Statistical potentials to assess protein structure and stability

The files localpots.tar.gz and distancepots.tar.gz contain all values of the local and distance potentials described in the article:

"A new generation of statistical potentials for proteins"
Dehouck Y., Gilis D., and Rooman M.
Submitted (2005).

All details about the derivation of these potentials, their significance and the analysis of their performances can be found in this paper. The definitions of the sequence an structure descriptors underlying both local and distance potentials, as well as precisions concerning each of these two types of potentials, are given below.

The potentials, separated into coupling terms (see the above paper for definitions), are given as flat files, with a short description of the columns as remarks (denoted by #) on top of each file. These files are distributed in subdirectories according to the order of the coupling terms (e.g. subdirectory "n3" corresponds to coupling terms of order n=3), and named according to the type of potential (e.g. the files DWaa and DWsds contain the values of the coupling terms $\Delta\tilde{W}_{aa}$ and $\Delta\tilde{W}_{sds}$).

1. Sequence and structure descriptors

- Amino acid type *(s)*: one of twenty amino acids.

The conventional one-letter code is used.

- Backbone conformation *(t)*: one of seven domains on the Ramachandran map [1,2].

The position of each residue on this map is given by the torsion angles of its backbone (ϕ,φ,ω). The seven domains are called A, C, B, P, G, E and O. Domain O groups all *cis*-conformations ($\omega \approx 0°$). The other six domains correspond to *trans*-conformations ($\omega \approx 180°$); A corresponds to α-helical and C to 3_{10}-helical structures, B corresponds to β-like and P to polyproline-like extended conformations and G and E have negative ϕ angles, mirror-symetrical with A/C and B/P. More detail can be found at the address http://babylone.ulb.ac.be/Prelude_and_Fugue/confos.php .

- Solvent accessibility *(a)*: one of five accessibility bins.

The accessibility is defined as the ratio of the solvent accessible surface of the residue in the considered structure (as computed by DSSP [3]) and in an extended tripeptide Gly-X-Gly [4]. These values are grouped in five discrete domains: $a \leq 5\%$ (bin number = 0), $5\% < a \leq 15\%$ (bin number = 1), $15\% < a \leq 30\%$ (bin number = 2), $30\% < a \leq 50\%$ (bin number = 3) and $50\% < a$ (bin number = 4).

- Spatial distance between two residues *(d)*: one of 27 distance bins.

The distance is computed between the average side-chain centroids, noted C^{μ}, of the two residues. The C^{μ} corresponds to the geometric center of heavy side-chain atoms of a given amino acid type, averaged over all side chain conformations in a dataset of known structures [5]. More detail can be found at the address http://babylone.ulb.ac.be/centroids . The first bin groups all distances d comprised between 0 and 3Å. The next 25 bins range from 3 to 8Å, with a 0.2Å width. The last bin includes all distances larger than 8Å. Each bin is denoted by the lower limit of its distance range.

2. Local potentials

These potentials reflect the correlations between the amino acid types *(s)*, the backbone conformations *(t)* and the solvent accessibilities *(a)* of residues close to each other in the sequence. The coupling terms are classified by their order *n*, which corresponds to the number of sequence and structure descriptors taken into account.

- $n = 2$: $\Delta\tilde{W}_{aa}$, $\Delta\tilde{W}_{as}$, $\Delta\tilde{W}_{at}$, $\Delta\tilde{W}_{ts}$ and $\Delta\tilde{W}_{tt}$.
- $n = 3$: $\Delta\tilde{W}_{aaa}$, $\Delta\tilde{W}_{aas}$, $\Delta\tilde{W}_{aat}$, $\Delta\tilde{W}_{ass}$, $\Delta\tilde{W}_{ats}$, $\Delta\tilde{W}_{att}$, $\Delta\tilde{W}_{tss}$, $\Delta\tilde{W}_{tts}$ and $\Delta\tilde{W}_{ttt}$.
- $n = 4$: $\Delta\tilde{W}_{aaaa}$, $\Delta\tilde{W}_{aaas}$, $\Delta\tilde{W}_{aaat}$, $\Delta\tilde{W}_{aass}$, $\Delta\tilde{W}_{aats}$, $\Delta\tilde{W}_{aatt}$, $\Delta\tilde{W}_{asss}$, $\Delta\tilde{W}_{atss}$, $\Delta\tilde{W}_{atts}$, $\Delta\tilde{W}_{attt}$, $\Delta\tilde{W}_{tsss}$, $\Delta\tilde{W}_{ttss}$, $\Delta\tilde{W}_{ttts}$ and $\Delta\tilde{W}_{tttt}$.

