective function or if the difference in objective function satisfies a similar Metropolis criterion. The exchanged weights are then served as a new set of initial weights and the iteration continues until the weights from the lowest-temperature replica R_1 are converged.

The parameter search was performed using SUR Blue Gene/L cluster of RPI (www.scorec.rpi.edu/wiki/SUR_Blue_Gene). 1024 replicas are used in the training and the temperatures of replicas range from 10^6 to 10^{-7}. A temperature annealing factor of 0.97 is used to build the profile of the temperature gradient among replicas. The replica-exchange Monte Carlo simulation algorithm is coded in fortran90 using message passing interface (MPI). The optimization process was repeated independently three times to confirm convergence.

2.5 Evaluation of the energy function

To evaluate the energy function derived from the parameter search, prediction accuracy of native side chain conformations on fixed backbones was analyzed. For each residue of the test set proteins, the prediction is performed by identifying the lowest energy rotamer through the calculation of total energy from derived energy function in the context of neighboring side chains. The neighboring side chains are assigned as described above. The rest of the structure is kept in the native conformation for each prediction, and only the neighboring side chains of the predicting residue are considered in the rotamer-rotamer and rotamer-backbone energy calculation. The accuracy of the prediction was assessed by asking 1) if the lowest energy rotamer also is the native residue, 2) if the lowest energy rotamer also is the native rotamer, or 3) if the native rotamer conformation is within the top 1.5, 3.0 or 6.0% of the lowest energy rotamers.

3 Results

Reproducing native-like protein sequences given desired backbone structures is a good test for a protein design algorithm. Here we have tested four ways of formulating this definition of success in protein design, validating optimized energy functions using the accuracy in predicting the native sequence or side chain structure on a test set.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

3.1 Qualitative analysis of the objective functions

Objective functions guide the parameter optimization process and define the criteria of success. We don't have access to true energies in general, since very few of the mutations used in this process have actually been made and evaluated [49]. Instead we rely on an assumption that is generally accepted as more often right than wrong, that the native amino acid at each position has a relatively low energy as compared to all other amino acids, in the context of its local structural environment. This is likely to be a good assumption, since during the process of evolution side chains have been sampled and accepted or rejected based on the stability of protein structure and the survival of function.

Here, we inspect two ways of formulating the objective function, (15) and (16), and two ways of targeting it, either structure oriented (17) or sequence oriented (18). Each objective function was used to optimize the parameters of the energy model to convergence. The resulting energy models were validated using side chain prediction on a test set. Note that Gly residues were not included in training, and weights were forced to sum to one during the training.

3.1.1 Rotamer probabilities.—Energy terms may be converted to probabilities using the Boltzmann distribution equation. The partition function in the denominator of the equation (19) is the sum of frequencies over all rotamers at residue *i* of protein *t*, while the numerator *S* is the sum of frequencies of the rotamers having native amino acids at that residue.

$$\frac{S}{\sum_{r \in rotamers} \exp\left(-E_{t,i}^{r}\right)} \quad (19)$$

The objective function for all residues over all proteins in the training set may be expressed by: 1) summing the probabilities over all residues, ΣP (15), or by 2) summing the log of the probabilities over all residues, ΣlnP (16). Both functions reach a maximum when all of the native side chains are lowest in energy, but they differ in the relative contributions of high-energy terms.

As shown in Table 3, the performance of the resulting energy function is slightly better when the weights were trained using ΣlnP objective function, indicating that the ΣlnP objective function gives an energy model that better separates native rotamer conformations from non-native conformations.

The rationale for this result becomes clear when considering the relative effect on the objective function of a high-energy native rotamer, as would occur more often if the weights deviated more from optimal. A high energy produces a near-zero Boltzmann probability *S*, contributing little to the ΣP objective function, but produces a large negative lnP, making the ΣlnP objective function more sensitive to errors.

3.1.2 Sequence recovery or structure recovery—The target for optimization, the numerator in the objective function, may be defined to be the native amino acid (18), or the native rotamer of the native amino acid (17). Sequence recovery (18) trains the energy function to reproduce the native sequence, allowing a native side chain with a non-native rotamer. Structure recovery (17) trains the energy function to discriminate between the native and all other rotamers.

Although the meaning of the objective function has been altered, both the optimized weights (Table 2) and the accuracy of sidechain prediction (Table 3) show only minor differences,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

indicating that when the native amino acid is selected it is generally also the native rotamer. Structure recovery is a more powerful training strategy by a small margin when the ΣlnP functional form is used, again because high-energy native rotamers add a greater penalty.

Because the aim of computational protein design is to find lowest energy sequences and finding sequences fits better to the downstream process of creating molecules, we decided to use objective functions that targets sequence recovery (18) and works in log space (16) for the following studies. Excluding Gly from the training resulted in false positive and false negative predictions of glycines and glycine positions (data not shown), suggesting that all possibilities must be considered during training so as to have a general model. Gly is included in the following studies.

3.2 Dependency between energy terms

The energy model often consists of physics-based pairwise energy terms that are combined linearly (1), assuming additivity. However, it is known that different energy terms are correlated. This could result from the physics behind energy terms or simply from artificial effects of correction terms in the equation. The correlations and anti-correlations between terms implies an over- and under-counting in energy calculation, making a balance between energy terms hard to achieve. And it can never be achieved if a linear combination of correlated energy terms is used without some type of correction.

