Supplementary Material for the Web

Native Protein Sequences are Close to Optimal for Their Structures
Brian Kuhlman and David Baker

The energy of a protein is calculated as follows.

$$E_{protein} = W_{rot}E_{rot} + W_{atr}E_{atr} + W_{rep}E_{rep} + W_{solv}E_{solv} + W_{pair}E_{pair} + W_{hbond}E_{hbond} - E_{ref}$$

The *W* are scale factors for each energy term and are determined as described below.

$$E_{rot} = \sum_{i}^{nres} -\ln(P(rot(i) \mid phi(i), psi(i)) - \ln(\frac{P(aa(i) \mid phi(i), psi(i))}{P(aa(i))})$$

E_{rot} is related to the self-energy of a rotamer/amino acid and is derived from Protein Data Bank statistics by observing the probability of a particular rotamer and amino acid for a given phi angle and psi angle. The probability for a rotamer given phi and psi is taken directly from Dunbrack and Cohen (1).

$$E_{atr} = \sum_{i}^{natom}\sum_{j>i}^{natom}\left[\left(\frac{r_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{r_{ij}}{d_{ij}}\right)^{6}\right]e_{ij} \quad if \quad \frac{r_{ij}}{d_{ij}} < 1.12$$

$$E_{rep} = \sum_{i}^{natom}\sum_{j>i}^{natom}\left[10.0 - 11.2\left(\frac{d_{ij}}{r_{ij}}\right)\right] \quad if \quad \frac{r_{ij}}{d_{ij}} > 1.12$$

E_{atr} is the attractive portion of a 12-6 Lennard-Jones potential. d_{ij} is the distance between the two atoms, r_{ij} is the sum of the van der Waals radii and e_{ij} the square root of the product of the well depths. The values for r_{ij} and e_{ij} are taken from the CHARMM19 parameter set(2). Explicit hydrogens are not included except when looking for sidechain-backbone hydrogen bonds. E_{rep} expresses the repulsive energy between two atoms and is dampened in comparison to the typical 12-6 potential because a fixed backbone and rotamer set is being used in this model. E_{atr} and E_{rep} are not evaluated between atoms in the same amino acid because these energies are already part of the E_{rot} term.

$$E_{solv} = \sum_{i}^{natom} \left(\Delta G_i^{ref} - \sum_{j>i}^{natom} \left(\frac{2\Delta G_i^{free}}{4\pi\sqrt{\pi}\lambda_i r_{ij}^2} \exp(-d_{ij}^2) V_j + \frac{2\Delta G_j^{free}}{4\pi\sqrt{\pi}\lambda_j r_{ij}^2} \exp(-d_{ji}^2) V_i \right) \right)$$

E_{solv} is the solvation energy of an atom calculated by using an implicit solvent model developed by Lazaridis and Karplus (3). d_{ij} and r_{ij} are the same as in E_{atr}, ΔG^{ref} are ΔG^{free} related to the solvation energy of the fully solvated atom, λ_i is a correlation length, and V is atomic volume. The values for the parameters are taken from Lazaridis and Karplus.

$$E_{pair} = \sum_{i}^{nres} \sum_{j>i}^{nres} \frac{P(aa_i, a\,a_j \mid d_{ij}, env_i, env_j)}{P(aa_i \mid d_{ij}, env_i) P(aa_j \mid d_{ij}, env_j)}$$

E_{pair} is derived from the probability of seeing two amino acids close together in space in the PDB database after accounting for the intrinsic probabilities of these amino acids to be in that environment. This term primarily reflects electrostatic effects (4).

$$E_{hbond} = \left[5\left(\frac{r_{ij}}{d_{ij}}\right)^{12} - 6\left(\frac{r_{ij}}{d_{ij}}\right)^{10} \right] F(q) \quad \text{between side chain and backbone only}$$

E_{hbond} is the energy of side chain backbone hydrogen bonds. The potential is taken directly from Gordon *et al*. (5). d_{ij} is the distance between the acceptor and the donor and r_{ij} is the optimal hydrogen bonding distance. $F(\theta)$ describes the angular dependence of the hydrogen bond. If the backbone atom is already participating in a backbone backbone hydrogen bond the side chain backbone hydrogen bond energy is set to 0.

Last, every amino acid has a reference energy, E_{ref}.

$$E_{ref} = \sum_{i}^{nres} W_{ref}(aa(i))$$

The weights on these terms and the 20 reference energies were determined by maximizing the product of $\exp(-E(aa_{obs}))/(\sum \exp(-E(aa_i)))$ over a training set of 30 proteins using a conjugate-gradient-based optimization method, where $E(aa_{obs})$ is the energy of the native amino acid at a position and the partition function in the denominator is over all 20 amino acids at each position. In this process only one residue was changed at a time and all other residues were kept in their native conformation. Subsequently the parameters were refined slightly on the basis of the results of complete redesign calculations on the training-set proteins.

Table 2: weights and reference energies (kcal / mol).

W_{atr}	0.77	W_{ref}(Lys)	-12.6
W_{rep}	0.54	W_{ref}(Leu)	-1.31
W_{sol}	0.61	W_{ref}(Met)	-3.45
W_{pair}	0.23	W_{ref}(Asn)	-7.20
W_{hbond}	2.59	W_{ref}(Pro)	-1.49
W_{rot}	0.75	W_{ref}(Gln)	-7.72
W_{ref}(Ala)	-0.93	W_{ref}(Arg)	-17.2
W_{ref}(Asp)	-12.7	W_{ref}(Ser)	-4.08
W_{ref}(Glu)	-13.1	W_{ref}(Thr)	-4.30
W_{ref}(Phe)	-3.36	W_{ref}(Val)	-1.55
W_{ref}(Gly)	-0.65	W_{ref}(Trp)	-8.83
W_{ref}(His)	-8.73	W_{ref}(Tyr)	-7.11
W_{ref}(Ile)	-1.83		

1.Dunbrack, R. L. & Cohen, F. E. (1997). *Protein Sci* **6**, 1661-1681.

2. Neria, E., Fischer, S. & Karplus, M. (1996). *J. Chem. Phys.* **105**, 1902-1921.

3. Lazaridis, T. & Karplus, M. (1999). *Proteins: Struct. Func. Genet.* **35**, 133-152.

4. Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). *Proteins: Struct. Func. Genet.* **34**, 82-95.

5. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). *Curr. Opin. Struct. Biol.* **9**, 509-513.