Prof. Dr. Daniel Huson

Institute for Bioinformatics
and Medical Informatics
Faculty of Science (MNF)

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sequence Bioinformatics **WS 2025/26**

Assignment 3 **Due: Nov-5, 10 am**

In this assignment, please implement a Java program `ILP4MSA_YOUR_NAME.java` that writes out an "Integer Linear Program" to compute a multiple sequence alignment for these there sequences (hard-wired into the program):

```
S0: TCAG
S1: TCCTA
S2: CTAG
```

Please use the outline of `ILP4MSA_YOUR_NAME.java` provided on Ilias.

1 Objective function (1 point, together with task 4)

In the following, use $\texttt{X}ij\_pq$ to denote the variable that represents the edge connecting the nucleotide $s_i(j)$ in sequence s_i, at position j, with the nucleotide $s_p(q)$ in sequence s_p, at position q.

Using a match score of 4 and a mismatch score of 1, set up the objective function for the ILP, in the format:

```
max: +4*x_00_10+1*x_00_11+...
```

(Should contain 50-100 terms).

2 All simple-mixed cycles involving two sequences (2 points)

Generate the list of all simple mixed cycles that use only two of the three sequences, using the following format (which can be parsed by `lp_solve`, note that `<` means "$\leq$"), e.g.:

```
x_00_11 + x_00_10 < 1;
x_00_12 + x_00_10 < 1;
...
```

(Should be several hundred constraints).

Note that here we allow the two edges to start or end at the same node (but they can't both start and end at the same node, they must be different edges). This ensures what we called "column uniqueness" in Section 3.13.2 of the script.

3 All simple-mixed cycles involving three sequences (3 points)

Generate the list of all simple mixed cycles that use all three sequences, using the following format (which can be parsed by `lp_solve`, note that `<` means "$\leq$"), e.g.:

```
x_00_21 + x_00_10 + x_10_20 < 2;
x_00_22 + x_00_10 + x_10_20 < 2;
x_00_23 + x_00_10 + x_10_20 < 2;
x_00_22 + x_00_10 + x_10_21 < 2;
x_00_23 + x_00_10 + x_10_21 < 2;
...
```

(Should be several thousand constraints).

The above comment about ensuring column uniqueness also applies here.

4 Declare all variables to be integer variables

Generate a line declaring all variables as binary:

```
bin x_00_10, x_00_11, x_00_12, ... ;
```

(Should be same number of terms as in task 1.)

5 Run the ILP (1 point)

Download the program `lp_solve`, from `https://sourceforge.net/projects/lpsolve/` and install it.
Save the output of your Java program to a file.

```
max: +4*x_00_10+1*x_00_11+...

x_00_11 + x_00_10 < 1;
x_00_12 + x_00_10 < 1;
     ...    (all simple mixed cycle constraints involving two sequences)

x_00_21 + x_00_10 + x_10_20 < 2;
x_00_22 + x_00_10 + x_10_20 < 2;
  ...     (all simple mixed cycle constraints involving three sequences)


bin x_00_10, x_00_11, x_00_12, ... ;
        (specify all variables as binary)
```

6 Interpret the output (2 points)

Show how to interpret the output of the ILP solver; draw the extended alignment graph and indicate chosen edges, and also draw the associated multiple-sequence alignment.

7 Revision (exam-style question) (1 point)

Draw the extended alignment graph for the following three sequences. List four different *simple mixed cycles* (such as, but other than, $\mathtt{A}_1 - \mathtt{A}_5 \to \mathtt{A}_8 - \mathtt{A}_1$), of which two involve only two sequences and two involve all three sequences.

$$S_1 = \quad \mathtt{T}_1 \quad \mathtt{G}_2 \quad \mathtt{C}_3 \quad \mathtt{A}_4$$

$$S_2 = \quad \mathtt{T}_5 \quad \mathtt{C}_6 \quad \mathtt{A}_7 \quad \mathtt{T}_8$$

$$S_3 = \quad \mathtt{G}_9 \quad \mathtt{T}_{10} \quad \mathtt{C}_{11}$$