BCB 569: Bioinformatics III

Lecture 7: Knowledge-based Energy Functions

Guang Song

Why energy functions are important

- Anfinsen (Nobel Prize in chemistry, 1972)
  - a protein's primary structure has all the information it needs to fold to the native state
- Thermodynamic hypothesis of protein folding
  - the native state has global minimum free energy
- Implication: free energy can be used to locate the native state, the tertiary structure

Why energy functions are important

- In protein structure prediction, structure refinement, and molecular docking, there are two essential components
  - conformation sampling
  - scoring function
- The challenge in sampling is to be able to generate conformations close to the native state
- The challenge in scoring is that, given that those conformations are sampling, can you pick them out?

Why knowledge-based potentials

- Physics-based (or semi-empirical)
  - the parameters are based on fitting experimental data for small molecules or quantum-mechanics calculations
- Knowledge-based
  - parameters are based on statistical analysis of a database of known structures

- Physics-based potentials are computationally more expensive

- Knowledge-based potentials are much simpler, and can yield results comparable to semi-empirical potentials in structure predictions, fold recognition, docking and binding

Knowledge-based Potentials

- Statistical potential
  - based on statistical analysis of a dataset of known structures
  - the potential of an interacting pair is determined by its relative frequency in the database
- Optimization-based
  - the sets of parameters for potential functions are optimized based on some criterion
    - e.g., by maximizing the energy gap between known native state and a set of “decoy” conformations

Statistical Potentials (SPs)

- Three main ingredients
  - a protein descriptor
  - a function form of the potential function
  - a method to derive the values of the parameters

- SP categories
  - distance-independent vs. distance-dependent
  - residue-level vs. atomic level
  - orientation-dependent vs. orientation-independent

Protein descriptor

- Protein descriptor
  - a description of the shape of a protein that best characterizes its features
- Example:
  - First, we define two residues are in contact if their side chain center are within, say, 6 angstroms
  - Given a conformation, there are many pairs of residues are in contact
  - one possible protein descriptor is total number of contacts

A better descriptor

- Since there are 20 types of amino acid and 210 types of amino acid pairs
- Another descriptor would be to count how many contacts for each type of amino acid pairs
  - the descriptor is a vector: $[p_{1,1}, p_{1,2}, \ldots p_{20,20}]$
  - For residues that like to form contacts, such as hydrophobic residues, we thus expect a good conformation will give higher frequency for such types of contacts

Function form

- The easiest and mostly used form for energy function is a linear function

$$F([p_{1,1}, p_{1,2}, \ldots p_{20,20}]) = \sum \sum e_{i,j} * p_{i,j}$$

- As soon as the weights $e_{i,j}$ are determined, the potential is known
  - $e_{i,j}$ is the contact energy for amino acid pair pi

Deriving the parameters

- The contact energy $e_{i,j}$ is normally derived by characterizing the frequency distributions of the structural descriptors
  - e.g., how often two Cystine residues are in contact
- A database of experimentally determined structures are used
  - such as Protein Data Bank (PDB)

Statistical Potentials: Background

- In statistical potentials, the observed frequency of various structure features are converted to effective free energy (or potential of mean field)

- Based on the assumption that frequently observed features corresponds to low free energy state
  - Boltzmann distribution

Boltzmann Distribution

- Boltzmann distribution basically says that a particle has a higher propensity to stay in lows energy states
  - The probability is proportional to e^(-E/KT), where E is the energy of the state, T is temperature, K is the Boltzmann constant
-

$$\frac{N_i}{N} = \frac{g_i e^{-E_i/k_B T}}{Z(T)}$$

Therefore, the free energy of a state is proportional to $-\ln(N_i)$ => high frequency means low energy

Miyazawa-Jernigan Potential

- References
  - Miyazawa & Jernigan (1985)
  - Miyazawa & Jernigan (1996)

