Solutions Exercise Set 3
Author: Charmi Panchal

Problem 1: Suppose we have following fragments:
f1 = ATCCGTTGAAGCCGCGGGC
f2 = TTAACTCGAGG
f3 = TTAAGTACTGCCCG
f4 = ATCTGTGTCGGG
f5 = CGACTCCCGACACA
f6 = CACAGATCCGTTGAAGCCGCGGG
f7 = CTCGAGTTAAGTA
f8 = CGCGGGCAGTACTT
We know that the total length of the target molecule is about 50bp and may be ready to accept a solution of length between 45 and 55 bp. Assemble these fragments and obtain a consensus sequence. Be prepared to deal with errors. You may also have to use the reverse complement of some of the fragments.

Solution:

Fragments	Reversed and complemented
f1 = ATCCGTTGAAGCCGCGGGC	GCCCGCGGCTTCAACGGAT
f2 = TTAACTCGAGG	CCTCGAGTTAA
f3 = TTAAGTACTGCCCG	CGGGCAGTACTTAA
f4 = ATCTGTGTCGGG	CCCGACACAGAT
f5 = CGACTCCC GACACA	TGTGTCGGGAGTCG
f6 = CACAGATCCGTTGAAGCCGCGGG	CCCGCGGCTTCAACGGATCTGTG
f7 = CTCGAGTTAAGTA	TACTTAACTCGAG
f8 = CGCGGGCAGTACTT	AAGTACTGCCCGCG

Starting with f2' and f3
CCTCGAGTTAA
TTAAGTACTGCCCG
CCTCGAGTTAAGTACT GCCCG
AAGTACTGCCCGCG (f8')
CCTCGAGTTAAGTACTGCCCGCG
CTCGAGTTAAGTA (f7)
CCTCGAGTTAACTACTGCCCGCG
GCCCGCGGCTTCAACGGAT (f1')
CCTCGAGTTAACTACTGCCCGCGGCTTCAACGGAT
CCCGCGGCTTCAACGGATCTGTG (f6')
CCTCGAGTTAACTACTGCCCGCGGCTTCAACGGATCTGTG

CCTCGAGTTAACTACTGCCCGCGGCTTCAACGGATCTGTG (last superstring)
ATCTGTGTCGGG (f4)
TGTGTCGGGAGTCG (f5)
CCTCGAGTTAACTACTGCCCGCGGCTTCAACGGATCTGTGTCGGGAGTCG

Problem 2: Let F={ATC, TCG, AACG}. Find the best layout for this collection according to the sequence reconstruction model, with error level e=0.1. The same problem for e=0.25. Be sure to consider also reverse complements.

Solution:

With e = 0.1

-For f1= ATC it matches well with substring a = ATC, d(f1,a) = 0<0.3 (because|f1|= 3 and e*|f1|=0.3).

-For f2= TCG it matches well with substring a = ATC, d(f2,a) = 0<0.3 (because|f2|= 3 and e*|f2|=0.3).

-For f3= AACG, e*|f3|=0.4.

If f3= AACG is considered and it is matched with substring a = ATCG, then d(f3,a) = 1≮0.4 (because|f3|=4).

Taking reverse complement f3' = CGTT.

f3' matches well with a=CGTT (see the table with error level 0.1), d(f3',a) = 0.

A	T	C	
	T	C	G
A	A	C	G
A	T	C	G

Error level 0.25
One error is allowed.

A	T	C			
	T	C	G		
		C	G	T	T
A	T	C	G	T	T

Error level 0.1
No errors are allowed.

Considering e= 0.25.

-For f1= ATC it matches well with substring a = ATC, d(f1,a) = 0<0.75 (because|f1|= 3 and e*|f1|=0.75).

-For f2= TCG it matches well with substring a = ATC, d(f2,a) = 0<0.75 (because|f2|= 3 and e*|f2|=0.75).

-For f3= AACG,

If f3= AACG is considered and it is matched with substring a = ATCG, then d(f3,a) = 1=1 (because|f3|=4 and e = 0.25, e*|f3| = 1).

Problem 3:

a) You want to use the sequencing by hybridization method (SBH) to sequence a DNA fragment. For this, you are using a DNA array containing all DNA sequences of length 3 and test which of these sequences bind to your target. As a result, you find out that the target sequence has the following substrings of length 3:

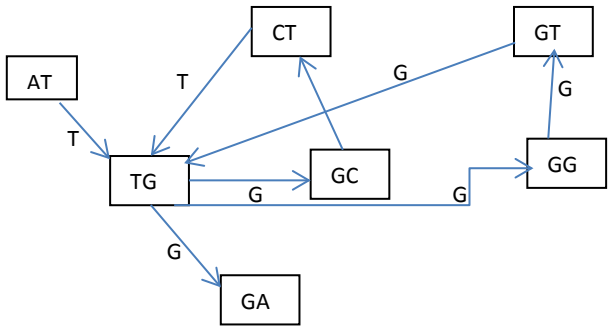
{ ATG, CTG, GCT, GGT, GTG, TGA, TGC, TGG } Find at least 2 DNA sequences validating this data.

b) How many solutions do you have if, using a DNA array containing all sequences of length 4, you obtain that the target sequence has the following substrings of length 4:

{ATGG, CTGA, GCTG, GGTG, GTGC, TGCT, TGGT } ?

