Try the following exercises from pages 139–140 of the Setubal-Meidanis book:

3, 6, 7, 12, 13, 18, 1

Answers: `http://www.liacs.nl/home/kosters/bio/`

Exercise 3 from Setubal-Meidanis, p. 139:

> What is the smallest value of ϵ such that the layout below is valid under the Reconstruction model?

```
ACCGT          --ACCGT--
CGTGC          ----CGTGC
TTAC           TTAC-----
TGCCGT         -TGCCGT--
                --------
               TTACCGTGC
```

Note that there is one mistake present. So we compute $d_s(\texttt{TGCCGT}, \texttt{TTACCGTGC}) = 1$.

We should have that

$$d_s(\texttt{TGCCGT}, \texttt{TTACCGTGC}) \leq \epsilon \, |\texttt{TGCCGT}| \quad ,$$

or $1 \leq \epsilon \, 6$.

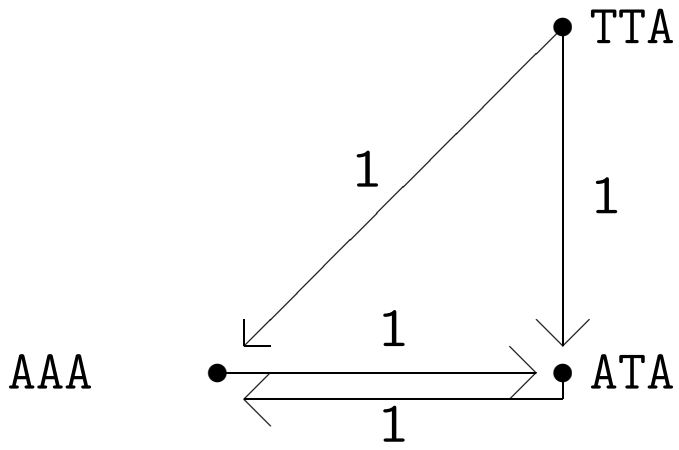
So the smallest admissible value is

$$\epsilon = 1/6 \quad .$$

Exercise 6 from Setubal-Meidanis, p. 140:

> Construct the overlap graph for $\mathcal{F} = \{\texttt{AAA}, \texttt{TTA}, \texttt{ATA}\}$. Find a shortest common superstring for this collection.

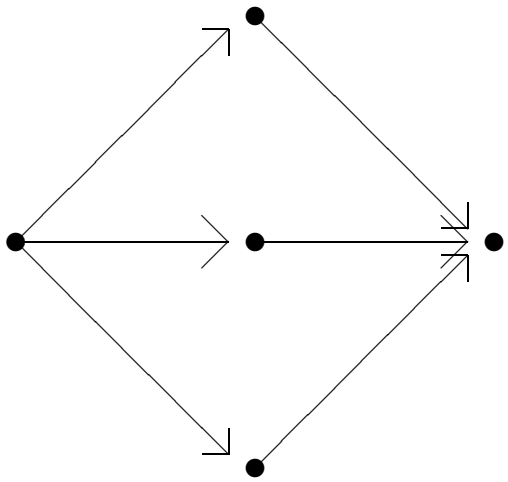
Overlap graph

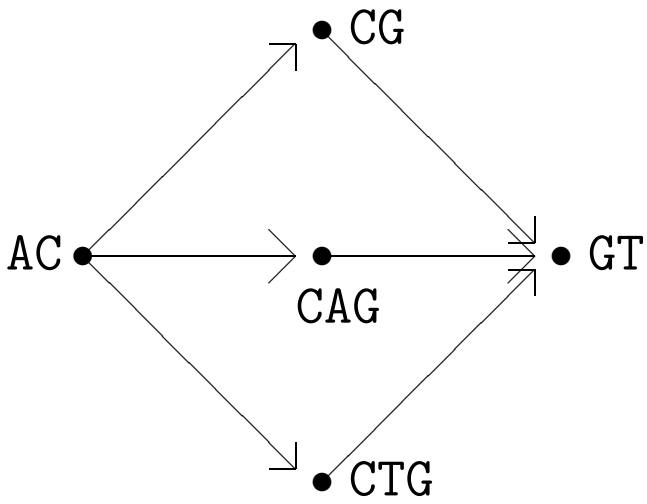


with shortest common superstring `TTAAATA` or `TTATAAA`.