3.2.1 Correlations of energies.—The individual unweighted energy terms were calculated for all 7191 residues in the training protein set, using the neighboring side chains as described above. All the side chains were kept in native conformations when calculating energies. Correlation coefficients (Table 4) between any two terms were calculated and only those that have strong correlations (>0.5 or <−0.5) are further discussed. Among those with strong correlations, the void space energy (E_{Vol}) used to model tighter packing has the highest number of correlations to other energy terms.

Some of the correlations are easy to explain. For example, higher VDW repulsion energy (E_{Rep}) often occurs where the buried side chain volume is large, translating to a lower energy E_{Vol}, and giving a negative correlation between E_{Rep}-E_{Vol}. This also explains the positive correlation between VDW attraction (E_{Att}) and E_{Vol}. Larger side chains tend to block the dummy atoms (DUM or LP) from H-bond donor or acceptor atoms in hydrogen bonding interactions, and result in unfavorable E_{NonHb} energies, giving a negative correlation between E_{NonHb}-E_{Vol}. The positive correlation between E_{Rep}-E_{NonHb} can also be understood considering the explanations for E_{Rep}-E_{Vol} and E_{NonHb}-E_{Vol}. The positive correlation between E_{Elect} and E_{Hb} agrees with the fact that partial charges on H-bond donor and acceptor atoms also favor the charge-charge interactions. Moreover, forming hydrogen bonds should compensate for the increased VDW interactions, which rationalizes the relationships of E_{Rep}-E_{Hb} and E_{Att}-E_{NonHb}. It is expected that both E_{Rep}-E_{Att} and E_{Hb}-E_{NonHb} have negative correlations since these are pairs of opposite energies. It is not obvious how the correlations of solvation energies (E_{Solv1}, E_{Solv2}) to other terms can be elucidated, although the correlation analysis captures these.

3.2.2 Cross terms—Seeing that some energy terms heavily depend on others, it poses a question about the correctness of a linear combination of energies. Thermodynamic additivity is a fundamental principle of chemical systems, and to imply that the true energy function should include terms where energy is multiplied by energy seems to be a violation of this principle. But when we consider that each term is simply a function of the atoms and their coordinates, multiplying terms is a way of creating multi-body energy terms. By defining the functional form, we are asking the machine learning protocol to find new

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

multibody terms that correct for the non-additivity of the simple terms. If we did not consider this cross talk between energy terms, some energies may be over-counted and others under-estimated.

Here, we introduce cross terms for each of the pairs of energetic terms that are observed to be highly correlated or anti-correlated. Given two correlated energy term E_a and E_b, the cross term is expressed as:

$$w_{ab} \times E_a \times E_b \quad (20)$$

where w_{ab} is the cross weight for the cross term, which can be determined by machine learning. Since the energy used in computational protein design is further decomposed into the energy of backbone-rotamer (self energy, E^{self}) and the energy of rotamer-rotamer (pairwise energy, E^{pair}), the cross term can be further expressed either as *separate*,

$$w_{ab} \times \left(E_a^{pair} \times E_b^{pair}\right) + w_{ab} \times \left(E_a^{self} \times E_b^{self}\right), \quad (21)$$

or as *combined*,

$$w_{ab} \times \left(E_a^{pair} \times E_a^{self}\right) \times \left(E_b^{pair} \times E_b^{self}\right), \quad (22)$$

terms, where the cross term is applied either to the pairwise energy and the self energy separately (21) or to the sum of E^{pair} and E^{self} for both term E_a and E_b (22). Using the cross terms with equation 9 requires N times more memory space to store pairwise and self energies for N energy terms considered in making corrections for over-counting. The cross terms are then added to the equation 1 to obtain the total energy.

Parameter optimizations were performed as described above with extra twelve cross terms, one for each correlation found in Table 4. The resulting energy functions were cross validated by side chain prediction as before. As shown in Table 3, using the cross terms significantly increases the side chain prediction accuracy, both for structure and for sequence recovery. The “combined” mode (22) performs better than the “separate” mode (21), and the native residues predicted correctly increase from ~25% to ~44% when the “combined” cross term is used.

Further analysis of the prediction accuracy across twenty amino acids (Figure 1) shows a dramatic rescue of correct predictions for polar side chains, such as Asp, Glu, Lys, Asn, Gln, Arg, Ser and Thr, when cross terms are applied. Phe, His, Pro, Trp and Tyr also have different degree of increase in correct prediction. The “combined” mode tends to be slightly better over twenty amino acids than the “separate” mode.

3.3 Correction for false prediction biases

The overall false prediction (Table 6) shows a strong bias towards Gly when the cross terms are not used. Although the false prediction is more evenly distributed across twenty amino acids while the energy function is corrected by the cross terms, false prediction biases largely in favor of Cys, Gly and His can still be observed. It may not be possible to have a perfect energy function that completely eliminates false prediction, but having strong false prediction biases to certain amino acids shows the energy model needs to be fine turned. The result indicates energy terms specific for individual amino acids, which capture the energetic difference of residues and correct the biases observed, are missing in our energy model. The missing term often is treated as the amino acid reference energy, and it is interpreted to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

capture the energetic differences between individual amino acid types in the unfolded state [18], [50], [51].

