6.581J / 20.482J
Foundations of Algorithms and Computational Techniques in Systems Biology
Professor Bruce Tidor
Professor Jacob K. White

6.581/ 20. 482J Problem Set #1

Due 5pm, Thursday, 24 February 2006

Discrete conformational search

Consider a system with N variable residues, each of which has M_i possible conformations. We can denote any single configuration of the system by the set $\{m\} = \{m_1, m_2, ..., m_N\}$, where each m_i is the conformation adopted by residue i. The energy of any given configuration can be computed by:

$$E_{\{m\}} = \sum_{i=1}^{N} E^{\text{self}}_{m_i} + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} E^{\text{pair}}_{m_i,m_j} \tag{1}$$

The energy terms can be computed in many ways, but are provided as a given in this problem. In this problem you will explore the configuration space of a system with 5 variable positions, each of which may adopt 306 possible conformations. The file `pair.dat` is a plain text, tab-delimited file, containing the energies as a matrix of the following form:

$$\begin{array}{ccccc}
\mathbf{E}^{\text{self}}_1 & \mathbf{E}^{\text{pair}}_{1,2} & \mathbf{E}^{\text{pair}}_{1,3} & \cdots & \mathbf{E}^{\text{pair}}_{1,N} \\
\mathbf{0} & \mathbf{E}^{\text{self}}_2 & \mathbf{E}^{\text{pair}}_{2,3} & \cdots & \mathbf{E}^{\text{pair}}_{2,N} \\
\mathbf{0} & \mathbf{0} & \mathbf{E}^{\text{self}}_3 & \cdots & \mathbf{E}^{\text{pair}}_{3,N} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{E}^{\text{self}}_N
\end{array} \tag{2}$$

where each sub-matrix $\mathbf{E}^{\text{self}}_i$ is an $M_i \times M_i$ matrix of the form:

$$\begin{array}{ccccc}
E^{\text{self}}_{m_i^1} & 0 & 0 & \cdots & 0 \\
0 & E^{\text{self}}_{m_i^2} & 0 & \cdots & 0 \\
0 & 0 & E^{\text{self}}_{m_i^3} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & E^{\text{self}}_{m_i^{M_i}}
\end{array} \tag{3}$$

and each sub-matrix $\mathbf{E}^{\text{pair}}_{i,j}$ is an $M_i \times M_j$ matrix of the form:

$$\begin{array}{ccccc}
E^{\text{pair}}_{m_i^1,m_j^1} & E^{\text{pair}}_{m_i^1,m_j^2} & E^{\text{pair}}_{m_i^1,m_j^3} & \cdots & E^{\text{pair}}_{m_i^1,m_j^{M_j}} \\
E^{\text{pair}}_{m_i^2,m_j^1} & E^{\text{pair}}_{m_i^2,m_j^2} & E^{\text{pair}}_{m_i^2,m_j^3} & \cdots & E^{\text{pair}}_{m_i^2,m_j^{M_j}} \\
E^{\text{pair}}_{m_i^3,m_j^1} & E^{\text{pair}}_{m_i^3,m_j^2} & E^{\text{pair}}_{m_i^3,m_j^3} & \cdots & E^{\text{pair}}_{m_i^3,m_j^{M_j}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
E^{\text{pair}}_{m_i^{M_i},m_j^1} & E^{\text{pair}}_{m_i^{M_i},m_j^2} & E^{\text{pair}}_{m_i^{M_i},m_j^3} & \cdots & E^{\text{pair}}_{m_i^{M_i},m_j^{M_j}}
\end{array} \tag{4}$$

Note that two additional files are provided `pair-tiny.dat` and `pair-small.dat`. The former desribes a system of 3 residues, each with 135 possible conformations; the latter a system of 4 residues with 216 possible conformations. The tiny file may be useful in debugging your code. The small file can be used if you are working on a small computer with not much memory and are unable to use the plain `pair.dat`, but we would rather you not take this route.

1. How many operations are required to evaluate the energy of a single configuration, given that the set of E^{self} and E^{pair} are precalculated. Calculate the total number of possible configurations of

this system. How many operations would be required to find the global minimum configuration through enumeration of configurational states.

2. The dead-end elimination (DEE) algorithm provides a powerful tool for reducing the complexity of this discrete search. In its simplest form, the DEE algorithm involves two statements of "eliminating power". The first applies to individual residue conformations, and the second to pairs of residue conformations.

   Singles elimination criterion: A given conformation m_i^s for position i can **not** be present in the global mimimum solution if:

$$E^{\text{self}}_{m_i^s} + \sum_{j=1}^{N} \min_u (E^{\text{pair}}_{m_i^s, m_j^u}) > E^{\text{self}}_{m_i^t} + \sum_{j=1}^{N} \max_u (E^{\text{pair}}_{m_i^t, m_j^u}) \tag{5}$$

   for **any** conformation $t \neq s$ at position i $(i \neq j)$.

