Ding F, Dokholyan NV (2006) Emergence of protein fold families through rational design. PLoS Comput Biol 2(7): e85.
DOI: 10.1371/journal.pcbi.0020085

# Supporting Information

**Figure S1. The Protein Sequence/Structure Optimization with Backbone Relaxation**

The sequence identities, backbone deviation, and the energy at the end of each iteration are shown.

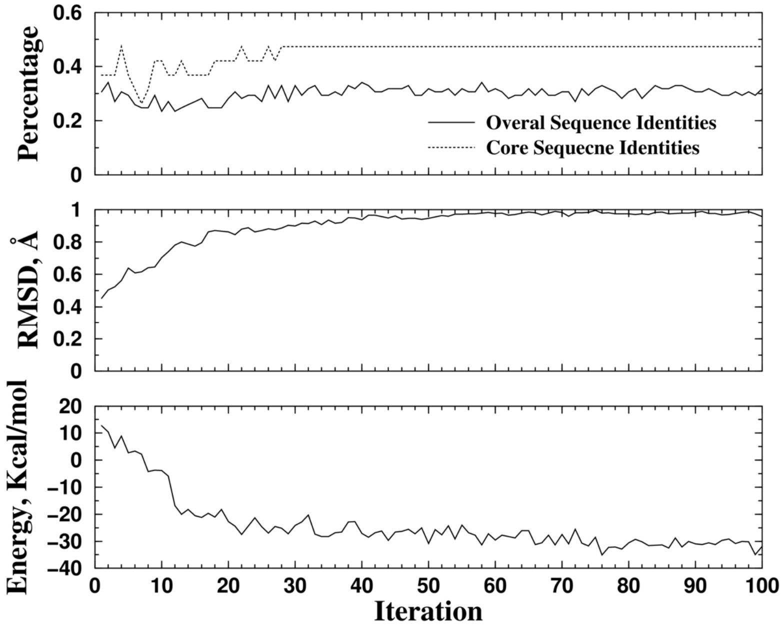

**Figure S1**. The protein sequence/structure optimization with backbone relaxation. The sequence identities, backbone deviation, and the energy at the end of each iteration are shown.

**Protocol S1. Supporting Materials and Methods**

**Energy terms.** The energy of a given sequence and structure is evaluated by the following terms:

*Van der Waals Potentials ($E_{vdw\_attr}$, $E_{vdw\_rep}$)*. We use a standard 12-6 Lennard-Jones (LJ) potential to model the attractive portion of Van der Waals interactions. For simplicity, we use a linear extrapolation to model VDW repulsion to damp the quick increase of regular LJ repulsions.

$$E_{vdw-attr} = \sum_{i,j>i} 4\varepsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}], r_{ij} > \sigma_{ij}$$

$$E_{vdw\_rep} = \begin{cases} \sum_{i,j>i} 4\varepsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}], \alpha_{cutoff}\sigma_{ij} < r_{ij} \le \sigma_{ij} \\ K_{slope} r_{ij} + 4\varepsilon_{ij}(\alpha_{cutoff}{}^{-12} - \alpha_{cutoff}{}^{-6}) - \alpha_{cutoff} K_{slope}\sigma_{ij}, r_{ij} \le \alpha_{cutoff}\sigma_{ij} \\ Here, \alpha_{cutoff} = 0.93; K_{slope} = -24\varepsilon_{ij}(2\alpha_{cutoff}{}^{-13} - \alpha_{cutoff}{}^{-7})/\sigma_{ij} \end{cases} . \quad \text{(S2)}$$

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}; \sigma_{ij} = \sigma_i + \sigma_j$$

Here, $r_{ij}$ is the distance between two atoms *i* and *j*. The energy parameters $\varepsilon$, $\sigma$ are taken from the CHARMM19 force field of united atoms [1].

