Sequencing and Mapping

The genome of an organism is the complete set of the DNA sequences that constitutes its total genetic information content in a cell. This information is wrapped up in a set of **chromosomes** in the cell. The eukaryotes also have an additional set of **extrachromosomal** genes. These are located outside the nucleus of the cell within the energy producing organelles called **mitochondria**. For plants and algae, there are genes located in the **chloroplasts**. By the word genome, we usually mean the **nuclear genome**. For prokaryotic cell, the genome is a circular DNA molecule. For eukaryotes, like human, the genome consists of a set of linear DNA molecules contained in different chromosomes. In most eukaryotes, there are two copies of each chromosome, and hence two copies of each gene. This is called the **diploid** complement. The nucleus of a **haploid** contains only one copy of each chromosome, found only in reproductive cells. The number of chromosomes in a genome is characteristic of a given species. The following table gives examples.

Organism	Genome Size(kb)	No. of Chromosomes	Avg. no. of DNA/chromosome
Prokaryotes			
E.Coli	4 000	1	4000
Eukaroytes			
Yeast	20 000	16	1250
Fruit Fly	165 000	4	41 250
Human	3 200 000	23	130 000
Mouse	3 454 200		
Maize	15000 000	10	1 500 000
Salamander	90 000 000	12	7 500 000
Puffer Fish	375 000		

Obviously, genome size does not predict the complexity of the organism and also there is no direct correlation between the genome size and the number of chromosomes. It is generally true that it takes more genes to make the species more complex but there are also other factors. About 2-3% of the human nuclear genome actually takes part in the production of proteins. Even if we ignore the introns, apparently 70 to 80% of the genome is unused. This paradox may be due to the existence of highly repetitive DNAs.

In order to understand the structure and functions of the genome, we need to first extract the complete base-pair sequence in the chromosomes. The goal of the Human Genome project was to obtain this complete DNA sequence information. The process of obtaining this information is called **sequencing**. Current available biotechnology does not allow sequencing a DNA molecule having more than a few hundred bp (less than 1000 bp).

Before the genome project was started, biologists started sequencing thousands of mRNAs corresponding to coding genes. The process iinvolved first purifying mRNA, then obtaining complementary DNA (cDNA) by reverse transcriptase. Sequencing the

cDNA gives immediate information of the DNA of the original gene. However, the cDNA fragment containing a gene is considerably smaller than the genomic DNA.
This difficulty has given rise to several challenging problems in computational biology. We will discuss these issues in this chapter.

Before we do that we need to briefly review some of the molecular biology laboratory techniques that have been developed during the last few decades.

DNA Sequencing : separating DNA segments according to size (Gel Electrophoresis)

The DNA sequence can be read by a technique called *gel electrophoresis* which separates DNA molecules into groups depending on their lengths. Gel electrophoresis has high resolution; even fragments which differ by a single nucleotide can be separated. The sample molecules are placed in a gel under the influence of an electric field. The DNA or RNA molecules (which are slightly negatively charged) can migrate towards the positive electric field. The speed of migration is inversely proportional to the length of the molecule; longer molecules move slow, shorter move faster. All molecules are initially placed at the top of the 'well' and after a few hours, the molecules move to different locations depending on its length. If the molecules are labeled with radioactive isotopes, their positions can be photographed on a film.

DNA or a RNA molecule can be sequenced using these techniques as follows. Given a DNA molecule, obtain all fragments that end in a single letter *A.* Similarly, obtain all sequences ending in *T, C* and *G.* For example, if the sequence is GATTCGGATTTACT the fragments that end in *T* are *GATGATT, GATTCGGAT, GATTCGGATT, GATTCGGATTT* and the whole sequence *GATTCGGATTTACT.* These subsequences are formed by special enzymatic chemical reactions in presence of DNA polymerase and ddATP, ddTTP, ddCTP and ddTGP which are used as ingredients to start copying the DNA sequence. The replication of the DNA sequence is stopped at positions occupied by the four bases *A,T,C* and *G* by base analogs for each of the individual wells. The sequences are also labeled with a primer at the beginning. In modern automated sequencing, the primer is replaced by a different fluorescent probes and the signals from the probes are detected by special detectors. After a period of incubation, these sequences are then placed in four wells, the *A*-well, the *T*-well, the *C*-well and the *G*-well and subjected to electric field simultaneously. We can conclude the precise sequence of the original fragment. The figure below illustrates the principle. We assume here that the positive terminal is on the top and the shorter fragments leave their mark near the top.

	A	*T*	*C*	*G*
G				-------
A	-------			
T		-------		
T		-------		
C			-------	

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

If you now read the horizontal bars from top to bottom corresponding to the wells, you will get the entire sequence *GATTCGGATTTACT* For further details, see http://web.utk.edu/~khughes/main.htm

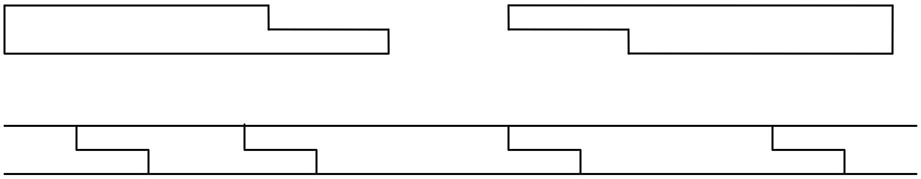
The gel electrophoresis technique was developed in 1970 by Maxam and Gilbert and Sanger. Since the method obtained the DNA fragments by chemical degradation of part of the sequence, it was not very reliable. A more efficient and reliable method is to use PCR which we describe next.