Solution:

(a)



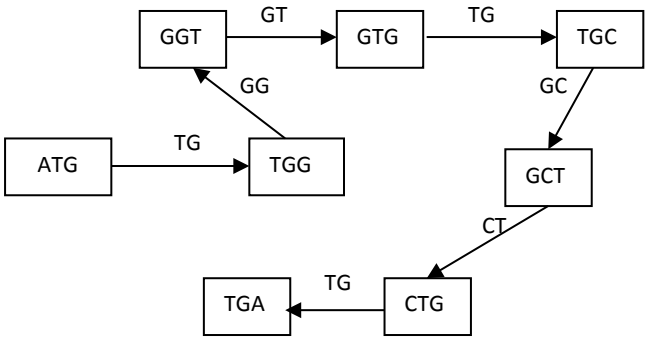
Following the Eulerian paths

AT→TG →GC→ CT→ TG→ GG→GT→TG→GA One possible sequence ATGCTGGTGA

AT→TG→ GG→ GT→TG→ GC→ CT→TG→GA Another possible sequence ATGGTGCTGA

(b)

{ATGG, CTGA, GCTG, GGTG, GTGC, TGCT, TGGT}



One Eulerian path possible

ATG→TGG→GGT→GTG→TGC→GCT→CTG→TGA

Solution:

ATGGTGCTGA

Problem 4: You are assembling a DNA sequence containing a repeat of the form XXX. Having given the fragments ATG, CTTGAT, TGT, TGTCA, TCAGAT, TGTAACT, find at least two such DNA sequences knowing that

- No fragment is included into some other

- The fragments provide "good linkage", in the sense that all fragments (except the one covering the ends of the sequence) overlap with at least one fragment at left and with another at right.

Solution:

Fragments	Reversed and complemented
ATG	CAT
CTTGAT	ATCAAG
TGT	ACA
TGTCA	TGACA
TCAGAT	ATCTGA
TGTAACT	AGTTACA

One possible solution:

```
A T C A A G
_ _ _ _ A G T T A C A _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ C A T _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ T G T _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ T G A C A _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ A T C T G A
A T C A A G T T A C A T G T G A C A T C T G A
```

Other possible solutions :

Fragment 4 and 5 – Alignment 1

```
T G T C A _ _ _
_ _ T C A G A T
T G T C A G A T
```

Fragments 2 and 6 – Alignment 2

```
_ _ _ _ _ C T T G A T
T G T A A C T _ _ _ _
T G T A A C T T G A T
```

Consider above two alignments. It is clear that TGT must be part of the repeat.
One solution could be :

```
T G T C A G A T _ _ _ _ _ _ _ _ _ _ _ _
            T G T A A C T T G A T
_ _ _ _ _ _ _ _ _ _ _ _ _ _ - - A T G _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ T G T
T G T C A G A T G T A A C T T G A T G T
```

Another solution :

```
T G T A A C T T G A T _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ A T G _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ T G T C A G A T _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ T G T
T G T A A C T T G A T G T C A G A T G T
```

More possibilities:

Result 3: ATG + Alignment 1 + TGT + Alignment 2: **AT**GTCAG**AT**GTGTAACTTG**AT** – length 21.
Result 4: ATG + Alignment 2 + TGT + Alignment 1: **AT**GTAACTTG**AT**GTGTCAG**AT** – length 21.

Problem 5.

a) Assemble the following error-free fragments using the shotgun approach: ATGTG, GCCGCA, GTGCCG, TGTGCC.

b) The same problem as above, replacing the second fragment above with CCCGCA. Assemble the fragments using the shotgun approach. Assemble the fragments using also the SBH-style shotgun approach. Compare the results and also with the result obtained at a), knowing that fragment CCCGCA had one substitution error – the correct one was the second fragment in a).

Solution:

(a)

```
    T G T G C C _ _ _
_ _ G T G C C
A T G T G _ _ _ _ _
_ _ _ _ G C C G C A
A T G T G C C G C A
```

(b)

Replacing the second fragment with CCCGCA.

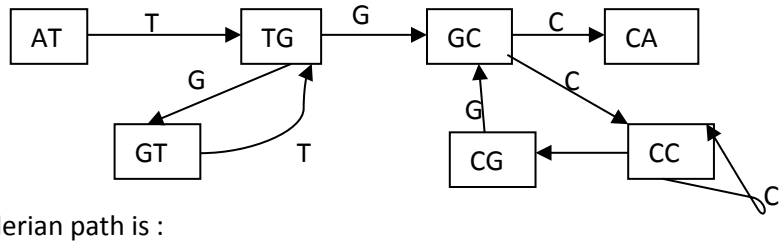
```
        _ _ _ T G T G C C _
        _ _ _ _ G T G C C G
        _ _ A T G T G
C C C G C A _ _ _ _ _ _
C C C G C A T G T G C C G
```

SBH-style shotgun approach:

First making substrings of length 3.

{ATG, TGT, GTG, CCC, CCG, CGC, GCA, TGC, GCC}.

Build a graph with nodes all substrings of length 2. Corresponding graph is as follows:



Eulerian path is :

AT→TG→GT→TG→GC→CC→CC→CG→GC→CA

ATGTGCCCGCA

Comparison	
Result obtain in (a)	ATGTGCCGCA
Same as (a) with replacing the second fragment with CCCGCA	CCCGCATGTGCCG
SBH-style shotgun approach	ATGTG**C**CCGCA

Solution obtain with SBH-style is very close to the one obtain in (a).
SBH approach is less sensitive to the errors than shotgun approach.