*Solvation energy ($E_{solv}$).* We approximate the solvation energy with EEF1 model proposed by Lizaridis and Karplus [2]**:**

$$E_{solv} = \sum_{i,j>i} [-\frac{2\Delta G_i^{free}}{4\pi\sqrt{\pi}\lambda_i r_{ij}^2} \exp(-x_{ij}^2)V_i - \frac{2\Delta G_j^{free}}{4\pi\sqrt{\pi}\lambda_j r_{ij}^2} \exp(-x_{ji}^2)V_j] . \quad \text{(S3)}$$

$$x_{ij} = (r_{ij} - 1.12\sigma_i)/\lambda_i; x_{ji} = (r_{ij} - 1.12\sigma_j)/\lambda_j$$

Here, the parameters of reference solvation energy ($\Delta G^{free}$), volume of atoms (V), correlation length ($\lambda$), and atom radius ($\sigma$) are taken from Lizaridis and Karplus. We do not include the intrinsic solvation energy [2] of each atom, since the sum of these terms for each amino acid is a constant that can be incorporated into the reference energies. The calculation of van der Waals and solvation energies uses a distance cutoff of 9Å.

To avoid the over-estimation of the solvation energy when the atom pair *i,j* are too close, we define a cutoff distance $d_{ij}=0.95\sigma_{ij}$. When the distance $r_{ij}$ is smaller than $d_{ij}$, the solvation energy will be a constant of $E(d_{ij})$.

*Hydrogen bond interactions ($E_{bb\_hbond}$, $E_{sc\_hbond}$ and $E_{bb\_sc\_hbond}$).* To model the electrostatic interactions by hydrogen bond formation, we use an orientation-dependant hydrogen bond model proposed by Kortemme and Baker [3]. The energy of backbone-backbone ($E_{bb\_hbond}$), sidechain-backbone ($E_{bb\_sc\_hbond}$) and sidechain-sidechain ($E_{sc\_hbond}$) hydrogen bonds were derived from the statistics of naturally-occurring hydrogen bonds in the PDB database.

In the hydrogen bond model, the optimal distance between donor and acceptor in a hydrogen bond is approximately 2.95Å, which is often much smaller than the VDW radii between these atoms in CHARMM19 force field. To favor the formation the hydrogen bonds, we assign the VDW radii between all the potential donor and acceptor pairs $\sigma_{donor-acceptor} = 2.95/2^{1/6}$Å.

*Backbone dependent internal energies ($E_{\phi.\psi|aa}$, $E_{\phi.\psi,aa|rot}$).* We use a backbone dependent rotamer library from Dundrack and Cohen [4] to perform the search of optimal sidechian packing. To account for the internal energy for the rotamer state, we use the log normal of the natural occurring probability for a specific rotamer in the PDB, derived by Dunbrack et al., as its internal energy:

$$E_{\phi,\psi,aa|rot} = -\ln(P_{\phi,\psi,aa|rot})\,. \qquad \text{(S4)}$$

To account for the variation of amino acids in a specific backbone $\phi.\psi$ conformation, we also include the internal energy for different amino acids. We compute from the non-redundant PDB library the natural occurring probability of different amino acids in a given backbone dihedral angles $\phi,\psi$, $P_{\phi,\psi|aa}$. To be consistent with the Dunbrack backbone dependent rotamer library, we use a 10 degree bin size for $\phi,\psi$ angles. We use the log normal of $P_{\ \phi,\psi|aa}$ as the energy contribution:

$$E_{phi,psi|aa} = -\ln(P_{phi,psi|aa})\,. \qquad \text{(S5)}$$

Thus, $E_{\phi,\psi|aa}$ and $E_{\phi,\psi,aa|rot}$ will account for the internal energy between the (*aa,rot*) pair in the sidechain with the neighboring backbone atoms. In order to avoid overcounting in the energy calculation, the calculation of van der Waals and solvation energies will exclude interactions between *i*th residue and its neighboring backbone atoms $C_{\alpha i-1}$, $C_{i-1}$, $O_{i-1}$, $N_{i+1}$, $HN_{i+1}$, and $C_{\alpha i+1}$.