Exercise 7 from Setubal-Meidanis, p. 140:

> Find sequences that give rise to the following overlap graph, where only edges with positive weight are shown. The weights are yours to find/choose.





All weights are 1.

Exercise 12 from Setubal-Meidanis, p. 140:

> Let $\mathcal{F} = \{\texttt{ATC}, \texttt{TCG}, \texttt{AACG}\}$. Find the best layout for this collection according to the Reconstruction model with $\epsilon = 0.1$ and $\epsilon = 0.25$. Be sure to consider reverse complements.

If $\epsilon = 0.1$, mistakes are not allowed. If $\epsilon = 0.25$, a string of length 4 allows for one insertion, deletion or substitution. We find

```
ATC---          ATC-
-TCG--          -TCG
--CGTT          AACG
------          ----
ATCGTT          ATCG
```

respectively.

Exercise 13 from Setubal-Meidanis, p. 140:

> Let $\mathcal{F} = \{\texttt{TCCCTACTT}, \texttt{AATCCGGTT}, \texttt{GACATCGGT}\}$. Find the best set of contigs for this collection according to the Multicontig model with $\epsilon = 0.3$ and $t = 5$. (No reverse complements.)

The “minimum” overlap should have length at least 5. And at most 2 mistakes per string (of length 9).

There is a solution with one contig:

```
TCCCTACTT------
------AATCCGGTT
----GACATC-GGT-
```

```
TCCCGACATCCGGTT
    T     -
```

The consensus can be chosen as presented here, but there are some variations possible.

Exercise 18 from Setubal-Meidanis, p. 140:

> Find a polynomial time reduction of SCS to Reconstruction.

Or: transform a problem instance for the Shortest Common Superstring problem into a problem for the Reconstruction problem, in such a way that solutions "correspond" with each other.

Of course, take $\epsilon = 0$ to force exact matches.

Now we have to deal with the reverse complements. To this end, in every fragment $f \in \mathcal{F}$ (where $\mathcal{F}$ is the collection for the SCS problem), replace all bases x by uxv for suitable strings u and v.

Now take care to choose u and v such that uxv has no overlap with the reverse complement of uyv *and* uxv has no "proper" overlap with uyv.

So, in order to deal with the reverse complements, in every fragment $f \in \mathcal{F}$ (where $\mathcal{F}$ is the collection for the SCS problem), we replace all bases x by `AAC`x`CC`. So, e.g., `AT` becomes `AACACCAACTCC`.

Now, in all contigs for Reconstruction, fragments must occur in the same direction: note that the reverse complement of any fragment will contain `TT` and `GG` at the ends (that never occur in the adapted fragments in the original direction), and there is no "self-intersection".

Exercise 1 from Setubal-Meidanis, p. 139:

Suppose we have the following fragments:

$f_1 =$ ATCCGTTGAAGCCGCGGGC
$f_2 =$ TTAACTCGAGG
$f_3 =$ TTAAGTACTGCCCG
$f_4 =$ ATCTGTGTCGGG
$f_5 =$ CGACTCCCGACACA
$f_6 =$ CACAGATCCGTTGAAGCCGCGGG
$f_7 =$ CTCGAGTTAAGTA
$f_8 =$ CGCGGGCAGTACTT

And we know that the length of the target molecule is about 55. Assemble these fragments and obtain a consensus sequence. Think of reverse complements.

One possible assembly is:

```
CCTCGAGTTAA-----GCCCGCGGCTTCAACGGAT---------------
-------TTAAGTACTGCCCG------------ATCTGTGTCGGG-----
---------AAGTACTGCCCGCG-------------TGTGTCGGGAGTCG
-CTCGAGTTAAGTA---CCCGCGGCTTCAACGGATCTGTG----------
```

```
CCTCGAGTTAAGTACTGCCCGCGGCTTCAACGGATCTGTGTCGGGAGTCG
```