3.3.1 Amino acid reference energies—To balance the composition of false prediction in protein design, the amino acid reference energy (14) was added to our energy model. Parameter optimization was performed with twenty new energy terms, ref_{aa}, one for each amino acid. The parameters are trained with and without the "combined" cross terms (22) for comparison.

As the result showed in Table 4, adding the amino acid reference energy into the energy model improves the prediction accuracy by ~25% when the native side chain conformations are within top 1.5~6.0% of the lowest energy conformations, although both the prediction of native residues as the lowest energy conformations (Best) and as the lowest energy sequences (Sequence) don't improved as much as using the cross terms. The energy function with both the reference terms and the cross terms shows the best performance in all our measurements. Above 80% of the true conformations are predicted with favorable energies that are within top 6%. About 53% of the native residues are predicted as the lowest energy sequences, which is comparable to the previous study; however, adding the reference energy does not correct all of the false prediction biases, either with or without the cross terms. Interestingly, the retraining moved these biases to other amino acids, now overpredicting Ala instead of Cys or Gly.

3.3.2 Leaving out energy terms.—An alternative way to resolve the non-additivity of energy terms is to leave out the most highly correlated terms from the model. The correlation study (Table 4, upper-right) shows that the most non-additive energy terms are E_{NonHb} and E_{Vol}, which are highly correlated with VDW interactions. E_{NonHb} is designed to disfavor unsatisfied H-bond donor and acceptor atoms, but penalizing atoms other than H-bond donors and acceptors that occupies the dummy positions is also captured by the VDW interactions from the donors and acceptors. E_{Vol} is introduced to help the formation of tighter packing in protein cores by favoring side chains with larger volume; however, a well optimized VDW attraction term E_{Att} should have the same effect. Leaving out E_{NonHb} and E_{Vol} may therefore have a minor effect on accuracy but with the advantage of potentially removing eight of the high correlations. In addition, VDW repulsion interactions would need to be softened after removing these two terms. We chose a linearized Lennard-Jones potential [60], where the twelfth power term has been changed to a linear relationship with distance.

Knowing that the amino acid reference energy does not eliminate the false prediction biases, a two-step optimization strategy was adopted with the aim of having the designed sequences reproduce the composition of naturally occurred protein sequences. That is, we minimize the difference between native and predicted amino acid distribution,

$$\sum_{aa} |N_{aa,prdicted} - N_{aa,native}|, \quad (23)$$

in a second, after step all parameters are optimized with the objective function described above. In the second step, an extra term is added to the objective function,

$$-\sum_{aa} (N_{aa,prdicted} - N_{aa,native})^2, \quad (24)$$

where $N_{aa,predicted}$ is the number of the residues that have the amino acid *aa* being predicted as the lowest energy residue, and $N_{aa,native}$ is the number of the residues that have the amino

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

acid *aa* in the training protein set. Only twenty ref_{aa} are allowed to be refined in the second-step optimization with the goal of minimizing (23).

As shown in Table 4 (lower-left), the modified energy function eliminates nine out of twelve correlations originally found and brings two correlations, E_{Att}-E_{Hb}, and E_{Elect}-E_{Rep} above 0.5. The optimization of the modified energy function is performed without the cross terms, unneeded since the most important correlations have now been removed. The second step of optimization decreased the value of (23) from 3314 to 8 on the training set and from 3662 to 556 on the test set, making the distribution of designed amino acids closer to the true, natural distribution (Figure 2). The prediction accuracy has slightly decreased by 8~9% in all measures (Table 5, last row) but the false predictions (Table 6, last column) are now more evenly distributed.

Careful inspection revealed a concern that oppositely charged atoms were too close in some of the false positive rotamers. This occurred because the hyperbolic attractive Coulomb potential overwhelms the linearized Lennard-Jones potential at very short distances. The problem is easily fixed, fortunately. This artifact explains the strong negative correlation, E_{Elect}-E_{Rep}.

4 Discussion

We first examined different ways of summing the probabilities in objective functions, either using the sum of the probabilities of wild-type amino acids ΣP (15) or the sum of the logs of those probabilities ΣlnP (16). Although both can find weights that maximize the probability of recovering wild-type residues in designs, the ΣlnP function treats each residue as an independent event and gives all events equal weight, while the ΣP function focuses the training in proportion to the degree that the residue is already correctly predicted, downplaying the incorrectly predicted residues. The implicit assumption for ΣlnP is that *all residues have evolved to their lowest energy rotamer*, while the implicit assumption for ΣP is that *most but not necessarilly all residues have evolved to their lowest energy rotamer*.

The accuracy of side chain prediction using energy weights trained by ΣlnP is significantly better than the weights trained by ΣP. This result is explained by observing that poorly predicted residues contribute little to ΣP but they greatly affect ΣlnP, specifically in the numerator (19). We also observed that weights trained by ΣP favored VDW repulsion energy and void space energy, whereas weights trained using ΣlnP also favored solvation energy. The steepness of the VDW repulsion energy function means that a change in its weight has a larger effect on P when P is approaching one, since erroneous rotamers (in the denominator) are more likely to have collisions than correct ones. But many of these collisions are an artifact of the graininess of the rotamer library. Differences in solvation energy, parameterized on buried surface area, are insensitive to the graininess of the library and probably more relevant to natural selection. The objective function that better diffuses the contributions over all residues (i.e. ΣlnP) favors the descriminating power of the solvation energy in predicting the wild-type rotamer. When using ΣP, more of the weights converged on zero, as compared to using ΣlnP, probably due to the domination of the VDW repulsion term.