*Reference energy ($E_{ref}$).* We use an empirical energy for each amino acid to model the unfolded the unfolded state.

$$E_{ref} = \sum_i W(aa_i) = \sum_{aa} n_{aa} W(aa)\,. \qquad \text{(S6)}$$

**Weights determination**. Since the energy terms we use are taken from different sources, we use weighted coefficients, *W*, to scale the contribution for each of them to total energy. In order to determine these coefficients, we use a procedure similar to the one proposed by Kuhlman and Baker [5]. For a given position, we substitute the native amino acid/rotamer, ($aa_{nat}$,$rot_{nat}$), with all possible amino acid/rotamer (*aa,rot*). For each (*aa,rot*), we compute the corresponding energy, *E*(aa,rot). We use a training set of 34 high-resolution x-ray crystal structures and maximize the probability to observe the native (*aa,rot*) at each position,

$$P(aa_{nat}, rot_{nat}) = \exp(-E(aa_{nat}, rot_{nat}))\,/\sum_{(aa,rot)} \exp(-E(aa,rot))\,. \qquad \text{(S7)}$$

Then, we maximize the product of $\prod_{prot,i} P(aa_{nat}, rot_{nat})^{i}_{prot}$ at each position for all proteins in the training set. In addition, we introduce a term into the optimization function to minimize the deviation of the twenty amino acids distributions between the prediction and observation. For each position, we select the amino acid with lowest energy as the prediction.

$$S = -\sum_{prot,i} \ln(P(aa_{nat}, rot_{nat})^{i}_{prot} + \sum_{aa} [n^{pred}_{aa} - (\sum_{aa'} n^{pred}_{aa'}) p^{PDB}_{aa}]^2\,. \qquad \text{(S8)}$$

Here, $n^{pred}_{aa}$ is the number of amino acid *aa* that are predicted for all the positions in the training set. $P^{PDB}_{aa}$ represents the distribution of twenty amino acids in PDB.

*Sub-rotamers.* For a given discrete rotamer state, there is an allowed dihedral angle variations with its standard deviations tabulated in the rotamer library [4]. Thus, a small deviation from the average dihedrals results in a different energy. To compute the energy contributions for a specific rotamer, we use the lowest energy sidechain conformation with sidechain dihedral angles limited within one standard deviation around the average value. We confine dihedral angles by assigning a half-harmonic potential near the two edges:

$$E_{confinement}(\theta) = \begin{cases} E_0(\theta - (\bar{\theta} - \Delta\theta))^2, \theta < \bar{\theta} - \Delta\theta \\ 0, \bar{\theta} - \Delta\theta < \theta < \bar{\theta} + \Delta\theta \\ E_0(\theta - (\bar{\theta} + \Delta\theta))^2, \theta > \bar{\theta} + \Delta\theta \end{cases} . \qquad \text{(S9)}$$

Here, $E_0$ is a constant chosen be 100 kcal/mol to impose a strong penalization for deviations. Parameters $\bar{\theta}, \Delta\theta$are the average sidechain dihedral angle and its standard deviation taken from Dundrack rotamer library. We use a conjugate-gradient based minimization algorithm to find the lowest energy sub-rotamer conformations and then to compute the corresponding energy terms.

To perform the minimization, we start with a pre-assigned set of weights to calculate the energies for each rotamer. Based on computed energies, we determine the optimal weights by minimizing the scoring function of Eq. (S8). The new weights are used for the next round of calculations. With such an iterative procedure in determining the optimal weights, we find that the weights rapidly converge after two iterations. We list the determined weights in Table S1.

**Energy calculation.** In our protein model, each residue position is fitted with all twenty amino acid types, although only one amino acid type is active when the energy is evaluated. A point mutation is achieved by simply deactivating and activating the related amino acids at a given position. The rotamer state of an amino acid is obtained by appropriate rotations of the sidechain dihedral angles to the values given in the rotamer library.

