Algoritmer for bioinformatikk - TDT4287 - Øving 2

Ola Kristoffer Hoff

30. September 2022

Heuristics & expected running times

1 * Consider the Exact pattern matching problem introduced in the lecture. Consider the brute force algorithm using branch-and-cut. That is, over all positions of the text, the algorithm tries to match the pattern until the first mismatch is encountered, at which point it starts from scratch at a new position.

What is the expected running time of this brute force algorithm for:

A Text and Pattern symbols distributed uniformly at random.

B Text distributed uniformly, but each symbol of the Pattern is indepen- dently distributed according to the discrete distribution specified by: $p_A = 0.2, p_C = 0.4, p_T = 0.3, p_G = 0.1$**.**

C Both Text and Pattern distributed unevenly.

2 Solve the k-difference global alignment problem, as well as standard global alignment for Edit Distance scoring function δ, sequences ω_1 = GGCTCTA and ω_2 = CTCTAGC, and $k = 2$.

A Do you get the correct solution?

Yes, in this case we would not get a solution to the global alginment. Since it has at least four mismatches/indels.

+	¢	G	G	C	T	C	T	A
¢	0	-1	-2	-3	-4	-5	-6	-7
C	-1	-1	-2	-2	-100	-100	-100	-100
T	-2	-2	-2	-3	-2	-100	-100	-100
C	-3	-3	-3	-2	-3	-2	-100	-100
T	-4	-100	-4	-3	-2	-3	-2	-100
A	-5	-100	-100	-4	-3	-3	-3	-2
G	-6	-100	-100	-100	-4	-4	-4	-3
C	-7	-100	-100	-100	-100	-4	-5	-4

Table 1: Table of GA with k-difference (k = 2).

B When are you guaranteed that the algorithm's solution is the actual solution for the ED problem?

The standard global alginment always reflect the edit distance. Since the global alignment is the best alignment, hens fewest mismatches/indels. The edit distance is then the length of the string minus the global aligment. This all assumes that the δ-function is the same for the ED and GA.

+	¢	G	G	C	T	C	T	A
¢	0	1	2	3	4	5	6	7
C	1	1	2	2	3	4	5	6
T	2	2	2	3	2	3	4	5
C	3	3	3	2	3	2	3	4
T	4	4	4	3	2	3	2	3
A	5	5	5	4	3	3	3	2
G	6	5	5	5	4	4	4	3
C	7	6	6	5	5	4	5	4

Table 2: Table of ED with standard GA.

3 * Suppose that you know that GA-score distribution follows the following probability distributions. Can you design an algorithm that computes Edit Distance GA, and uses *on average* asymptotically less than O(n · m) time?

A **Binomial distribution** $Bi(\lfloor \log(n) \rfloor, \frac{1}{2})$.

B **Probability distribution specified by probability mass function** $p(k) = \frac{1}{2} p_{Bi(\lfloor \log(n) \rfloor, \frac{1}{2})}(k) + \frac{1}{2} p_{Geo(\lambda)}(k - \lfloor \log(n) \rfloor)$.

where $p_{Bi(\lfloor \log(n) \rfloor, \frac{1}{2})}(p_{Geo(\lambda)})$ **denotes the probability mass function of the respective Binomial (Geometric) distribution.**

C **Will your solution work for the following distributions?**

- **Uniform probability distribution over all possible GA values.**
- **Probability distribution specified by probability mass function** $p(k) = \frac{1}{2} p_{Bi(C \cdot n, \frac{1}{2})}(k) + \frac{1}{2} p_{Geo(\lambda)}(k - \lfloor \log(n) \rfloor), p$ *where* $C \in (0, \frac{1}{2})$ *is a constant.*

p-values

4 You have found the LA-score for sequences ω_1, ω_2 to be 8. Based on your parameters for the LA problem, and composition of the sequences, you model the LA-score distribution for unrelated ω_1, ω_2 by the following distributions.

What are the *p*-values for the LA alignment you have found? Would you consider the sequences to be homologous?

A Poisson distribution with probability mass function $f_S(s) = \frac{\lambda^s e^{-\lambda}}{s!}$, *with parameter* = 4.

We have $s = 8 => f_S(8) = \frac{4^8 e^{-4}}{8!} = \frac{8*4*2*2*2*4^4}{8*4*2*7*(3*2)*5*3*e^4} = \frac{2*4^4}{7*3*5*3*e^4} = \frac{512}{315e^4} \approx 0.0298$

Given that I have found the p-value $0.0298 < 0.05$ that is less than the normal 5% p-value test. I thus conclude that the sequences probably are homologous.

B Probability distribution with cumulative distribution function $P(S \leq s) = e^{-100 \cdot e^{-s}}$.

We have hare $s = 8 => P(S \leq 8) = e^{-100 \cdot e^{-8}} \approx 0.9670$

We could do the same test here, but I really don't know whats going on here at all.

