Fragment Assembly of DNA (Ch 4 of Setubal and Meidanis)

Goal: Sequence DNA of length 30,000 to 100,000 base pairs

Problem: Can only sequence DNA of length less than 700 base pairs.

Solution: Break the long DNA sequence into 500 to 2000 fragments of length 200-700 base pairs, sequence these shorter fragments, and use discrete mathematics to determine the sequence of the original long piece of DNA.

We will illustrate this process with an example. Note, however, the problem and techniques have been greatly simplified and altered from actual practice.

Example 1:

The DNA sequence
AGACCGCGTATAGCATCAGCATGACGCGTAGCACTA
is randomly broken into the following 5 pieces

AGACCGCGTATAG
CGTATAGCATCAG
TAGCATCAGCATGACGCGT
GCGTAG
TAGCACTA

In order to sequence a long piece of DNA, it must be cut into many smaller pieces. These smaller pieces are then sequenced. The order of the smaller pieces within the longer original piece must then be determined in order to determine the sequence of the original longer piece of DNA.

Suppose you did not know the original DNA sequence, but you did know the sequences of the 5 pieces. Could you reconstruct the original DNA sequence?

We can construct a directed graph as follows:

i.) The vertices represent the DNA pieces

ii.) If there is an overlap between the end of piece A and the beginning of piece B, then an arc is drawn from piece A to piece B.

iii.) Each arc is given a weight corresponding to the length of the overlap.

AGACCGCGTATAG

CGTATAGCATCAG

TAGCATCAGCATGACGCGT

GCGTAG

TAGCACTA

To find the original DNA sequence, we look for a Hamiltonian path (a Hamiltonian path is a path that visits each vertex exactly once). We will assume that the Hamiltonian path with the largest sum of arc weights corresponds to the correct sequence. You can read more about how fragment assembly is really done in Ch 4 of Setubal and Meidanis.

Fragment Assembly of DNA Homework

1.) Draw the directed graph corresponding to the following fragments. Find a Hamiltonian path with the largest weight sum. Use this path to determine the sequence of the original DNA sequence. Is there more than one Hamiltonian path with the largest weight sum?

AGCAGTAG
GAGCGAT
AGTCGTAG
GATCACAG

2.) Draw the directed graph corresponding to the following fragments. Find a Hamiltonian path with the largest weight sum. Use this path to determine the sequence of the original DNA sequence. Is there more than one Hamiltonian path with the largest weight sum?

ACTGGATT
ATTATG
TGTT
TTAACT
TTCTACT

3.) To find the original DNA sequence, what are you looking for (i.e., what kind of graph theory problem is this?)?

4.) Is the solution always unique?

5.) How is this problem similar to and different from the traveling salesperson problem?