Bioinformatics III

Prof. Dr. Volkhard Helms — Saarland University
Nadine Schaadt, Christian Spaniol, Ruslan Akulenko — Chair for Computational Biology
Winter Semester 2011/2012

Exercise Sheet 5

Due: November 25th, 2011 12:45

Submit your solutions on paper, hand-written or printed at the *beginning* of the lecture or in building E2 1, Room 3.09. Alternatively you may send an email with a single PDF attachment. If possible, please include source code listings. Additionally hand in all source code via mail to s9ruakul@stud.uni-saarland.de.

Bayesian Analysis of (fake) Protein Complexes

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 (likelihood ratios) from known protein complexes based on their properties. These likelihood 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>
```

For instance, a protein labelled "`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 1000 are identified as complexes.
To determine the likelihood 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.

Hints:

- There is more than one way to solve this exercise sheet.
  - **Either:** you can write a Python script/class to represent proteins and store their properties, i.e. function, subclass, gene, and genome position. Then write a `BayesianNetwork` to read the gold standard sets and count the occurrence of different characteristics – positive and negative, respectively. From those values, calculate your probabilities and likelihoods for the different categories. Also write a method that reads the experimental data from part 2 to determine P_{exp}. Finally use a method to calculate the final probabilities for the test sets in part 3.
  - **or:** if you are familiar with shell scripting, 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.

Exercise 5.1: Likelihood ratios from the theoretical properties (30 pts)

Use the gold standard data sets to determine likelihood 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 occurrences of these four classes in the gold standard data sets determine the corresponding likelihood ratio P_{func}.

For each category, give the numbers, conditional probabilities, and likelihood 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 ($|\text{pos}_1 - \text{pos}_2|$) is < 10, < 100, or < 1000. This gives you six different categories with their respective likelyhood ratios P_{gen}.

Exercise 5.2: Likelihood ratios from the "experiments" (30pts)

Use a fully connected Bayesian scheme to determine the likelihood 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?

Exercise 5.3: Identifying complexes (40pts)

(a) **Small test set**

For all the potential complexes in the small test set of `test1.dat` give the likelyhood ratios for all properties P_{func}, P_{gen} and P_{exp} and the final probability O_{post} that it is a true complex. Also give $\log(O_{\text{post}})$ (what for?)

Start from a reasonable O_{prior}. Indicate the probable complexes.

(b) **Gold standard revisited**

For all complexes of the Gold standard sets, determine the odds O_{post} that it is a true complex (do not list the values).

In one plot show the histograms of the two distributions of the occuring values of $\log(O_{\text{post}})$. In this plot also indicate the initial O_{prior}. Do you get a separation between the true and false candidates? At which value O_{cut} of the final O_{post} would you place the cut between true and false complexes? How many false positives and false negatives do you get from the Gold standard sets when you cut at O_{prior} and O_{cut}, respectively? What are the corresponding ratios of true positives to false positives? How large are the detection probabilities for the Gold positives?

(c) **Large test set**

For the large test set `test2.dat`, also determine P_{exp}, P_{theo}, and O_{post}. The probability P_{theo} is the product of the likelyhood ratios from the two theoretical criteria (function and genome position). With the value of O_{cut} from (b), split up the test data set into two sets of true and false candidates, respectively. Create a 2D scatter plot where you place a point for the pair (P_{theo}, P_{exp}) for each complex in the two sets of true and false complexes. Put both data sets into the same plot, but with different colors (markers). Choose a reasonable axis scaling and ranges so that the resulting plot is not too crowded or too sparse in some place. What do you observe? (Are the two sets already well separated? Is there a correlation between P_{theo} and P_{exp}? Anything else worth mentioning?)

Have fun!