The parameter F_{LOC} is set to 8 for all n=2 and n=3 potentials and coupling terms, which implies that no energies are computed for residues separated by more than 8 positions along the sequence. In the case of n=4 coupling terms, F_{LOC} is set to 4. On the basis of the information contained in these files, it is possible to use the potentials with any value of $F_{LOC} \leq 8$ (≤ 4 for n=4 coupling terms) by neglecting the energies corresponding to residues separated by more than F_{LOC} positions along the sequence. The parameter σ, involved in the correction for sparse data, is set to 20.

3. Distance potentials

These potentials reflect the propensity of a pair of residues to be separated by a given spatial distance *(d)*, knowing their amino acid types *(s)*, backbone conformations *(t)* and solvent accessibilities *(a)*. The coupling terms are classified by their order *n*, which corresponds to the number of sequence and structure descriptors taken into account.

- $n = 2$: $\Delta\tilde{W}_{ad}$, $\Delta\tilde{W}_{sd}$, $\Delta\tilde{W}_{td}$.
- $n = 3$: $\Delta\tilde{W}_{ada}$, $\Delta\tilde{W}_{sds}$, $\Delta\tilde{W}_{tdt}$, $\Delta\tilde{W}_{asd}$, $\Delta\tilde{W}_{atd}$, $\Delta\tilde{W}_{tsd}$
- $n = 4$: $\Delta\tilde{W}_{atsc}$.
- $n = 5$: $\Delta\hat{W}_{asdas}$, $\Delta\hat{W}_{atdat}$, $\Delta\hat{W}_{tsdts}$.
- $n = 7$: $\Delta\hat{W}_{atscats}$.

$\Delta\tilde{W}_{atsc}$ and $\Delta\hat{W}_{atscats}$ are contact potentials, which means that only two distance bins are considered : $d \leq 8$Å and $d > 8$Å. In the latter case, all energies are set to 0. Also note that $\Delta\hat{W}_{asdas}$, $\Delta\hat{W}_{atdat}$, $\Delta\hat{W}_{tsdts}$ and $\Delta\hat{W}_{atscats}$ are not equivalent to $\Delta\tilde{W}_{asdas}$, $\Delta\tilde{W}_{atdat}$, $\Delta\tilde{W}_{tsdts}$ and $\Delta\tilde{W}_{atscats}$. $\Delta\hat{W}_{asdas}$, for example, is defined as the sum of $\Delta\tilde{W}_{asdas}$ and all the lower order asymmetrical 2-body terms. More precisely, we have :

- $\Delta\hat{W}_{asdas} = \Delta\tilde{W}_{ads} + \Delta\tilde{W}_{asda} + \Delta\tilde{W}_{asds} + \Delta\tilde{W}_{asdas}$.
- $\Delta\hat{W}_{atdat} = \Delta\tilde{W}_{adt} + \Delta\tilde{W}_{atda} + \Delta\tilde{W}_{atdt} + \Delta\tilde{W}_{atdat}$.
- $\Delta\hat{W}_{tsdts} = \Delta\tilde{W}_{tds} + \Delta\tilde{W}_{tsdt} + \Delta\tilde{W}_{tsds} + \Delta\tilde{W}_{tsdts}$.
- $\Delta\hat{W}_{atscats} = \Delta\tilde{W}_{atcs} + \Delta\tilde{W}_{asct} + \Delta\tilde{W}_{tsca} + \Delta\tilde{W}_{atsca} + \Delta\tilde{W}_{atsct} + \Delta\tilde{W}_{atscs} + \Delta\tilde{W}_{atcas} + \Delta\tilde{W}_{atcts} + \Delta\tilde{W}_{ascts} + \Delta\tilde{W}_{atscat} + \Delta\tilde{W}_{atscas} + \Delta\tilde{W}_{atscts} + \Delta\tilde{W}_{atscats}$.

The parameter F_{DIST} is set to 8 for all distance potentials and coupling terms, which implies that the relative positions along the sequence are explicitly taken into account only for pairs of residues separated by ≤ 8 positions along the sequence. All other pair of residues are grouped in a single class. The parameter σ, involved in the correction for sparse data, is set to 10.

References

1. Ramachandran, G., Sasiekharan, V. (1968) *Adv. Prot. Chem.* **23**, 283-437.
2. Rooman, M.J., Kocher, J.-P.A., Wodak, S.J. (1991) *J. Mol. Biol.* **221**, 961-979.
3. Kabsch, W., Sander, C. (1983) *Biopolymers* **22**, 2577-2637.
4. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H. (1985) *Science* **229**, 834-838.
5. Kocher, J.-P., Rooman, M.J., Wodak S.J. (1994) *J. Mol. Biol.* **235**, 1598-1613.