Supporting Text

Reference Energy Calculation

The reference state for each residue is chosen to be an abstraction of the denatured state that captures contextual residue interactions. The denatured state is typically conceptualized as an ensemble of partially unfolded structures. Here we use the "expanded" state of Elcock (1) to represent the denatured state ensemble (see Fig. 4 for the expanded state of the 1rx2 crystal). The Elcock expanded state is generated by increasing the van der Waals radii of each atom, causing considerable repulsive interaction throughout the protein, leading to expansion in subsequent energy-minimization steps. After a number of radii expansions (up to 3 Å), the radii are restored so that bond lengths and angles can relax to their equilibrium values. This type of denatured state approximation has two advantages over a dipeptide/tripeptide system, as used in ref. 2: first, the number and type of atoms remain constant, and second, the topology of the protein fold is retained so that atoms that are in close proximity in the native state remain relatively close to each other in the denatured state.

After the expansion, the side-chains are removed (with the exception of proline) to generate the denatured state backbone. Rotamer-backbone energies $\hat{e}_i^{bb}(r)$, rotamer-intrinsic energies $\hat{e}_i^{int}(r)$, and rotamer-rotamer energies $\hat{e}_{ij}(rs)$ in the denatured state are calculated just as described for the native state. A single reference energy $\epsilon_i(a)$ for each residue type a is then obtained by Boltzmann-averaging the rotamer energies (2). Reference energies $\epsilon_{ij}(ab)$ for all residue pairs of types a, b are calculated similarly by using Boltzmann-scaled probabilities. With reference energies in place, conformational energies for rotamers in the native state can be standardized to achieve consistent comparisons between rotamers of different types. Standardized rotamer energy differences $\delta e_i(r)$ (where rotamer r is of type a) are defined as

$$\delta e_i(r) \;=\; e_i^{bb}(r) + e_i^{int}(r) - \epsilon_i(a), \tag{1}$$

and standardized rotamer-rotamer energy differences $\delta e_{ij}(rs)$ (where rotamers r, s are of types a, b) are

expressed as

$$\delta e_{ij}(rs) \;=\; e_{ij}(rs) - \epsilon_{ij}(ab) \tag{2}$$

Therefore, the standardized conformational energy ΔE for a specific rotamer combination can be written as

$$\Delta E_{\text{combination}} \;=\; \sum_{i=1}^{N} \delta e_i(r) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \delta e_{ij}(rs). \tag{3}$$

Identifying the Ensemble Temperature

The temperature parameter T in the ensemble free energy function balances energy and entropy and is chosen so that the entropy of the ensemble matches that of the protein family of interest (Fig. 6*a*). Specifically, average residue entropy, represented by $\bar{S}$, is used as the metric for residue-type variation for both the ensemble and the protein family, and it is calculated by the following expression.

$$\bar{S} \;=\; -\frac{R}{N} \sum_{i=1}^{N} \sum_{a=1}^{20} p_i(a) \log p_i(a) \tag{4}$$

The calculation of $\bar{S}_{\text{ensemble}}$ is straightforward after the completion of the second-order mean-field analysis; however, the calculation of $\bar{S}$ for the protein family presents a challenge due to the inherent sampling limitations of databases [i.e., Pfam (3)]. Because databases only sample a small fraction of the sequence space available, $\bar{S}_{\text{database}}$ is adjusted by calculating $\bar{S}$ for a subset of n randomly selected sequences, as n increases from 2 to the total number of database sequences. The results for $\bar{S}$ are then extrapolated from finite database sizes to $n \to \infty$, allowing the estimation of database entropy $\bar{S}_\infty$ describing a database without sampling limitation. Figure 6*b* shows how $\bar{S}$ increases with increasing sequence set size for the dihydrofolate reductase (DHFR) ($\bar{S}_\infty = 1.6R$) and transformylase ($\bar{S}_\infty = 1.8R$) families. The smaller $\bar{S}_\infty$ for the DHFR family indicates a higher degree of conservation. $\bar{S}_\infty$ is then used as a target for the ensemble calculation. Temperatures are chosen by trial and error so that the resulting $\bar{S}_{\text{ensemble}}$ is near that of $\bar{S}_\infty$ for the appropriate protein family (see flowchart, Fig. 6*c*). For the DHFR and transformylase

families, this temperature is equal to 1,000 and 1,200 K, respectively. It is important to note that the T identified has no physical meaning but rather sets the appropriate balance between energy and entropy in the ensemble.

References

1. Elcock, A.H. (1999) *J. Mol. Biol.* **294,** 1051–1062.

2. Wernisch, L., Hery, S. & Wodak, S.J. (2000) *J. Mol. Biol.* **301,** 713–736.

3. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. & Sonnhammer, E.L. (2002) *Nucleic Acids Res.* **20,** 276–280.