Miyazawa-Jernigan Potential

- MJ potential is a *distance-independent residue-based* statistical potential

- The protein descriptor: $[p_{1,1}, p_{1,2}, \ldots p_{20,20}]$
  - the frequency of residue contacts of each type of amino acid pair
- The energy function:

  $F([p_{1,1}, p_{1,2}, \ldots p_{20,20}]) = \sum \sum e_{i,j} * p_{i,j}$

- $e_{i,j}$: the free energy change when residues of type i and j form a contact

determine the contacts

- Given a conformation, all the atom coordinates are known
- For each residue, compute the coordinates of its side chain center
- Compute the pairwise distances of all side chain centers, for residue pairs whose center side centers are within 6.5 angstrom, consider them in contact
- group the contacts based on the type of amino acid pairs and count the frequencies

Determine $e_{i,j}$

- First, look at the following chemical reaction (o denotes solvent):

  i-O + j-O ⇔ i-j + O-O

- $e_{i,j}$ is the free energy change after forming contacts between residue of type i and type j
  - Therefore, $e_{i,j} = -\ln([m_{ij}][m_{oo}]) / [m_{io}][m_{jo}])$
  - $[m_{ij}]$ is the numbers of i-j contacts

- Thus, if we can estimate the values of $[m_{ij}]$, $[m_{oo}]$, $[m_{io}]$, and $[m_{jo}]$, we can determine $e_{i,j}$
  - These values can be estimated from some statistical analysis of a database

What the potential looks like

Table 3. Contact energies in RT units; ε_{ij} for upper half and diagonal and ε_{ji} for lower half

	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	
Cys	−5.44	−4.99	−5.80	−5.50	−5.83	−4.96	−4.95	−4.16	−3.57	−3.16	−3.11	−2.86	−2.59	−2.85	−2
Met	0.46	−5.46	−6.56	−6.02	−6.41	−5.32	−5.55	−4.91	−3.94	−3.39	−3.51	−3.03	−2.95	−3.30	−2
Phe	0.54	−0.20	−7.26	−6.84	−7.28	−6.29	−6.16	−5.66	−4.81	−4.13	−4.28	−4.02	−3.75	−4.10	−3
Ile	0.49	−0.01	0.06	−6.54	−7.04	−6.05	−5.78	−5.25	−4.58	−3.78	−4.03	−3.52	−3.24	−3.67	−3
Leu	0.57	0.01	0.03	−0.06	−7.37	−6.48	−6.14	−5.67	−4.91	−4.16	−4.34	−3.92	−3.74	−4.04	−3
Val	0.52	0.18	0.10	−0.01	−0.04	−5.52	−5.18	−4.62	−4.04	−3.38	−3.46	−3.05	−2.83	−3.07	−2
Trp	0.30	−0.29	0.00	0.02	0.08	0.11	−5.06	−4.66	−3.82	−3.42	−3.22	−2.99	−3.07	−3.11	−2
Tyr	0.64	−0.10	0.05	0.11	0.10	0.23	−0.04	−4.17	−3.36	−3.01	−3.01	−2.78	−2.76	−2.97	−2
Ala	0.51	0.15	0.17	0.05	0.13	0.08	0.07	0.05	−2.72	−2.31	−2.32	−2.01	−1.84	−1.89	−1
Gly	0.68	0.46	0.62	0.62	0.65	0.51	0.24	0.20	0.18	−2.24	−2.08	−1.82	−1.74	−1.86	−1
Thr	0.67	0.28	0.41	0.30	0.40	0.36	0.37	0.13	0.10	0.10	−2.12	−1.96	−1.88	−1.90	−1.
Ser	0.69	0.53	0.44	0.59	0.69	0.55	0.38	0.14	0.18	0.14	−0.06	−1.67	−1.58	−1.49	−1.
Asn	0.97	0.62	0.72	0.87	0.79	0.77	0.30	0.17	0.36	0.22	0.02	0.10	−1.68	−1.71	−1.
Gln	0.64	0.20	0.30	0.37	0.42	0.46	0.13	−0.12	0.24	0.24	−0.08	0.11	−0.10	−1.54	−1.
Asp	0.91	0.77	0.75	0.71	0.80	0.89	0.30	−0.07	0.26	0.13	−0.14	−0.19	−0.24	−0.09	−1.
Glu	0.91	0.30	0.52	0.46	0.56	0.55	0.00	−0.25	0.30	0.35	−0.22	−0.19	−0.21	−0.19	0
His	0.65	0.28	0.39	0.66	0.67	0.70	0.08	0.09	0.47	0.56	0.16	0.26	0.29	0.31	−0
Arg	0.93	0.38	0.42	0.41	0.43	0.47	−0.11	−0.30	0.30	0.18	−0.07	−0.01	−0.02	−0.26	−0
Lys	0.85	0.31	0.33	0.32	0.37	0.33	−0.10	−0.46	0.11	0.08	−0.19	−0.15	−0.30	−0.46	−1
Pro	0.53	0.16	0.25	0.39	0.35	0.31	−0.33	−0.25	0.20	0.15	0.04	0.14	0.18	−0.06	0