To facilitate the calculation of energy, we put the protein into a three dimensional grid, where each cell is a cubic box with the dimension equal to the interaction cutoff distance, beyond which the interaction is zero. Each cell keeps a list of atoms within it. Thus, the energy evaluation for a given atom is only between itself and atoms in the neighboring $3^3$=27 cells. The hydrogen bond interaction has a maximum interaction distance of 3 Å [3], shorter than that of the VDW and solvation interaction, which is set as 9 Å in our study. Therefore, we use two different grids for hydrogen bond interaction and for VDW + solvation interactions, respectively: for VDW and solvation interactions we use a grid with a cell-dimension of 9 Å; for hydrogen bond interaction, we use a grid with a cell-dimension of 3 Å.

## Reference List

1. Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. Journal of Chemical Physics 105: 1902-1921.

2. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. Proteins 35: 133-152.

3. Kortemme T, Morozov AV, 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.

4. Dunbrack RL, Jr., Cohen FE (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 6: 1661-1681.

5. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97: 10383-10388.

**Table S1. The Weight of Each Energy Term (Materials and Methods)**

**Table S1**. The weight of each energy terms (Methods).

| Energy wrights | The parameterized value |
|---|---|
| $W_{VDW\text{-}attr}$ | 1.00 |
| $W_{VDW\text{-}rep}$ | 0.87 |
| $W_{solv}$ | 0.73 |
| $W_{bb\text{-}hbond}$ | 2.00 |
| $W_{bb\text{-}sc\text{-}hbond}$ | 1.63 |
| $W_{sc\text{-}hbond}$ | 1.84 |
| $W_{\phi,\psi\|aa}$ | 0.63 |
| $W_{\phi,\psi,aa\|rot}$ | 0.62 |
| $W_{CYS}$ | 2.00 |
| $W_{MET}$ | -0.49 |
| $W_{PHE}$ | 2.74 |
| $W_{ILE}$ | 0.52 |
| $W_{LEU}$ | 0.57 |
| $W_{VAL}$ | -0.07 |
| $W_{TRP}$ | 4.32 |
| $W_{TYR}$ | 2.46 |
| $W_{ALA}$ | -1.07 |

| | |
|---|---|
| $W_{GLY}$ | -1.41 |
| $W_{THR}$ | -0.86 |
| $W_{SER}$ | -1.20 |
| $W_{GLN}$ | -0.22 |
| $W_{ASN}$ | -1.15 |
| $W_{GLU}$ | -2.16 |
| $W_{ASP}$ | -2.43 |
| $W_{HIS}$ | 0.88 |
| $W_{ARG}$ | -0.04 |
| $W_{LYS}$ | -1.33 |
| $W_{PRO}$ | 0.94 |

**Table S2. The Average Native Sequence Recapitulation Rate between Native and Redesigned Sequences, with Different Subsets of Energy Terms**

**Table S2**. The average native sequence recapitulation rate between native and redesigned sequences with different subsets of energy terms. To turn off an energy term, we set the corresponding weight, W, zero. The detail description of different weights can be found in Method section.

| **Control Simulations** | **Core** | **Overall** |
|---|---|---|
| Full coefficient set $\{W\}$ | 58.6% | 37.8% |
| $W_{solv}=0$ | 15.2% | 8.1% |
| $W_{solv}=0;\ W_{ref}=0$ | 13.0% | 12.3% |
| $W_{solv}=0;\ W_{HB}=0$ | 16.0% | 11.9% |
| $W_{solv}=0;\ W_{HB}=0;\ W_{ref}=0$ | 13.3% | 20.1% |
| $W_{\phi.\psi\vert aa}=0;\ W_{\phi.\psi,aa\vert rot}=0$ | 11.6% | 19.8% |