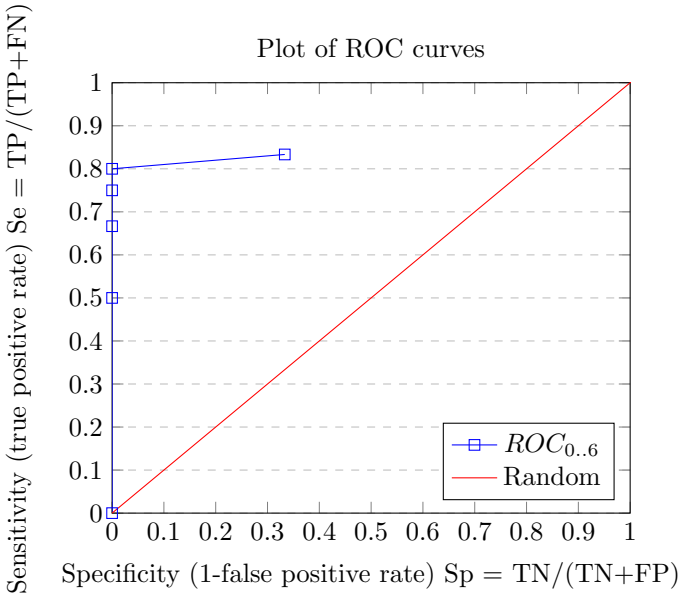
C * Consider your LA-score was 30, and you model the LA values with Poisson distribution with $\lambda = 25$. Approximate the Poisson distribution with normal distribution, and determine the p-value. That is, use $X \sim N(\lambda,\lambda) instead of X \sim Po(\lambda)$. Use the table on the next page with values of $CDF\ N(0,1)$. You thus need to transform the LA-value to standard score, also known as Z-score. Compare your results with and without the continuity correction, with the real value of Poisson CDF. Try the same with LA-score of 20.

Classifiers & ROC-curves

5 You have developed a test to determine homologous sequences. Homologous sequences get score X with distribution $X \sim F_X$, whereas non- homologous get $X \sim G_X$. Here, F_X and G_X are the respective cumulative distribution functions.

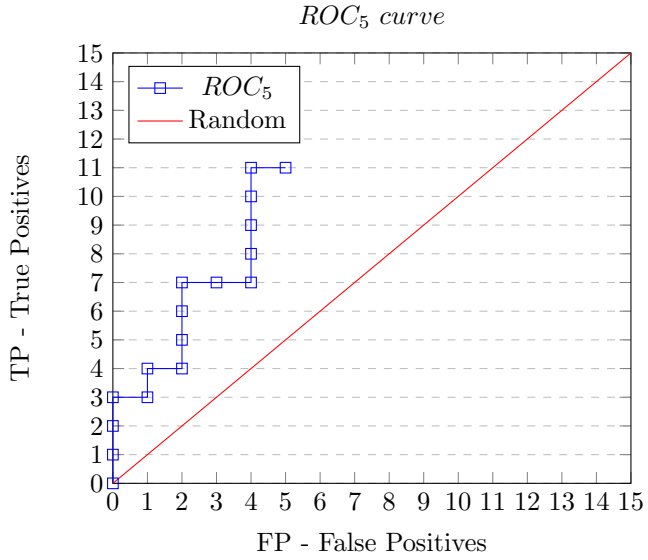
A You created a small control set to see how your clasifier performs. Your positive cases got scores {4, 6, 8, 9, 10, 12} and your negative cases got scores {1, 2, 3, 3, 5, 7}. Draw the ROC curve for these cases, and determine the confusion matrix for threshold of 5,5.

Based on the text I am assuming that the threshold is 5,5 for the first task as well. $\{4,6,8,9,10,12\} \xrightarrow{CM} \{FN\ TP\ TP\ TP\ TP\ TP\}$ and $\{1,2,3,3,5,7\} \xrightarrow{CM} \{TN\ TN\ TN\ TN\ TN\ FP\}$



	Homologous sequences	Non-homologous sequences
Score at least threshold T	TP	FP
Score under threshold T	FN	TN

B **Next, you run your algorithm on all the sequences in the database. Each sequence got a unique score, and the sequences with the highest scores were labelled, in decreasing-score order, as follows: (P, P, P, N, P, N, P, P, P, N, N, P, P, P, P, N, ...) Compute ROC for these results. What is its interpretation?**



ROC_6 value is equal to $3 \cdot 1 + 4 \cdot 1 + 7 \cdot 2 + 11 \cdot 1 = 3 + 4 + 14 + 11 = 42$

C **You want to share the sequences with your colleagues. There is a good distinction between the distributions, so you decide your priority is that the sequences you share include at least 90% of all homologous sequences.**

Express in terms of F_X and G_X how to choose the cut-off value. Can you interpret the task graphically on the ROC curve?

Graphically this must mean that the cut-off would be at number of Relative TP equal to 0.9.

D **Assume that, from experience, you know that only 1 in 1000 sequences are homologous. You have tested all the sequences that were available in the database, order of 10^6, so you can believe that your results will hardly deviate from the pattern. You want to make sure your colleagues use their time wisely looking at sequences that indeed are homologous, but at same time you want to have them look at as many as possible. Thereby, you want to provide them with a set of sequences, where you can expect that at most 5% of them not to be homologous.**

Express in terms of F_X and G_X how you would choose the cut-off value. Can you interpret the task graphically on the ROC curve?

E *** Your method seems to work nicely, and you decided to share it with your colleagues. You don't think they would know themselves what threshold to choose, so you decided to choose the threshold that correctly labels the highest portion of sequences, whether that correct label is positive or negative.**

- **How do you find the threshold, and what does it represent on the ROC curve? Consider both discrete and continuous case.**
- **How does the answer change if negatives and positives are not distributed evenly?**