

Exercises Algorithmic Systems Biology

Freie Universität Berlin, SoSe 2010
Roland Krause · Martin Vingron

Blatt 1 · Durchführung am 21.5.2010

Aufgabe 1 (Differential genes and false discovery rate). This exercise builds on exercise 45, Zettel 11 of the course *Algorithmic Bioinformatics*, WiSe 2008/09.

1. Simulate the gene expression of 1000 genes with 50 significantly changed between two samples a, b .
 - Draw the non-differentially expressed genes from a distribution with $\mathcal{X} \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 10$ and $\sigma = 5$.
 - The differentially expressed genes are generated for sample a as before. For sample b , add values from $\mathcal{N}(5, 1)$.
 - Repeat the procedure to obtain 4 replicates. Use the following with 2, 3 and 4 replicates.
2. Calculate a t-test for each gene. Do you need to use a paired t-test?
3. Plot the resulting test statistic in ascending order. Include a line from the origin to the highest point.
4. Let's use an $\alpha = 0.05$. How many non-differentially expressed genes do we expect to see by chance and how many do we see?
5. Apply the Bonferroni correction. How many genes are significantly expressed now?
6. How many genes are significantly expressed if we apply the FDR as introduced by Benjamini and Hochberg?
7. Graphically interpret the procedure in the plot of the p-values.
8. What is a reasonable FDR in this setting? What are the practical consequences?
9. How would you have to select the FDR when using two replicates?

Notes

1. Label all plots (axis, legend, title) as for publication.

Questions

1. What would happen if you normalize the data?

(45min)

Aufgabe 2 (Markov Chain Monte Carlo). Sample from a distribution using the Metropolis algorithm.

Metropolis algorithm

1. Initialize $x^{(0)}$
2. For $i = 0$ to $N - 1$
 - Sample $u \sim \mathcal{U}$
 - Sample $x^* \sim q(x^*|x^{(i)})$
 - If $u < \mathcal{A}(x^{(i)}, x^*) = \min\{1, \frac{p(x^*)}{p(x^i)}\}$
 - $x^{(i+1)} = x^*$
 - else
 - $x^{(i+1)} = x^i$

Use the *proposal distribution* $q(x^*|x^{(i)}) = \mathcal{N}(x^{(i)}, \sigma^2)$.

- Instead of simulating a complex scheme, we use a simple bi-modal *target distribution*
 $p(x) \propto 0.3\exp(-0.2x^2) + 0.7\exp(-0.2(x - 10)^2)$
- Implement and run the analysis for 5000 iterations with $\sigma = 1, 10, 100, 1000$.
- Plot the results in a scatter plot. Assess the convergence after 300, 1000 and 3000 iterations.
- Generate a histogram of the results. Does it resemble the target distribution?

(45min)