Cutting a DNA Sequence -- Restriction Enzymes

DNA is a long molecule. In order to be able to sequence it piece by piece, a biological pair of scissors is needed. Around 1973, Smith et. Al [] made a startling discovery in the course of their study of defense mechanism of bacterial cells from viral attack. They observed that certain bacteria produced enzymes that can cut or break a double stranded DNA at specific points. These proteins, called **restriction enzymes** can catalyze the **hydrolysis** of DNA (the process of breaking a molecule by adding water) at specific points called **restriction sites** that are determined by a specific sequence of base pair. The first such enzyme discovered called *Eco*RI could **cleave** or **digest** DNA molecules between *G* and *A* whenever it encountered the sequence 5'-*GAATTC-3'*. Note that the sequence is its own reverse complement, that is, if you read the single strand in the 3'-5' direction, you get the same sequence *GAATTC.* Such reverse complement sequences are called **palindromes.** So, whenever such a sequence appears in one strand, it also appears in the other strand. Since the cuts are made in both strands between *G* and *A*, the remaining DNA pieces have **sticky ends.**

5' ....*GAATTC*............*3'*
3'..........CTTAAG............5'

5'...........G   AATTC.........3'
3'...........CTTAA   G.........5'



Chromosomal DNA and restriction enzyme cutting sites

The sticky ends themselves are naturally complementary to each other. This favors re-linking with another DNA piece cut with the same enzyme with the help of another glue enzyme called **ligase**. It is also possible to mix DNA from two different sources that have both been cut by using the same restriction enzyme. This allows combining fragments from two distinct DNA. Thus, restriction and ligase enzymes are nature's way of providing "cut and paste" editing facility for DNA sequences and have been used in **genetic engineering for recombinant DNA**. Even for the same DNA, the cut pieces may join together in different combinations generating overlapping DNA fragments. These are also recombinant DNA and can be cloned for further processing.

There are also restriction enzymes that do not create sticky ends, they create **blunt ends** and such blunt cuts can also be ligated with other blunt-ended DNA molecules. In particular, small oligonucleotides can be legated at the blunt ends to have almost arbitrary combination of DNA ends. Since the discovery of *EcoR*I, more than 300 restriction enzymes have been found in other bacterial species and have been used in laboratories. These are mostly 4-, 6- or 8-cutters; it is rare to find an odd cutter since the palindromes must be of even length. Finally, the restriction enzymes are sometimes called **endonucleases** because they cut the DNA in the middle of the sequence. There are enzymes called **exonulceases** that cut a DNA from only one end.
Recently, PCR technology has replaced the restriction endonucleases in many applications, they are still used extensively in laboratories for routine subcloning and diagnostic purposes. Restriction enzymes cut the DNA into many different sizes of fragments ranging from 256 bp to 1 million bp.

DNA molecules can also be broken down into *random* pieces by subjecting a solution of purified DNA to rapid mechanical vibrations. The fragments are then filtered; multiple copies are made by cloning, and then sequenced by gel electrophoresis (or microarrays). Finally, the fragments are assembled to get the entire DNA sequence. We will describe each of these processes now.

DNA Cloning

In order to study a specific fragment of DNA sequence, we need to select the fragment and *amplify* it so that the solution contains a purified near-homogeneous population. The technique inserts the DNA piece in a ***vector***. A naturally occurring vector is a ***plasmid*** which is a circular DNA found in bacteria. Plasmids can infect bacteria such as *E.Coli*.Cutting plasmids with a restriction enzyme that has also been used to cut the DNA creating compatible sticky end. This allows formation of **recombinant plasmids**. The resulting molecule is then inserted into a suitable host (a bacteria or yeast cell) and the organism multiply under suitable conditions (temperature and nutrients), producing a colony of identical cell clones. The host is then killed and the resulting DNA pieces extracted and sequenced. The method is explained in the following slide:

Vectors for cloning vary depending on the size of the DNA to be cloned. There are many types of cloning vectors available allowing varying sizes of DNA inserts to be amplified. The table below gives a partial list. This includes plasmids, viruses, yeast artificial chromosomes (YAC) and bacterial artificial chromosome (BAC) which were used to create overlapping clones for sequencing human genome. The details of the laboratory techniques to produce a purified clone set containing a specific DNA fragment as an insert in the vectors are not covered in this course.