The whole training process can be viewed as a test of the validity of the assumptions built into the objective functions. To reiterate, the ΣlnP assumes all rotamers are in their lowest energy state, while for ΣP assumes that some of them are not in their lowest energy state, given the imperfect way energy is calculated. Both objective functions have their theoretical justification, as described above. ΣP suggests that true rotamers with poor energies can be “swept under the rug” when searching for the optimum energy function. ΣlnP says we

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

cannot afford to allow such energies. The conclusion of this study is it is best not to sweep anything under the rug!

We also considered two choices of criterea for correctness, in one case assigning correctness when the correct amino acid is predicted (Seq) and in the other case assigning correctness only when the correct amino acids and correct rotamer are both predicted (Struct). But we found little difference in the accuracy of the energy function when trained under these two criterea. We conclude that when we predict the correct amino acid, we also predict the correct rotamer. In other words, when the amino acid prediction is right, it is right for the right reason.

Our analysis of the dependencies of energy terms showed several strong correlations and anti-correlations. These relationships come from the imperfect separation of physical interaction types. For instance, the volume-based void space energy rewards tighter packing, but VDW attraction also rewards tighter packing. Also, forming a hydrogen bond between donor-acceptor pairs usually accompanies the increase of VDW repulsion from atoms near those pairs. Having stronger VDW repulsion usually means weaker VDW attraction. Hydrogen bonding interactions can be treated as a special type of charge-charge interactions. Existence of these correlations reflects non-additivity of terms and will impair efforts to balance the energy function, since linearly combined energy terms cannot capture the cross-talk between terms. Applying cross terms in the energy function shows a promising improvement in overall side chain prediction. Twenty amino acids, except Gly, all show enhancement of prediction to some extent when cross terms are used. Moreover, polar side chains gain the most benefit with cross terms, and this can be explained by the fact that seven out of thirteen cross terms are related to hydrogen bonding. The improvement of prediction for nonpolar side chains, Phe, Pro and Trp, and His can also be explained by the correction for overcounting by the VDW energy and void space energy terms.

Our study of the influence of the amino acid reference energy shows the best energy function requires both the corrections for the dependency of energy terms and for individual amino acids. With both, the prediction accuracy surpasses 80% when considering rotamers that are within top 6% of the lowest energy conformations. The over-counting of the energy can also be reverted by simply removing problematic energy terms, suggesting that highly correlated energy calculations should be avoided when creating a energy model. However, neither the cross terms nor the amino acid reference energy could completely correct the false prediction biases towards some amino acids, which produces non-native like amino acid compositions in designed proteins. The distribution of predicted amino acids is comparable to the distribution of all proteins found in NCBI database only when distribution itself is introduced into the objective function (Figure 3), again showing the importance of a carefully constructed objective function.

Finally, a concern was raised that the design test was too easy. We only required that the energy function to select the correct rotamer for a single position, leaving all neighboring side chains fixed. This is unrealistic in the sense that a real design problem would have several neighboring side chains unfixed. But this was the test we selected for two reasons. First, updating the energy for a single side chain after changing the parameters of the energy function would require repacking teh side chains. If this were done, the whole optimization process, which took tens of thousands of CPU hours as it is, would have taken millions of CPU hours. As such it was impossible even for the CCNI cluster. Second, we believed it was important to solve the easy problem first. If the energy function cannot predict the correct side chain given fixed neighbors, then there is no hope that it could solve the problem when the neighbors were not fixed.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Others have sought to optimize the energy function by machine learning, each with a different slant. Yanover [59] used an objective function similar to our ΣlnP and trained a weighted sum of term (similar to Eq. 1) on the problem of side chain prediction. In Yanover's case, the identity of the amino acid was given. Even so, the prediction accuracy for the best method was only 82.6%, using a fine-grained rotamer library. As in our case, Ser and other polar side chains were the most problematic. Sharabi [60] optimized a linear sum of terms against a database of interface side chains. In this case, the energy terms that were up-weighted by the learning procedure made sense in the context of protein-protein interfaces. VDW was downweighted, as was the polar side chain burial (similar to our Eq. 7). Sharabi reasoned that since explicit water was not being considered, polar burial must be allowed. By optimizing for the recovery of side chain conformation, they were able to also recover the native sequence, which mirrors our result in which sequence-based training recovered the side chain structure. In both cases, the sequence answer was right for the right structural reason. Leaver-Fay [61] used a variety of structural metrics to evaluate three specific hypotheses to improve the Rosetta energy function. One metric was over/under prediction numbers, which we also used. This led to improvements in sequence recovery by smoothing the knowledge-based φ,ψ potential, a change that could not have been discovered by machine learning. Rosetta's sequence recovery rates (~38%) are better than ours (25%) when the energy is expressed as a linear sum (Eq. 1), but we were able to recover as much as 53% of the sequence when we added cross-terms with reference weights (see Table 5).

We do not expect to develop an ideal energy function that eliminates all false designs, but having the ability in an energy function to accurately reduce the enormous space of all possible sequences down to a small set of suboptimal sequences that solve the design problem, makes feasible a faster and easier library screening in the wet lab.

