How do proteins fold?

Biology offers many opportunities for doing interesting physics. Proteins are important players in our cells, and their functionality depends strongly on their structure. This structure in turn depends on the sequence of amino acids. So, how do we get from sequence to structure? In this project you will attempt to answer this question.

One part of this exercise will be a competition where the goal is to find the protein configuration that has minimal energy for a set of proteins. This will be a local version of an existing biannual competition for researchers working on protein folding algorithms (http://predictioncenter.org/casp9/).

You will learn: Monte Carlo techniques. Global optimization.

Suggested reading: G&N chapter 8 and 12.1.

The Model

The model to be investigated is a simplistic protein model where only two different amino acids exist (hydrophobic (H) and hydrophilic (P)), and where the spatial localization of the amino acids is on a two-dimensional lattice (as described in the lecture notes).

Since the pivot moves (suggested protein updates) and their relation to self-avoidance is not trivial, a c-code with an example implementation is provided at the course webpage. It is allowed to call these functions directly or translate the functions into your favorite programming language. Note also the data structure used, which might be useful (x-position, y-position, and sequence vectors and an occupancy matrix for the lattice).

Physical Considerations (some examples)

Several aspects of protein folding can be investigated using Monte Carlo simulations of the HP-model (see e.g. Chapter 12.1 in G&N). Some examples are to investigate average end-to-end distances or energy as a function of temperature for a given sequence (e.g. HHHHP HPHPH PPHHP HHH). Another test could be to investigate the self-avoiding properties of a chain of all P's (always E=0), and how this scales with sequence length.

The Competition

The goal of the competition is to find configurations with minimal energy for the following three sequences of length 25:

HHPHH PHPPH HPHPP HPPPH HPHPH
PHHHH PHHHP PPPPH HHPHH HPHHH
PHHHP PPHPP HHPHH PPPPP HPPHH

Although Simulated Annealing is suggested as an optimization algorithm, any algorithm is allowed. Note that it might be interesting to also investigate how stable the performance is, i.e. how often is the optimal configuration found if the algorithm is restarted, and how does this compare for different annealing rates, and for a local search algorithm (T=0).

The report should include:

- Brief introduction to the protein folding/HP-model.
- Description and motivation of the algorithm(s) used.
- Results of simulations (and possible measurements).
- Optimal energy configurations (found) for the competition sequences.
- Conclusions
- Appendix: include a listing of your program. Also, send your program by e-mail to henrik@thep.lu.se.

Deadline for the project is 23:59:59, 2010-10-22