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*

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