ITO Integrative Transcriptomics
Prof. K. Nieselt,
Institute for Bioinformatics and Medical Informatics Tübingen
Prof. S. Nahnsen,
Institute for Bioinformatics and Medical Informatics Tübingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Lecture: Grundlagen der Bioinformatik SoSe 2022

Assignment 10 (20 points)

Hand out: Thursday, July 7
Hand in due: Thursday, July 14 18:00
Direct inquiries via the ILIAS forum or to your respective tutor at:
Mathias Witte Paz: iizwi01@uni-tuebingen.de
theresa-anisja.harbig@uni-tuebingen.de
meret.haeusler@student.uni-tuebingen.de
jules.kreuer@student.uni-tuebingen.de
simon.heumos@qbic.uni-tuebingen.de

1. **Combinatorics of hairpins** (6P)

   We have seen that the structural components of an RNA secondary structure are hairpins, bulges, interior loops, and multi-branch loops. Just to make sure: a hairpin is a structure with at least one basepair and exactly one (end) loop of size of at least 1. (It can possibly have bulges and interior loops in the stacked region). Prove the following:

   Theorem: There are $2^{n-2} - 1$ hairpin structures for sequences $[1, ..., n]$ of length n, with $n \geq 3$.

   As examples, draw all hairpin structures for sequences of length $n = 4$.

2. **Visualisation of RNA secondary structure** (4P)

   For the following two RNA sequences, first derive the secondary structures (you do not need to formally compute the secondary structures, just deduce from the sequence) and write them in dot-bracket notation. Then draw their secondary structures as a graph (see Figure on p. 164) and the arc representation:

   CACGCUGAACGUACU (hairpin structure)

   UCCAGCAGGAAAGC (pseudoknot structure)

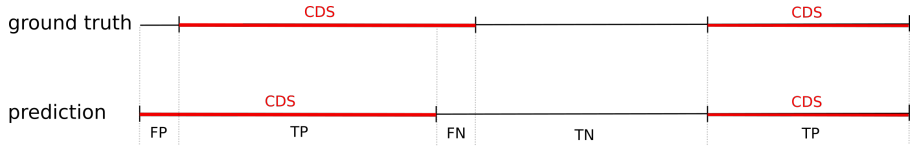
3. **Gene prediction** (10P)

   Background on the task: In this task we ask you to compare *PROKKA* and *GeneMark*, two gene prediction software tools that use either a dynamic programming approach (Prokka) or hidden Markov (GeneMark) to identify coding regions in prokaryotic DNA.

   Both, *PROKKA* and *GeneMark* output their feature predictions in an `.gff`-file. The GFF file format is used to describe predicted features like CDS or exon of DNA sequences. Inform yourself about the file format, especially what each line and what the nine different columns represent.

   Your task is to run *PROKKA* and *GeneMark* on the genome of *Pseudomonas aeruginosa* and compare their respective output to a given `.gff`-file (`PAO1_annotation.gff`), which we consider

as the ground truth. In more detail, we only want to examine the coding sequences (CDS) on the nucleotide resolution level. Therefore, we count nucleotides/positions that are predicted correctly (true positives, true negatives) and not correctly (false positives, false negatives) as either being 'coding' or 'not coding', which is indicated in the following scheme:



Note: The sum of true positives, true negatives, false positives and false negatives equals the length of the genome. Make sure that when calculating the sum of true positives, true negatives, false positives and false negatives you consider that the predictions are made for both strands of the genome.

First part: Run *PROKKA* and *GeneMark*

- In order to run *PROKKA* go to the website `https://usegalaxy.org/` and create an account there. After confirming your email address search for the *PROKKA* tool and upload the input genome file for analysis. Select only the `.gff`-file as an output.
- Run *GeneMark* using the web service `http://exon.gatech.edu/GeneMark/gmhmmp.cgi` on the genome of *Pseudomonas aeruginosa PAO1* (`PAO1_genome.fasta`). Make sure you select the correct species (strain) and output format (gff).
- Save the respective outputs on your computer.

Second part: Comparison of results

- Get familiar with the provided GFF parser[1] to include it in your program. You might want to read the README in the GitHub repository for the instructions. (See Bonus task for an alternative)
- Write a Python program which compares the `.gff`-files of the two programs to the provided ground truth. Focus on the CDS feature. Compute true positives, true negatives, false positives and false negatives as described in the introduction by comparing the CDS regions of the respective files. From these compute the sensitivity, specificity and accuracy of *PROKKA* and *GeneMark*, respectively.

- Write a small section with your results and discussion (max. 2 pages, incl. figures).
- **Bonus (+2P):** Write your own `.gff` parser. Indicate in the PDF if you have developed your own GFF.

Please read the questions carefully. If there are any questions, you may ask them during the tutorial session or in the forum of ILIAS. You will usually get an answer in time, but late e-mails (e.g. the evening of the hand-in) might not be answered in time. Please upload all your solutions to ILIAS. Don't forget to put your names on every sheet **and** in your source code files. Please pack both your source code as well as the theoretical part into one single archive file and give it a name using this scheme: `<name1>_<name2>_<Assignment>_<#>.zip`. The program should run without any modification needed.

[1] `https://github.com/Jverma/GFF-Parser`