Integrative Transcriptomics
Dr. K. Nieselt, Faculty of Science

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Lecture: Bioinformatics I — WS 2015/16

Assignment No. 3 (10 points)

Hand out:	Monday, October 26
Hand in due:	Monday, November 2, 10:15

Direct inquiries to `alexander.seitz@uni-tuebingen.de`

Theoretical Assignments

1. **Equivalence of distance and similarity alignments** (2P, 2-3h)

   In the lecture "Grundlagen der Bioinformatik" the Levenshtein (edit) distance was introduced that can be used to compute an optimal global alignment that minimizes the distance score. Here the score of the gaps introduced to the alignment is added (and not subtracted as in the case of similarity based alignments) to the score of the aligned characters in order to compute the optimal score.

   Using this information, prove the following

   Theorem: Let a similarity measure be given with $s(a, b)$ and gap penalties $\hat{\gamma}(k)$. Let a distance measure be given with $d(a, b)$ and gap weight $\gamma(k)$. Assume there is a constant c, such that $s(a, b) = c - d(a, b)$ and $\hat{\gamma}(k) = \gamma(k) - ck/2$. Then a global alignment of two sequences of length n and m respectively is optimal with respect to its similarity if and only if it is optimal with respect to its distance.

   (Hint: Use the equation $n + m = 2 \cdot \#a + \sum_k kg_k$ where a=aligned characters and g_k is the number of gaps of length k.)

Practical Assignments

2. **Using BLAT to align 454 reads to the *Helicobacter pylori* genome** (8P, 5-8h)

   454 is a next generation sequencing technology based on pyrosequencing, developed by 454 Life Sciences, later bought by Roche[1]. With this technology sequences (so-called 454 reads) of DNA with lengths up to 400bp can be produced. In this exercise you are supposed to use the program `BLAT` to align 454 reads to the genome of *Helicobacter pylori*. Afterwards perform a series of analysis steps on the resulting data.

   The 454 reads, the genome and the BLAT sources can be downloaded from our website in the material zip folder A03.zip.

   Application:

   - Compile the `blat` sources.
   - Use `blat` to align the reads in `reads.fasta` to the genome `Hpylori.fasta`.
     Please restrict the maximum intron size to 50. Document the command-line you used.
     Why do we want to limit the intron size?

[1] http://www.454.com/ no longer in use or supported by Roche, however note that sequencing technologies such as those sold by the company PacBio and others are already producing much longer reads.

- Inspect the output file. Write down the meaning of each column.

Analysis:

- Write a Java program that takes the output file of BLAT as input and answers the following questions:
- How many reads could be aligned by BLAT?
Your answer should contain the absolute number and the percentage wrt. all input reads.
- How many reads could be mapped *uniquely* by the program, i.e. how many reads map to exactly one locus in the genome?
- How many positions of the *H. pylori* genome have not been aligned by any read?

If you have a closer look on the output files, you will recognize that some of the reads show interesting results (e.g. `SRR031182.1, SRR031182.3, SRR031182.4, ...`). Try to find a biological explanation for this phenomenon.

In addition to your source code and report, hand in the resulting BLAT output file for this task!

3. ***Bonus: Use SSAHA2 to align 545 reads to the *Helicobacter pylori* genome** (4P, 4 h)

Please download `SSAHA2` from the web (`https://www.sanger.ac.uk/resources/software/ssaha2/#t_2`).

Application:

- Use `SSAHA2` to align the reads in `reads.fasta` to the genome `Hpylori.fasta`.
- Set the `-454` parameter to optimize the application for 454 reads.
- Try to find a parameter that results in a similar output format as `blat` to make the subsequent analysis easier.
- Document the command-line parameters you used.

Analysis:

- Extend your Java program from the previous task so that it takes the output files of the two programs as input and answers the the same questions for both programs as in the previous task.

Provide a short overview of the analysis results and briefly discuss them.
How do BLAT and SSAHA differ with respect to runtime and sensitivity? Which of the two programs would you recommend in this case and why?

Please read the questions carefully. If there are any questions, you may ask them during the tutorial session or via e-mail to your tutor. You will usually get an answer in time, but late e-mails (e.g. on Monday morning before class) might not be answered in time. Please send all your electronic solutions to `alexander.seitz@uni-tuebingen.de` or `alexander.peltzer@uni-tuebingen.de` (depending on your tutor). Please pack both your source code as well as the theoretical part into one single archive file. Source code should compile correctly. Make sure, that you export the source code and not only the binaries. Handwritten assignment solutions (e.g. for the theoretical part) can be turned in during the lecture.