How to use the potential

- Now for a given conformation of protein
  - we can first determine its corresponding protein descriptor vector: $[p_{1,1}, p_{1,2}, \ldots p_{20,20}]$
  - Then, its energy should be:

$$F([p_{1,1}, p_{1,2}, \ldots p_{20,20}]) = \sum \sum e_{i,j} * p_{i,j}$$

Distance-dependant potentials

- It may desirable to add another feature to the previous protein descriptor
- We may not only consider the type of amino acid pairs that are in contact, but also their separation distance
- This can be used to determine a more sensitive distance-based potentials

The DFIRE potential

- DFIRE is distance-dependent atom-based statistical potential [Zhou and Zhou, Protein Science, 11:2714, 2002]
- The contact potential between a pair of atoms is

$$\bar{u}(i,j,r) = -RT \ln \frac{N_{obs}(i,j,r)}{N_{exp}(i,j,r)}$$

R is the gas constant, *T* is the temperature,

Nobs(*i,j,r*) is the observed number of atomic pairs (*i,j*) within a distance shell $r - \Delta r/2$ to $r + \Delta r/2$ in a database of folded structures

Nexp(*i,j,r*) is the expected number of atomic pairs (*i,j*) in the same distance shell if there were no interactions between atoms (the reference state).

- There are a number of distance-dependent potentials. They differ mainly on how the reference state is selected

The DFIRE potential

- Now let us go through the derivation of the DFIRE potential

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2373736/?tool=pubmed

The dDFIRE

- The dDFIRE is an extended version of DFIRE with an orientation-dependence
  [Yang and Zhou, Proteins 72:793, 2008]

- Atoms are classified as polar or non-polar
- Polar atoms are treated as dipoles
- The potential is a function of the following form:

$$\overline{u}^{\mathrm{dDFIRE}}(r_{pq}, \theta_p, \theta_q, \theta_{pq})$$

Optimization method

- Another category of knowledge-based potential is based on optimization
- We may have a similar form of potential function as before, but the parameters are optimized according to some criteria
  - for example, the parameters may be set so that there is a energy gap between the native state and the “decoy” conformations
- Techniques used for optimization
  - support vector machine, neural network, etc

Recommended Readings

- Chapter 3 of the textbook by Xu, Xu, and Liang

- Miyazawa & Jernigan, Residue-Residue Potentials with ... *Journal of Molecular biology, (1996), 256, 623-644*
- DFIRE: distance-dependent atom-based statistical potential [Zhou and Zhou, Protein Science, 11:2714, 2002]
- dDFIRE: orientation-dependent potential [Yang and Zhou, Proteins 72:793, 2008]