   Pairs elimination criterion: A given pair of conformations m_i^s, m_j^t for positions i and j can **not** *simultaneously* be present in the global mimimum solution if:

$$\begin{gathered}(E^{\text{self}}_{m_i^s} + E^{\text{self}}_{m_j^t} + E^{\text{pair}}_{m_i^s, m_j^t}) + \textstyle\sum_{k=1}^{N} \min_w (E^{\text{pair}}_{m_i^s, m_k^w} + E^{\text{pair}}_{m_j^t, m_k^w}) > \\ (E^{\text{self}}_{m_i^u} + E^{\text{self}}_{m_j^v} + E^{\text{pair}}_{m_i^u, m_j^v}) + \textstyle\sum_{k=1}^{N} \max_w (E^{\text{pair}}_{m_i^u, m_k^w} + E^{\text{pair}}_{m_j^v, m_k^w})\end{gathered} \tag{6}$$

   for **any** pair of conformations $u \neq s$ and $v \neq t$ at positions i and j $(i \neq j)$.

   The general implementation of the DEE algorithm applies these criteria iteratively:

```
for i=1:iterations
  eliminate singles
  eliminate pairs
end
```

   with the number of single residues, and of pairs of residues, decreasing with each iteration. The number of iterations may be set to a fixed value, or may be implemented as a `while` loop, continuing until no more singles or pairs are eliminated.

   Implement the DEE algorithm in MATLAB or any programming language of your choice. Plot the number of possible conformations as a function of iteration number. How many iterations are required to make enumeration of the remaining conformations feasible? Perform this enumeration to find the global energy minimum configuration.

3. An alternative approach to the conformational search problem is the mean-field approach. In this method, all conformational states of every residue are considered simultaneously, with a probabilistic description of the relative populations of each state. In this model, the energy of the system is given by:

$$E_{\text{eff}} = \sum_{i=1}^{N} \sum_{s=1}^{M_i} (P(m_i^s) E^{\text{self}}_{m_i^s}) + \sum_{i=1}^{N-1} \sum_{s=1}^{M_i} \sum_{j=i+1}^{N} \sum_{t=1}^{M_j} (P(m_i^s) P(m_j^t) E^{\text{pair}}_{m_i^s, m_j^t}) \tag{7}$$

   where $P(m_i^s)$ is the probability of residue i being found in conformation s.

   The key goal of the algorithm is to obtain a set of probabilities that accurately describes the low energy configurations of the system. One method to do this is using a self-consistent approach (known as self-consistent mean-field, SCMF) — first assigning an initial probability distribution, then using this starting point to iteratively refine the probabilities until some level of convergence is reached. The update of each probablity is achieved using the relation:

$$P(m_i^s) = \frac{e^{\frac{-E(m_i^s)}{kT}}}{\sum_{t=1}^{M_i} e^{\frac{-E(m_i^t)}{kT}}} \tag{8}$$

where $E(m_i^s)$ is the energy of conformation s of residue i (m_i^s) in the "mean-field" of all conformations at other residues. This is given by:

$$E(m_i^s) = E^{\text{self}}_{m_i^s} + \sum_{j=1}^{N} \sum_{t=1}^{M_j} (P(m_j^t) E^{\text{pair}}_{m_i^s, m_j^t}) \tag{9}$$

Note that the energies provided are in kcal/mol, T should be in degrees Kelvin (K), and thus the Boltzmann constant $k = 1.987 \times 10^{-3}$ kcal/(mol·K).

Given these equations, we write the self-consistent mean-field algorithm as follows:

```
select initial probabilites
for i=1:iterations
  compute mean-field energies for each m(i,s)
  compute probability distribution
end
```

Implement the self-consistent mean-field approach and solve for the probability distribution, and its associated energy. Use a uniform probability density ($P(m_i^s) = 1/M_i$) as an initial guess, and a temperature of 298 K. How do your results change if you use a temperature of 100 K or of 1000 K (plot the distribution for each temperature)?

4. Only at very low temperatures will SCMF give a unique configuration, comparable to the result from DEE. Refine your results from the run with $T = 298$ K by succesively reducing the temperature, using the final probabilities from the previous run as initial conditions. Use the temperature set $\{298, 100, 10, 1\}$ K. Plot the distribution after each stage, and compare the final result to the answer from DEE.

Comments

1. As this assignment requires fair amount of programming, I would highly recommend people with less programming experience to start working on problems early. You can code in matlab or your favourite programming language. Please note that due to matlab loop overhead, a matlab implementation of Dead End Elimination algorithm (on a small dataset) may take hours to run, whereas a C++ implementation would run in seconds. On the other hand, it is very convenient to perform matrix manipulations with matlab. If you do decide to go ahead with matlab, it'll be fine to answer the questions using the smaller dataset.

2. You will submit all your work via MIT server . Please tar and zip your files. I would highly recommend you to include a simple README, which clearly describes each file in your directory. It is a lot easier for us to grade your assignment if you document your code, and use meaningful variable and function names.