Acknowledgments

We thank Robert Shaffer for helpful insights and experiments. We thank Dan DiTursi, Derek Pitman, Patrick Buck and Christian Schenkelberg for contributions to the DEEdesign program that was used in this study.

This work was supported by a grant from the NIH to C.B., R21 GM088838.

Biographies

Yao-ming Huang received a bacchelors degree and a national license in medical technology, and a masters degree in biochemistry from the National Technological University of Taiwan, where he was awarded the Presidential Award, the Academic Thesis Award of the Association of Laboratory Medicine, and the Graduate Prize Scholarship of the same group. He received his Ph.D. in biology from Renssealer Polytechnic Institute in 2008, where he studied with C.B. At Rensselaer he was awarded the Charles S. and Helen Humphrey Graduate Fellowship and the Founders Award of Excellence. He served as a postdoctoral scholar in the lab of C.B. from 2008-2011 and is currently a postdoctoral scholar at the University of California, San Francisco Department of Bioengineering and Therapeutive Sciences, working under Tanja Kortemme.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Christopher Bystroff is Associate Professor of Biology and Computer Science at Rensselaer Polytechnic Institute. He received his bacchelors degree from Carleton College, and his PhD. from the University of California, San Diego, both in Chemistry. He trained in crystallography with Joseph Kraut at UCSD, and Robert Fletterick at UCSF, and in bioinformatics with David Baker at the University of Washington. Chris spent three years as a Fulbright Fellow in Managua, Nicaragua. At RPI Chris has been funded by the NSF and NIH for the last ten years, including an NSF Early Career Award. He has published 47 journal articles, 8 book chapters and 1 book. Chris is an organizer of the Protein Society.

6 References

1. Park S, Yang X, Saven JG. Advances in computational protein design. Curr Opin Struct Biol. 2004; 14(4):487–94. [PubMed: 15313244]
2. Schueler-Furman O, et al. Progress in modeling of protein structures and interactions. Science. 2005; 310(5748):638–42. [PubMed: 16254179]
3. Pokala N, Handel TM. Review: protein design--where we were, where we are, where we're going. J Struct Biol. 2001; 134(2-3):269–81. [PubMed: 11551185]
4. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science. 1997; 278(5335):82–7. [PubMed: 9311930]
5. Desjarlais JR, Handel TM. De novo design of the hydrophobic cores of proteins. Protein Sci. 1995; 4(10):2006–18. [PubMed: 8535237]
6. Malakauskas SM, Mayo SL. Design, structure and stability of a hyperthermophilic protein variant. Nat Struct Biol. 1998; 5(6):470–5. [PubMed: 9628485]
7. Chevalier BS, et al. Design, activity, and structure of a highly specific artificial endonuclease. Mol Cell. 2002; 10(4):895–905. [PubMed: 12419232]
8. Jiang L, et al. De novo computational design of retro-aldol enzymes. Science. 2008; 319(5868): 1387–91. [PubMed: 18323453]
9. Rothlisberger D, et al. Kemp elimination catalysts by computational enzyme design. Nature. 2008
10. Lassila JK, et al. Computationally designed variants of Escherichia coli chorismate mutase show altered catalytic activity. Protein Eng Des Sel. 2005; 18(4):161–3. [PubMed: 15820980]
11. Allert M, et al. Computational design of receptors for an organophosphate surrogate of the nerve agent soman. Proc Natl Acad Sci U S A. 2004; 101(21):7907–12. [PubMed: 15148405]
12. Looger LL, et al. Computational design of receptor and sensor proteins with novel functions. Nature. 2003; 423(6936):185–90. [PubMed: 12736688]
13. Dwyer MA, Looger LL, Hellinga HW. Computational design of a biologically active enzyme. Science. 2004; 304(5679):1967–71. [PubMed: 15218149]
14. Reina J, et al. Computer-aided design of a PDZ domain to recognize new target sequences. Nat Struct Biol. 2002; 9(8):621–7. [PubMed: 12080331]
15. Song G, et al. Rational design of intercellular adhesion molecule-1 (ICAM-1) variants for antagonizing integrin lymphocyte function-associated antigen-1-dependent adhesion. J Biol Chem. 2006; 281(8):5042–9. [PubMed: 16354667]
16. Lippow SM, Wittrup KD, Tidor B. Computational design of antibody-affinity improvement beyond in vivo maturation. Nat Biotechnol. 2007; 25(10):1171–6. [PubMed: 17891135]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

