Assignment 5

for lecture "Bioinformatics III" WS 08/09

Return before lecture on Dec. 11, 2008 or by email to p.walter@bioinformatik.uni-saarland.de until **Dec. 14**. This assignment will be discussed in the tutorial on Dec. 15, 2008, room15, building E1 3

1. Graph layout (20 points)

As you learned in the lecture there are various methods to create a graphical image of a network. In this assignment you will deal with a physically motivated method, the force directed layout. For this the nodes of a network are modelled as charged mass points that repel each other. Each link, on the other hand, is modelled by a spring, pulling the respective nodes closer together. When such a system is left alone it tries to get into a state of minimal energy, where the nodes are as far apart from each other as possible — with the constraint that two connected nodes have to stay close together. Thus the distance on the network (number of links between two nodes) is transformed into a spatial distance.

a) **Configuration of minimal energy:**
Determine the equilibrium distance between two equally charged mass points, which are connected by a spring. At the equilibrium distance the total force vanishes. Verify that instead of calculating the forces explicitly, it is equivalent to determine the configuration of minimal energy.
Hint: The force equals the negative gradient of the energy, i.e., the force is a measure for how much the energy changes with an infinitesimal displacement:

$$\vec{F}(\vec{r}) = -\nabla E(\vec{r}) \quad \textit{with the gradient operator} \quad \nabla := \begin{pmatrix} d/dx \\ d/dy \\ d/dz \end{pmatrix}$$

In 1D this reduces to $\nabla = d/dr$ *, i.e. the simple derivatative with respect to the distance r. The gradient of a function can consequently be understood as a multidimensional slope.*
Hint: The interaction energy between two charges q_1 *and* q_2 *is given as:*

$$E_c(r) = \frac{1}{4\pi\epsilon_0\epsilon}\frac{q_1 q_2}{r} \quad ,$$

for the connecting spring use the harmonic potential:

$$E_h(r) = \frac{kr^2}{2}$$

Hint: To show the equivalence of vanishing force and minimal energy remember how the minimum of a function is defined. Also note that the distance between two particles is a one-dimensional measure.

b) **Force field from a sperically symmetric potential:**
Calculate the force fields $\vec{F}(\vec{r}) = -\nabla E(\vec{r})$ for both the Coulomb interaction E_c and the harmonic potential E_h from 1a).

Hint: Write ∇ *and the resulting force field* $\vec{F}(\vec{r}) = \begin{pmatrix} F_x(x) \\ F_y(y) \\ F_z(z) \end{pmatrix}$ *in component form.*

Then you get one equation for x, y and z each. Note that $r = \sqrt{x^2 + y^2 + z^2}$.

2. Conditional probability (15 points)

For a proteom about 2% of the possible protein pairs interact with each other. It was found that a new experimental method for determining protein-protein interactions indicates an interaction in 99,9% of the cases when there is a genuine interaction and indicates no interaction in 97% of the cases when there is no interaction. What is the probability that there is a true interaction provided that the method gives a positive result. Do you think this new method is reliable? Explain your answer.

3. Bayesian network (65 points)

One way to estimate whether a given combination of proteins is a potential complex or not, is to use a Bayesian analysis. It allows to determine probabilities (likelyhood ratios) from known protein complexes based on their properties. These likelyhood ratios can then be used to estimate a probability whether the candidate is a potential complex. For this assignment, we use fake binary complexes, where each of the two proteins has two properties: a "function" and a "genome position". The "function" is labelled with a letter from [A–D] denoting the primary functional class, followed by a number from [1–7] for the subclass. The "genome position" consists of a letter from [A–C] for one of the three genes of our hypothetical species and an integer position denoting the transcription start in the range [1–1000]. These two properties are encoded in the protein name as "FunctionClass"+"FunctionSubClass"+ "_"+"GeneLabel"+"GenePosition"; a protein labelled, e.g., "A4_C86" belongs to the functional class "A", subclass 4 and is located at position 86 on gene C. To estimate the initial probability O_{prior}, you can use the information that from a set of 20000 protein combinations about 700 are identified as complexes. To determine the likelyhood ratios, you have two "Gold standard" data sets, gold_pos.dat and gold_neg.dat, which contain complexes that occur and that do not occur definitely, respectively, plus three "experimental" sets. These sets, which have a certain overlap with the gold standard data sets, contain both true and false complexes at a variable ratio, i.e., each experiment was performed at a different level of accuracy. Correspondingly, these sets are of different size, too.

Use the gold standard data sets to determine likelyhood ratios for the following properties:

a) **Function:**
For each complex in the gold standard data sets compare the main functional classes and the subclasses. From these two comparisons you get four categories of (i) equal main class and equal subclass, (ii) equal main class, but different subclass, (iii) different main but equal sub class and (iv) both main and sub class different. From the relative occurances of these four classes in the gold standard data sets determine the corresponding likelyhood ratio P_{func}. For each category, give the numbers, conditional probabilities, and likelyhood ratios in a table as presented in the lecture.

b) **Genome position:**
For the position on the genome use the two criteria whether both partners are from the same gene (same letter) or not and whether the absolute distance ($|pos1 - pos2|$) is < 10, < 100, or < 1000. This gives you six different categories with their respective likelyhood ratios P_{gen}.

c) **Likelyhood ratios from the "experiments"**
Use a fully connected Bayesian scheme to determine the likelyhood ratios P_{exp} from the "experimental" data sets exp1.dat, exp2.dat and exp3.dat. This gives you eight categories depending on in which combination of experiments the complexes of the Gold standard positives and negatives appear. Can you judge the quality of the experiments? Hint: To count the respective numbers of complexes in the different categories you do not have to write a program yourself — check the manpages for grep and wc (and maybe also for uniq and sort). Note that the Gold standard sets are not sorted and that complexes XY and YX are in fact identical.

d) **Identifying complexes :**
For all the potential complexes in the small test set of test1.dat give the likelyhood ratios for all properties Pf_{unc}, P_{gen} and P_{exp} and the final probability O_{post} that it is a true complex. Also give $\log(O_{post})$ (what for?) Start from a reasonable O_{prior}. Indicate the probable complexes.

Contact: Peter Walter, phone -3613, email: p.walter@bioinformatik.uni-saarland.de