Bioinformatics I
WS 15/16
Tutor: Alexander Seitz

1	2	3	$\sum$

Jonas Ditz
& Benjamin Schroeder

Assignment 3

(Abgabe am 2. November 2015)

Theoretical Assignment - *Equivalence of distance and similarity alignments*

Assume there are two sequences of length n and m, respectively. The length of an alignment between this two sequences is of length n+m. As it is written on the assignment following equation is valid:

$$n + m = 2 * M + \sum_k kg_k, \tag{1}$$

where M is the number of aligned characters. Since later one of our sequences is called a, we changed the latter a, which is used on the assignment sheet, to M.
Using this equation, we can write the distance of our two sequences (let us call them a and b) as

$$\begin{aligned} D(a,b) &= min\{\sum_M d(a,b) + \sum_k kg_k\} \\ &= min\{\sum_M c + \sum_k kg_k c/2 - \sum_M s(a,b) + \sum_k \hat{\gamma}(k)g_k\} \\ &= min\{c(n+m)/2 - \sum_M s(a,b) + \sum_k \hat{\gamma}(k)g_k\} \\ &= c(n+m)/2 - max\{\sum_M s(a,b) - \sum_k \hat{\gamma}(k)g_k\} \\ &= c(n+m)/2 - S(a,b) \end{aligned} \tag{2}$$

Solving for S(a,b), we get:

$$S(a,b) = c(n+m)/2 - D(a,b) \tag{3}$$

So one can see that the Score is optimal if and only if the Distance is optimal.

Practical Assignment - *Using BLAT to align 454 reads to the Helicobacter pylori genome*

The blat algorithm was compiled in a virtual box using bio-linux. The output generated by the blat algorithm was saved as .plt and is attached. The maximal Intron size was set to 50, to limit the ammount of hits. Generally this is a good idea, because introns which are getting to large make no sense in a biological context. The analysis was accomplished by a java program, which is

also attached. The program was executed via eclipse and not via the command line. If you want to use it, you might change the path directions in the main class and it should work than. Sadly is the runtime very high and was around 45 mins for the given data set.

The complete Data set had 524427 entries. The absolute number of reads, which was be aligned is 437586 or 83 %. Of these reads 314896 were unique, which are 60%. A second part of the task was to figure out how many positions of the H. pylori were left unaligned. From the total number of positions 1691694, 1071574 which is a coverage of 63The repeating of sequences can happen very often in different contexts. For example in with noncoding sequence regions or with viral DNA which is stabil incorporated into the genome.

Practical Assignment - *Bonus: Use SSAHA2 to align 535 reads to the Helicobacter pylori genome*

As described on the assignment sheet, we downloaded the binaries of SSAHA2 from https://www.sanger.ac.uk/resources/software/ssaha2/#t_2. After unpacking the zip archive SSAHA2 was working, instantly. We ran SSAHA2 with the following command:

/bin/ssaha2_v2.5.5_x86_64/ssaha2 -454 -output psl ../data/Hpylori.fasta ../data/reads.fasta > hpylori_reads_ssaha2_output.psl

The parameter *-output psl* convert SSAHA2 output into psl format, which is the same format as used by BLAT. One can look over the results by opening hpylori_reads_ssaha2_output.psl, which we provide together with this pdf. Since it is a huge file, we used a short python script (analysis.py, also send as a attachment) to get a short overview of our results.

Output of analysis.py:
- number of all mapped reads:
  - absolute number: 321954
  - percentage: 100
- number of unique mapped reads:
  - absolute number: 72897
  - percentage: 22.642054455

With that short overview one can see that SSAHA2 mapped all provided reads to our database. In this case our database was Helicobacter pylori. That is interesting and might indicate a problem with the default parameters of SSAHA2. It looks like SSAHA2 tries to get a sensitivity of 100%, which was successful in our case. One has to decide from case to case, if that is really what their needs. We run SSAHA2 on a linux machine with 1 Gb RAM and a single core processor with 1.66 GHz. The runtime was about one hour on that computer.