17. Kuhlman B, et al. Design of a novel globular protein fold with atomic-level accuracy. Science. 2003; 302(5649):1364–8. [PubMed: 14631033]
18. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A. 2000; 97(19):10383–8. [PubMed: 10984534]
19. Dunbrack RL Jr. Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 1997; 6(8):1661–81. [PubMed: 9260279]
20. Dunbrack RL Jr. Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol. 1993; 230(2):543–74. [PubMed: 8464064]
21. Lovell SC, et al. The penultimate rotamer library. Proteins. 2000; 40(3):389–408. [PubMed: 10861930]
22. Allen BD, Mayo SL. An efficient algorithm for multistate protein design based on FASTER. J Comput Chem. 2010; 31(5):904–16. [PubMed: 19637210]
23. Humphris EL, Kortemme T. Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design. Structure. 2008; 16(12):1777–88. [PubMed: 19081054]
24. Saunders CT, Baker D. Recapitulation of protein family divergence using flexible backbone protein design. J Mol Biol. 2005; 346(2):631–44. [PubMed: 15670610]
25. Desmet J, et al. The dead-end elimination theorem and its use in protein side-chain positioning. Nature. 1992; 356(6369):539–42. [PubMed: 21488406]
26. Goldstein RF. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys J. 1994; 66(5):1335–40. [PubMed: 8061189]
27. Gordon DB, Mayo SL. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. J Comput Chem. 1998; 19(13):1505–14.
28. Keller DA, et al. Finding the global minimum: a fuzzy end elimination implementation. Protein Eng. 1995; 8(9):893–904. [PubMed: 8746727]
29. Looger LL, Hellinga HW. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. J Mol Biol. 2001; 307(1):429–45. [PubMed: 11243829]
30. Pierce NA, et al. Conformational splitting: A more powerful criterion for dead-end elimination. J Comput Chem. 2000; 21(11):999–1009.
31. Hellinga HW, Richards FM. Optimal sequence selection in proteins of known structure by simulated evolution. Proc Natl Acad Sci U S A. 1994; 91(13):5803–7. [PubMed: 8016069]
32. Lee C, Levitt M. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. Nature. 1991; 352(6334):448–51. [PubMed: 1861725]
33. Jiang X, et al. A new approach to the design of uniquely folded thermally stable proteins. Protein Sci. 2000; 9(2):403–16. [PubMed: 10716193]
34. Jones DT. De novo protein design using pairwise potentials and a genetic algorithm. Protein Sci. 1994; 3(4):567–74. [PubMed: 8003975]
35. Gordon DB, Mayo SL. Branch-and-terminate: a combinatorial optimization algorithm for protein design. Structure. 1999; 7(9):1089–98. [PubMed: 10508778]
36. Wernisch L, Hery S, Wodak SJ. Automatic protein design with all atom force-fields by exact and heuristic optimization. J Mol Biol. 2000; 301(3):713–36. [PubMed: 10966779]
37. Leach AR, Lemon AP. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. Proteins. 1998; 33(2):227–39. [PubMed: 9779790]
38. Davis IW, et al. The backrub motion: how protein backbone shrugs when a sidechain dances. Structure. 2006; 14(2):265–74. [PubMed: 16472746]
39. Georgiev I, et al. Algorithm for backrub motions in protein design. Bioinformatics. 2008; 24(13):i196–204. [PubMed: 18586714]
40. Smith CA, Kortemme T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J Mol Biol. 2008; 380(4): 742–56. [PubMed: 18547585]
41. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods. 2009; 6(8):551–2. [PubMed: 19644455]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

42. Cuff AL, Martin AC. Analysis of void volumes in proteins and application to stability of the p53 tumour suppressor protein. J Mol Biol. 2004; 344(5):1199–209. [PubMed: 15561139]
43. Song Y, et al. Structure-guided forcefield optimization. Proteins. 2011; 79(6):1898–909. [PubMed: 21488100]
44. Liang S, Grishin NV. Effective scoring function for protein sequence design. Proteins. 2004; 54(2): 271–81. [PubMed: 14696189]
45. Word JM, et al. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol. 1999; 285(4):1711–33. [PubMed: 9917407]
46. van Gunsteren WF, et al. Biomolecular Simulation: The GROMOS96 manual and user guide. Zurich, Switzerland: Hochschulverlag AG an der ETH Zurich. 1996
47. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci. 1994; 3(3): 522–4. [PubMed: 8019422]
48. Iba Y. Extended Ensemble Monte Carlo. International Journal of Modern Physics C: Computational Physics & Physical Computation. 2001; 12(5):623–56.
49. Eriksson AE, Baase WA, Matthews BW. Similar hydrophobic replacements of Leu99 and Phe153 within the core of T4 lysozyme have different structural and thermodynamic consequences. J Mol Biol. 1993; 229(3):747–69. [PubMed: 8433369]
50. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci U S A. 2002; 99(22):14116–21. [PubMed: 12381794]
51. Ding F, Dokholyan NV. Emergence of protein fold families through rational design. PLoS Comput Biol. 2006; 2(7):e85. [PubMed: 16839198]
52. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. Proc Natl Acad Sci U S A. 1997; 94(19):10172–7. [PubMed: 9294182]
53. Dahiyat BI, Gordon DB, Mayo SL. Automated design of the surface positions of protein helices. Protein Sci. 1997; 6(6):1333–7. [PubMed: 9194194]
54. Street AG, Mayo SL. Pairwise calculation of protein solvent-accessible surface areas. Fold Des. 1998; 3(4):253–8. [PubMed: 9710572]
55. Zhang N, Zeng C, Wingreen NS. Fast accurate evaluation of protein solvent exposure. Proteins. 2004; 57(3):565–76. [PubMed: 15382246]
56. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature. 1986; 319(6050):199–203. [PubMed: 3945310]
57. Dahiyat BI, Mayo SL. Protein design automation. Protein Sci. 1996; 5(5):895–903. [PubMed: 8732761]
58. Bystroff C. MASKER:improved solvent-excluded molecular surface area estimations using Boolean maks. Protein Eng. 2003; 15(12):959–65. [PubMed: 12601135]
59. Yanover C, Schueler-Furman O, et al. Minimizing and learning energy functions for side-chain prediction. J Comput Biol. 2008; 15(7):899–911. [PubMed: 18707538]
60. Sharabi O, Yanover C, et al. Optimizing energy functions for protein-protein interface design. J Comput Chem. 2011; 32(1):23–32. [PubMed: 20623647]
61. Leaver-Fay A, O'Meara MJ, et al. Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol. 2013; 523:109–143. [PubMed: 23422428]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

