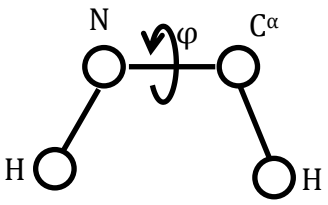
Revision – Free energy calculations, unusual Monte Carlo and Molecular Dynamics

SS 2017 Übung zu Struktur und Simulation, 27-Jun-2017

These are typical exam questions for discussion on 10 or 11 July.

1. I run an MD simulation of a protein without a temperature bath and find the temperature to be 310 K. I want to simulate at 300 K. I want to adjust the velocities to have the correct temperature. What factor should I multiply the velocities by ? Do not use a calculator. Give an answer as an exact expression.
2. I have the coordinates for a protein. At a certain site, it has an alanine. I would like to predict the stability if this alanine is mutated to a glycine. I suggest that we run an MD simulation of the native protein in water and the mutant in water. I then look at the average potential energy that my MD program prints. Why is this meaningless ?
3. I claim that the stability of the protein is given by ΔG_{fold}, which is the free energy change as the protein goes from the unfolded state(s) to the native state(s). When I try with either the alanine or glycine protein, the protein never folds, so I cannot estimate eitherΔG_{fold}^{ala} orΔG_{fold}^{gly}. Describe exactly what you could simulate in order to find the difference in stability of the native (ala) and mutant (gly) proteins.
4. I have the crystal structure of a protein bound to a small molecule. In the binding site, there is an alanine. I want to predict the change in binding energy if I mutate the alanine to a glycine. It is proposed to do a perturbation/free energy calculation by gradually changing the alanine to a glycine, integrating over the total energy (Hamiltonian) as I change the system.
   Why will this not give me the correct free energy difference ?
   Write a set of free energy differences which should, in principle, give me the correct answer.
5. I do not know whether to use simple Monte Carlo or simulated annealing when simulating a protein. Which would I use to find the density at 300 K ?
6. I have invented a physical measurement which detects H-bonds between atoms in a protein. In my protein simulation, I do not see the H-bonds being formed. I would like to push the simulation so it forms the measured H-bonds. Describe a simple quasi-energy term which would serve this purpose.

7. From NMR spectroscopy, I can use J-coupling constants to estimate the angle φ. The angle is related to the measured J value by $J = A\cos^2\varphi + B\cos\varphi + C$ for some constants A, B and C.



How could one implement an artificial (pseudo-) energy term which would persuade a protein to change its conformation and reflect the measured J coupling ?

8. I would like to use Monte Carlo / simulated annealing to optimize a protein sequence without disturbing the structure (design). I have a function $U(\mathbf{R}, S)$ which acts on the set of coordinates, **R**, and S the ordered set of residues and returns an estimate of the free energy of protein folding. Describe a Monte Carlo scheme which will find sequences that are compatible with a structure.

9. I have a drug "D" which binds to a protein "P". By every experimental method known to man, the binding appears instantaneous. It is suggested that I can mix the drug and protein, measure the concentrations and estimate the free energy of binding by saying $\Delta G = RT \ln \frac{[D][P]}{[DP]}$. If the binding / unbinding is really instantaneous, why may this be meaningless ?

10. I am simulating with a protein in a box of water with periodic boundary conditions using conventional molecular dynamics and with a cutoff for interactions of 15 Å. In "big O" notation, what is the running time ? Does the worst case running time change if I do not use cutoffs ? Does the running time change in practice ?

11. Exam examples only – no need for discussion

    a. Write down the energy due to electrostatics – use any reasonable choice of nomenclature.

    b. Write down the energy due to electrostatics if we use a distance dependent dielectric constant.

    c. Write down the force acting on a charge due to electrostatics if we use a distance dependent dielectric constant. Write it simply as a function of the distance between two particles (no need for vectors). Use nomenclature consistent with the first two parts of this question.

12. Describe a water model based on solvent accessible surface area. Explain in terms of simple equations why this will lead to a force acting on particles such as oxygen or nitrogen atoms. Describe an effect of water which will not be represented by this model (there are several sensible answers).

13. I am simulating a protein in a box with an explicit SPC water model (3 point charges, 3 mass sites and 1 Lennard-Jones site). An argon atom approaches the water. Later, a charged sodium ion

approaches the water. From the points of view of the argon atom and sodium ion, is the water spherical / isotropic (like a football / billiards ball) or some different shape ?

14. Exam example question with many possible answers:
    Describe a low-resolution protein model which can

    a. keep adjacent (successive) amino acids at the correct distance

    b. form regular α-helices and β-sheets

    c. has some mechanism to maintain roughly the correct density within the middle of the protein

Explain what you choose as interaction sites. Explain which features of your model are necessary for each feature.

