Statistical Potentials

part of “Bioinformatik von RNA- und Proteinstrukturen”

Sonja Prohaska

Computational EvoDevo
University Leipzig

Leipzig, SS 2011

Protein Structure – was bisher geschah

- **Primary structure**: sequence of amino acids
- **Secondary structure**: β-sheets and α-helices
- **Ternary structure**: arrangement of secondary structure elements into folds
- prefered conformations of an amino acid sequence are thoseof low energy
- physical, chemical and **thermodynamical rules** define a energy function or force field
- the force field can be used to predict protein structure from sequence

NEW!

A **statistical potential** can be derived from known structures and probabilistic theory.

Force Field

Thermodynamic Potential: Assisted Model Building with Energy Refinement (AMBER)

$$E^{total} = \sum E_{ij}^{bond} + \sum E_{ijk}^{angle} + \sum E_{ijkl}^{torsion} + \sum E_{ij}^{waals} + \sum E_{ij}^{coulomb}$$

Statistical Potential:

$$E^{total} = -kT \sum_{s} \ln \frac{P(s)}{P_R(s)}$$

Statistical Potential or Knowledge-Based Potential

Take proteins from Protein Data Bank (PDB), derive the probability $p(\phi, \psi)$ for the torsion angles ϕ and *psi* of each amino acid.

$$E(\phi, \psi) = -c \log \frac{p(\phi, \psi)}{p(\phi)p(\psi)} \tag{1}$$

results in a statistical potential resembling a Ramachandran plot.

Classic Application for Statistical Potentials

Statistical potential for pairwise amino acid contacts

- an interaction matrix assigns a weight or energy value to each possible pair of amino acids
- the values are determined using statistics on amino acid contacts in known proteins
- the energy of a structural model is the combined energy of all pairwise contacts (i.e. amino acids within a certain distance to each other)

Potential of Mean Force (PMF) by Sippl

The frequencies of amino acid interactions in dependence of their distance r can be transformed in a potential of mean force with the help of Bolzmann's law.

$$f(r) = \frac{1}{Z} e^{-\frac{E(r)}{kT}} \tag{2}$$

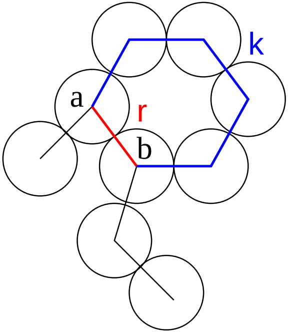
Here, k is the Bolzmann constant, T is the temperature, and

$$Z = \int e^{-\frac{E(r)}{kT}} \, \mathrm{d}r \tag{3}$$

is the partition function (over all distances r).
The free energy $E(r)$ can be given as a function of the probability $f(r)$ by

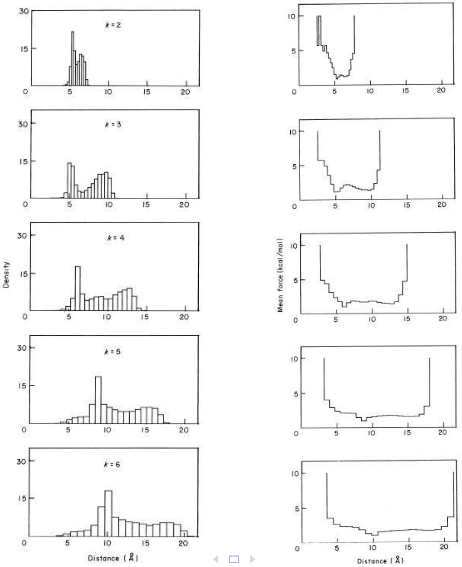
$$E(r) = -kT \ln f(r) - kT \ln Z \tag{4}$$

Transformation of $f(r)$ into $E(r)$



distance r in A^{o}

distance k in aa



Unfortunately, Z cannot be measured.

However, $E(r)$, derived from interval sampling, is an average over all conformations (and all their interactions). It can be used as a “reference state” when calculating the free energy contribution for a particular pair of amino acids ab at sequence distance k.

$$E_k^{ab}(r) = -kT \ln f_k^{ab}(r) - kT \ln Z_k^{ab} \tag{5}$$

$$\Delta E_k^{ab}(r) = E_k^{ab}(r) - E_k(r) \tag{6}$$

$$\Delta E_k^{ab} r = -kT \ln \frac{f_k^{ab}(r)}{f_k(r)} - kT \ln \frac{Z_k^{ab}}{Z_k} \tag{7}$$

Z_k^{ab} and Z_k cannot be obtained, however, they are constant and hence $-kT \ln(Z_k^{ab}/Z_k)$ does not depend on the variable r. Furhtermore, T corresponds to the average temperature at structure determination. It is set to 293K.

$$\Delta E_k^{ab} r = -kT \ln \frac{f_k^{ab}(r)}{f_k(r)} \tag{8}$$

The problem of small data sets – part I

We need potentials of mean force $E_k^{ab}(r)$ for amino acid distances in 1D for $1 <= k <= k_{max}$ and all possible pairs *ab* of amino acids. We need probability distributions for $k_{max} \times 20 \times 20 = 400k_{max}$ interactions.

The data set is too small to estimate $f_k^{ab}(r)$ accurately.

The problem of small data sets – part II

- set $f_k^{ab}(r)$ to $f_k(r)$ (in this case $E_k^{ab}r = 0$)
- iterpretation: if we don't know any specifics about interaction *ab* we assume it is average
- *a* and *b* of a single pair can be in distance r to each other ($\delta(r) = 1$) or not ($\delta(r) = 0$)
- a single measurement distorts $f_k(r)$ as follows

$$f'_k(r) = \frac{1}{z}(f_k(r) + \sigma\delta(r)) \tag{9}$$

- here $z = 1 + \sigma$ and σ is the weigth of the information $\delta(r)$
- for m measurements $f'_k(r)$ aproaches $f_k^{ab}(r)$

$$f_k^{ab}(r) \approx \frac{1}{1 + m\sigma} f_k(r) + \frac{\sigma}{1 + m\sigma} \sum_{i=1}^{m} \delta_i(r) \tag{10}$$

The total free energy difference (compared to the average) of a protein, ΔE_t, is claimed to be the sum over all pairwise free energies.

$$\Delta E_t = \sum_{i<j} \Delta E_k^{a_i a_j}(r_{ij}) \qquad (11)$$

Summary

potential difficulties

- interpretation of this "potential" as a true, physically valid potential of mean force
- nature of the reference state and its optimal formulation
- validity of generalizations beyond pairwise distances
- not accurate enough for protein structure prediction

Knowledge-base potential have been applied sucessfully in the prediction of protein in fold recognition.
Knowledge-base potentials can be used to derive energy parameters from known RNA structures for the prediction of secondary structures.

Literature

Manfred J. Sippl (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213: 859-883.