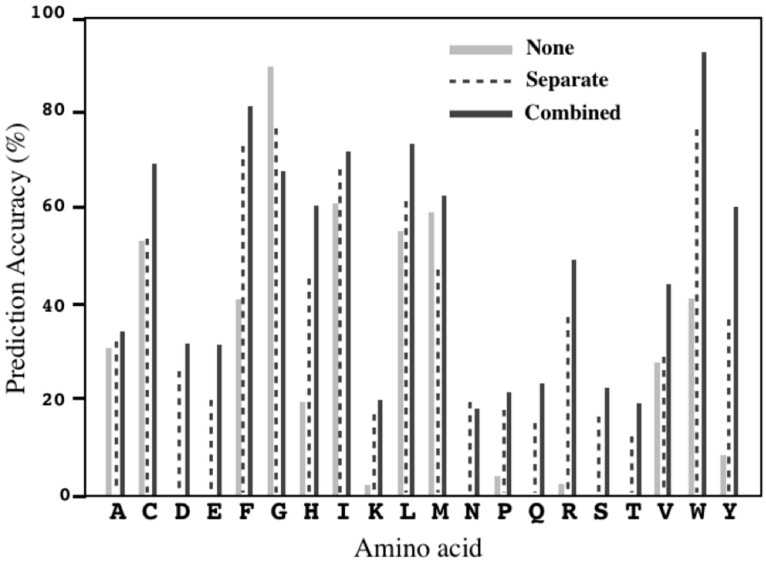

Figure 1.

Accuracy of predicting native residues as the lowest energy conformations. The side chain prediction is correct if the lowest energy conformations have the native residues. The prediction is made using optimized energy functions with “combined” (green), “separate” (red) or without cross terms (blue).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

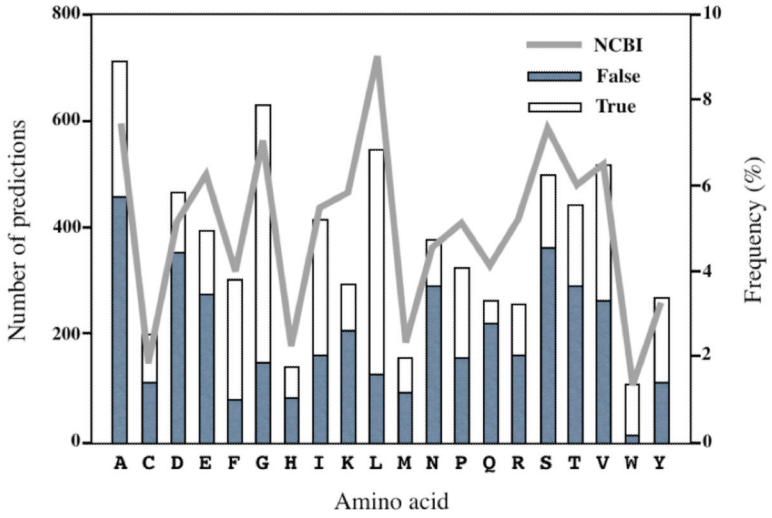

Figure 2.

Distribution of predicted amino acids, true positives (green) and false positives (blue). The amino acid distribution over all proteins is shown in red (NCBI). The prediction is made using the modified energy model and the two-step optimization with reference energies.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

The 80 proteins used for training and testing the energy function are listed as 4-letter PDB code followed by chain identifier, if any.

1aac	1aky	1amm	1arb	1aru	1bkf	1bpi	1cem
1cka	1cnr	1cpcB	1cse	1ctj	1cus	1cyo	1dad
1dif	1fnc	1fxd	1hfc	1ifc	1igd	1iro	1isuA
1jbc	1kap	1knb	1lam	1lit	1lkk	1mctI	1mla
1mrj	1nif	1osa	1phb	1php	1plc	1poa	1ptx
1ra9	1rcf	1rgeA	1rro	1sri	1tca	1ttaA	1whi
1xic	1xsoA	1xyzA	256bA	2ayh	2bopA	2cba	2cpl
2ctc	2end	2er7	2erl	2hft	2ihl	2mcm	2mhr
2msbA	2olb	2phy	2rhe	2rn2	2trxA	3chy	3ebx
3grs	3lzm	3pte	3sdhA	4fgf	5p21	7rsa	8abp

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Optimized weights from various objective functions.

