Sequence Bioinformatics: Assignment 03

Tobias Fehrenbach and Ignacio García Ribelles

October 2021

1 Counting

1.1 Edges between nucleotides

Three sequences were given, the first two consisted of four nucleotides and the latter of five. In an alignment graph, edges connect always two nucleotides of different sequences with each other. Using the binomial coefficient one can compute all possible k subsets of n possible elements:

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

Thus, in order to compute the number of edges N_e between k=2 nucleotides in different sequences, one has to subtract the combinations between each of the three sequences $n_{S1} = 4$, $n_{S2} = 4$, $n_{S3} = 5$ from the total number of nucleotides $n_{Tot} = n_{S1} + n_{S2} + n_{S3} = 13$ as shown below:

$$N_e = \binom{13}{2} - 2 \cdot \binom{4}{2} - \binom{5}{2} = 78 - 2 \cdot 6 - 10 = 56$$

The first two constraints for forming edges have thus been established, given that it starts at the position j of a sequence i and ends at the position q of a sequence p:

1. An edge cannot start and end at the same point:
   if $ij = pq$ then $x_{ij\_pq} \leq 1$. (e.g x_{00_00} ↯/ x_{00_11} ✓)
2. An edge cannot connect nucleotides in a same sequence:
   if $i = p$ then $x_{ij\_pq} \leq 1$. (e.g x_{00_02} ↯/ x_{00_20} ✓)

This adds up to a total of 56 possible edges, all of which are listed as binary constraints in the output of the ilp_alignment_Garcia_Fehrenbach.py program.

1.2 Number of simple mixed cycles

Mixed cycles form from the combination of edges and do not result in correct alignments. In the case of a three-sequence alignment with a maximal overlap between the three sequence lengths of 4, mixed cycles can be formed by any number of edges between 2 and 12. This sums up to a total of $\binom{56}{2} + ... + \binom{56}{12} =$ 752.157.638.737 theoretically possible alignment graphs, from which most can be simplified further to either simple mixed cycles or correct alignments. The first additional constraint relates to symmetry and further halves the range options:

3. Two edges that start at the same point as the other one ends are not allowed:

   if $ij = p'q'$ then $pq \neq i'j'$ else $x_{ij\_pq} + x_{i'j'\_p'q'} \leq 1$. (e.g $x_{00\_10} + x_{10\_00}$ ↯/ $x_{00\_10} + x_{00\_20}$ ✓)

For this reason we further-on decided to only loop through half the theoretically feasible possibilities, thus sparing computational space and time.

In order to form a correct alignment graph, crossing edges are also to be avoided, as they would result in letters of the nucleotide sequence being swapped or being aligned to more than one position, thus adding a fourth constraint:

4. The first edge cannot end at a higher position as the second one starts:

   if $j \leq q \leq j'$ then $j' \leq q'$ else $x_{ij\_pq} + x_{i'j'\_p'q'} \leq 1$. (e.g $x_{00\_11} + x_{02\_10}$ ↯/ $x_{00\_10} + x_{00\_20}$ ✓)

This further decreases the number of possible alignment graphs, which correspond to all possible crossing edges that form simple mixed cycles between any two sequences. However, as we are dealing with three sequences, simple mixed cycles can also be formed from three separate edges, given the following constraint:

5. Three edges can only be formed within one column of the alignment, thus form a closed triangular cycle:

   if $ij == p''q''$ then *j=q=j’=q’=j”=p”* else $x_{ij\_pq} + x_{i'j'\_p'q'} + x_{i''j''\_p''q''} \leq 2$. (e.g $x_{00\_11} + x_{11\_20} + x_{00\_20}$ ↯/ $x_{00\_10} + x_{10\_20} + x_{00\_20}$ ✓)

Our program calculated 344 simple mixed cycles between each pairs of the three sequence combinations: S1-S2, S2-S3 and S1-S3, corresponding to two-edged constraints. This adds up to the expected $\binom{|S_a|+1}{2} \cdot \binom{|S_b|+1}{2} - S_a \cdot S_b$ for each of the three combinations adding up to 84 + 130 + 130 = 344. Additionally, we found 264 three-edged simple mixed cycles between all three sequences at once: S1-S2-S3, however we expect at least $\binom{|S1|+1}{2} \cdot \binom{|S2|+1}{2} \cdot \binom{|S3|+1}{2} - |S1| \cdot |S2| \cdot |S3| = \binom{5}{2} \cdot \binom{5}{2} \cdot \binom{6}{2} - 4 \cdot 4 \cdot 5 = 1420$ simple mixed cycles, summing up to a total of 1764 simple mixed cycles. However, as we only attained 608, our error clearly lies on the

2 Simple mixed cycles

A program was written to output a list of all possible simple mixed cycles according to the five constraints listed on the former exercise, yielding 608 simple mixed cycles, which were printed to an output file in the required format. The function *calc_2constraints* gave us 344 simple mixed cycles between each two sequences, while as the function *calc_3constraints* calculated 264 simple mixed cycles for all three sequences combined.

3 Objective function

Using a match score of 4 and a mismatch score of 1, the objective function for lpsolve was written in the required format, by listing every possible edge and assigning a score depending on the corresponding connected nucleotides. This is realized in our program by the function *setup_objective_function*.

4 Run the ILP

The lpsolve program was downloaded and run with the the output file of our lp_aligment_Garcia_Fehrenbach.py program. The output form is done as specified in the assignment. The program needs an output path in order to produce the output file specified as the first command-line argument.

5 Report the alignment

In order to translate the output of lp-solve into an alignment, one first has to draw the alignment graph and fill in all of the given edges with an actual value of 1. The attained edges were the following:

> x00_10, x01_11, x01_20, x11_20, x02_22, x03_12, x03_23, x12_23 and x13_24

They corresponded to a total of 9 matches, thus yielding a score of 36. The resulting alignment graph and corresponding alignment are shown in figures 1 and 2.

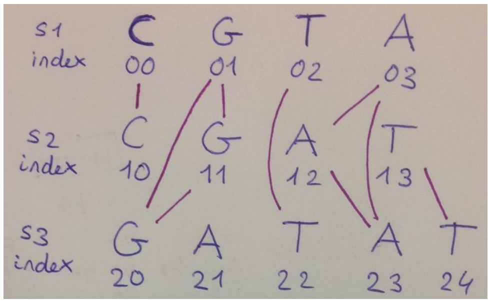
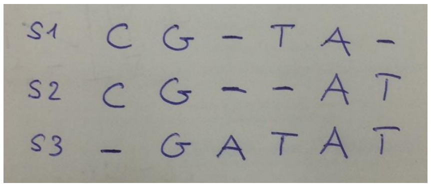


Figure 1: Alignment graph of the three given sequences drawn from the output of lp_solve.



```
s1  C G - T A -
s2  C G - - A T
s3  - G A T A T
```

Figure 2: Alignment of the three given sequences drawn from the output of lp_solve.