15. I have built a score function $S(r_{ij})$ for proteins based on the frequency with which certain pairs of amino acids are seen at certain distances. It allows me to give a score to any conformation of a protein by looking up scores in tables.

    - I am not able to use it for molecular dynamics simulations. Why ?
    - I try to make a continuous form of the functions by fitting the data to a polynomial function like $S(r_{ij}) = k_1 + k_2x + k_3x^2 + k_4x^3$
      Why would a sane person do this ? Where may it break down ?

16. I have built a Boltzmann / knowledge-based score function for proteins using the methodology based on potentials of mean force. It is based on C^{α}-C^{α} distances. I do not distinguish between amino acids which are separated by one residue (i, i+2) and those separated by many residues. Why will this be a very bad approximation ?

17. I am working with a lattice model for a protein. The model is simple so I can computationally visit all conformations. Describe how I would use the Boltzmann relation to work out the absolute probability for a certain configuration of points for a given sequence.

18. Explain why I could not do this with a continuous model for a molecule.

19. I would like to investigate protein conformational switches using a lattice model. These are proteins which can adopt more than one conformation. The hamburger Abendblatt has claimed that the protein switches are more likely to be found in proteins with less hydrophobic residues. Describe a set of steps to see if this is plausible. Describe calculations for a protein of length 18 in the HP model.

20. A popular view of protein folding is that as a protein folds, its potential energy decreases, but the entropy also decreases. How would you see this be reflected in a simple lattice based model and how could you check if it is true ?

21. Genetic algorithms a general optimization method. You have a population of n solutions. For some number of generations, you select the least fit individuals and remove them. From the remaining individuals, you apply "mutations" and "crossovers". You then copy individuals to bring the number back up to n.

    If you are working with protein coordinates, you may have the internal angles stored in an array and you have n copies of the array. A mutation may mean you change an internal angle. If you are working with binding a rigid ligand to a protein, a mutation might consist of a small change in the ligand coordinates. Consider the example of binding a ligand to a protein

    Generate n random poses (orientations of drug + protein). This is the initial set $\{\mathbf{m}\}$.

```
while (not_finished)
          foreach member m of {m}
          apply a small random change to m
     calculate energy of each of the n poses
     rank set of candidates
     discard  ½ of the candidates with the highest energy.
     duplicate the remaining candidates
```

    a. Why will this lead to low energy configurations ?

    b. Compare this with simulated annealing where have a temperature parameter and gradually cool the system. What could you change in the genetic algorithm to mimic the effect of high or low temperature ?

    c. You wish to implement a similar scheme with Monte Carlo / simulated annealing. You have a population / set $\{\boldsymbol{m}\}$ of size n and you think it is a good idea to discard some individuals at each generation. Describe a method to generate a new generation of individuals which

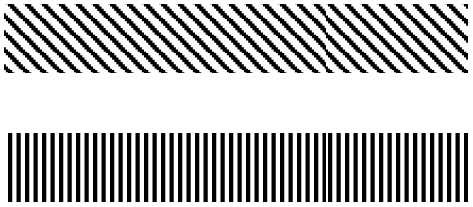
    - discards some individuals according to their probability
    - maintains a Boltzmann distribution

    There is more than one possible scheme which will do this. You keep some fraction f of the population at each step or this may be determined each step probabilistically.

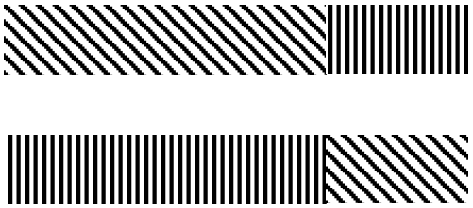
    How would you incorporate simulated annealing into this scheme ?

    What are advantages of the Monte Carlo based approach ?

22. Somebody says that the key to a genetic algorithm is not mutation, but crossover events. A possible scheme would be at each generation, first throw away some number of indivduals (maybe ½ of them). Of the remainder, pick ½, copy them and apply a mutation. Then pick pairs randomly. Copy them, but swap some characteristics within the pair.



before crossover



after crossover

This move is easy to implement if you can define your individuals as an array, such as a set of internal angles.

a. How would you implement crossover as a move in Monte Carlo ? Describe a simple method where there is no change in the number of individuals. If you start with two copies, you finish with two copies. To answer the question,
   - Write down the probability of the system (pair of individuals) before the crossover
   - Write down the probability of the system after the crossover
   - Write down an acceptance criterion

b. What is the disadvantage of simply using crossover events in a genetic algorithm, without any real acceptance criterion ?