Weight[a]	Objective function			
	Sequence oriented		Structure oriented	
	ΣP	ΣlnP	ΣP	ΣlnP
Rep	0.617	0.519	0.595	0.613
Att	<0.001	0.088	<0.001	0.127
Elect	0.073	0.003	0.049	0.003
Hb	0.170	0.084	0.197	0.078
NonHb	0.015	0.005	0.013	0.005
Solv1	<0.001	0.253	<0.001	0.121
Solv2	<0.001	<0.001	<0.001	<0.001
Rot	<0.001	<0.001	<0.001	0.002
Vol	0.124	0.048	0.146	0.050

[a]Converged weights used in side chain prediction are listed.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

Accuracy of side chain prediction given choice of objective function.

Obj. fun[b]	Prediction accuracy (%)[a]			
	Best	**1.5%**	**3.0%**	**6.0%**
Seq, ΣP	14.32	34.94	44.95	56.68
Seq, ΣlnP	19.32	39.86	49.96	62.64
Struc, ΣP	14.07	33.24	43.66	56.11
Struc, ΣlnP	19.78	40.96	51.34	63.04

[a]The accuracy of side chain prediction is evaluated if native side chains are the lowest energy conformations (Best) or within top 1.5, 3.0 or 6.0% of the lowest energy rotamers.

[b]Sequence, Seq (18), or structure, Struc (17), oriented objective functions using the sum of probability, ΣP (15), or the sum of log probability, ΣlnP (16), were used in the optimization of weights.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 4

Correlation coefficients between any two unweighted energy terms. Results from the original energy model (upper-right) and after leaving out $\mathbf{E_{NonHb}}$ and $\mathbf{E_{vol}}$ and adapting soften VDW repulsion (lower-left) are shown. Correlation coefficients greater than 0.5 or less than −0.5 are shown in bold.

	E_{Rep}	E_{Att}	E_{Elect}	E_{Hb}	E_{NonHb}	E_{Solv1}	E_{Solv2}	E_{Rot}	E_{Vol}
E_{Rep}		**−0.98**	−0.19	**−0.50**	**0.93**	−0.27	−0.30	0.05	**−0.83**
E_{Att}	−0.33		0.17	0.44	**−0.94**	0.36	0.36	−0.05	**0.89**
E_{Elect}	**−0.62**	0.28		**0.54**	−0.22	−0.24	−0.27	−0.12	0.03
E_{Hb}	**−0.60**	**0.55**	**0.53**		**−0.57**	−0.43	−0.28	−0.17	0.24
E_{NonHb}	-	-	-	-		−0.12	−0.17	0.12	**−0.77**
E_{Solv1}	0.19	0.35	−0.15	−0.38	-		**0.64**	0.21	**0.56**
E_{Solv2}	0.17	0.37	−0.15	−0.17	-	**0.58**		0.18	**0.53**
E_{Rot}	0.09	−0.09	−0.09	−0.13	-	0.18	0.14		0.03
E_{Vol}	-	-	-	-	-	-	-	-	

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 5

Accuracy of side chain prediction with cross terms and amino acid reference energy

Energy function	Prediction accuracy (%)[a]				
	Best	1.5%	3.0%	6.0%	Sequence
None[b]	20.24	34.85	43.41	54.83	25.06
S[c]	26.31	47.99	56.82	66.51	37.49
C[d]	33.61	54.07	62.15	71.42	44.18
Ref[e]	25.30	58.25	69.15	77.50	36.34
C+Ref[f]	38.79	67.56	75.90	83.47	53.10
2°Ref[g]	29.12	57.89	66.74	75.37	44.67

[a]The accuracy of side chain prediction is evaluated if native side chains are the lowest energy conformations (Best), within top 1.5, 3.0 or 6.0% of the lowest energy rotamers, or if the lowest energy rotamers are of native residues (Sequence).

[b]Only single terms are use (1). Glycine is included in the training.

[c]The cross terms are added as (21).

[d]The cross terms are added as (22).

[e]The amino acid reference energy terms are added as (14).

[f]Both the amino acid reference energy (14) and the cross terms (22) are added.

[g]The modified energy function with the amino acid reference energy is optimized using two-step optimization.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 6

The distribution of false predicted amino acids.

False AA[a]	Energy function					
	None[b]	S[c]	C[d]	Ref[e]	C+Ref[f]	2°Ref[g]
A	163	62	38	3352	1207	456
C	475	1038	1674	0	2	113
D	0	157	143	133	213	354
E	3	112	74	98	178	277
F	13	96	71	68	42	81
G	3806	986	448	38	490	153
H	128	444	529	12	59	86
I	241	242	131	45	61	165
K	74	170	121	47	58	209
L	41	42	35	292	122	129
M	245	131	151	1	6	95
N	1	341	97	23	62	294
P	3	18	14	50	244	158
Q	2	119	95	19	36	228
R	94	210	148	55	86	165
S	5	87	74	79	90	364
T	3	50	40	92	128	288
V	99	44	22	152	237	265
W	15	134	93	13	20	13
Y	6	35	37	32	49	106
S.D.[h]	841	289	371	738	270	114

[a]The type of amino acids is predicted when the sidechain prediction is not correct.

[b]The single terms only (1).

[c]The single terms (1) + the cross terms in “separate” mode (21).

[d]The single terms (1) + the cross terms in “combined” mode (22).

[e]The single terms (1) + the amino acid reference energy (14).

[f]The single terms (1) + the cross terms in “combined” mode (22) + the amino acid reference energy (14).

[g]The modified single terms + the amino acid reference energy; two-step optimization.

[h]Standard deviation of the number of false prediction over twenty amino acids.