Cloning Vector	Insert Size
Bacteriophage M13	1.5 kb
Plasmid	5 kb
Bacteriophage λ	25 kb
Cosmid	40 kb
BAC(bacterial artificial chromosome)	150 kb
YAC(yeast artificial chromosome)	500 kb

Polymerase Chain Reaction (PCR)

Restriction enzymes and plasmid cloning techniques are used routinely in many laboratory experiments. The discovery of PCR has replaced these techniques for large scale sequencing of genomes. Without PCR automated and fast sequencing technology would not have ben possible.

PCR is a cell-free method of amplifying a short (<15kb) fragment of a target DNA in large quantities. People have compared PCR with the Gutenberg printing press of DNA and Kary Mullis who invented PCR in 1983 got Nobel Prize. He thought it was a good idea because he “had been spending a lot of time writing computer programs”. PCR is a laboratory application of the concept of “recursion” in computer science.

PCR technique depends on the existence of a ***primer sequence*** of 15-30 nucleotides long at the end of a target DNA . When added to a ***denatured*** DNA (single stranded DNA at temperature >91^0 C), the primers will bind to complementary sequences if the temperature is now cooled to =50^0 C. This process is called ***annealing***. Under the presence of a DNA polymerase at =72^0 C, the synthesis of new DNA strands complementary to both strands of the target DNA will start. PCR is called a “chain reaction” because both the newly synthesized DNA strands now act as templates for future iterations, doubling the number if DNA fragments at every cycle. This results in a huge quantity of the DNA fragments in a short time.

For further details, see http://web.utk.edu/~khughes/main.htm.

Hybridization

Given a short (8- to 30-nuclreotides) synthetic fragment DNA, called *probe,* a target single stranded DNA molecule (produced by denaturing) will *hybridize* or bind to the probe if there is a substring in the target sequence that is complementary to the probe. For example, a target DNA sequence *CCCTGGCACCTA* will hybridize to a probe *ACCGTGGA* since the complementary sequence *TGGCACCT* is present in the target. In the mix of a DNA, the presence of a particular DNA can be tested by making the probe fluorescent or radioactive. This idea has lead to the development of DNA chips or *microarrays* that allows rapid DNA sequencing.

DNA Microchip: Microarrays- A large-scale Biological tool

There are a number of websites that give excellent presentations about microarray.
http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html
http://www.bio.davidson.edu/Biology/GCAT/ASCBpresentation/ASCB1.html
We will discuss here only the principle and show how a DNA sequence can be derived by using a microarray. The more important application of microarray is to determine the **expression level** of different genes rapidly and simultaneously. It can give information about gene expression of 30,000 or more genes in one experiment. It has also been used for proteomics and study of biological pathways. Another important application is diagnosis of a disease. For details read the websites. Microarray technology uses nanoscale machining and robotics. The data generated by microarrays have to be analyzed using mathematical techniques, pattern recognition, supervised learning, signal processing etc .Thus, the development microarrays has been an interdisciplinary project. The device is simply an array of small circular spots on a silicon surface. The spots contain contain "probes" which may be a DNA fragment, proteins or small molecules. We will discuss two basic techniques:

1. **Sequencing DNA using microarrays**
1) The microarray is loaded with all "probes" (possible DNA sequences) of length *l* in its array of spots.
2) Generate a hybridization solution containing many copies of fluorescently labeled DNA target fragment. The DNA fragment hybridizes with those probes that are complementary to substrings of length *l* of the fragment.
3) Detect probes that hybridize with the DNA fragment. This is usually done by using lasers.
4) Apply a combinatorial algorithm like shortest superstring problem to reconstruct the DNA sequence.
Example: *l*=4. The array looks like

	AA	AT	AG	AC	TA	TT	TG	TC	GA	GT	GG	GC	CA	CT	CG	CC
AA																
AT			atag													
AG																
AC												acgc				
TA											tagg					
TT																
TG																
TC																
GA																
GT																
GG													ggca			
GC	gcaa															
CA	caaa															
CT																
CG																
CC																

Target DNA TATCCGTTT (complement of ATAGGCAAA)

```
**
A T A G G C A A A
A T A G
  T A G G
    A G G C
      G G C A
        G C A A
          C A A A
```

The array provides only information about the *l-mers or l-grams* present but does not provide any clue where these *l-mers* are present in a sequence. Obtaining the above alignment is done by a **shortest superstring algorithm**. The shortest superstring problem is NP-complete but there is a greedy algorithm that is at most four times optimal.

Gene Expression Determination using DNA Micropchip

Although each cell of human body contains the same genetic material, not all cells produce the same proteins. For example the proteins synthesized by a muscle cells are quite different from those that create growth of hair. If a gene is active in a cell, it is **expressed**. By studying the expression levels of different genes in a cell the biologists can understand the function of a cell. Micrarrays can determine the expression levels of thousands of genes simultaneously. The set up is as follows.

The micorarry spots are filled up with functional DNA. Each spot may contain an oligonucleotide or a cDNA fragment which is characteristic of an mRNA. Then, mRNA from a cell or cell population are labeled with fluorescent tags and allowed to hybridize with the cDNAs. If the cDNA is complimentary to a substring of the mRNA, the mRNA will hybridize to that spot. A strong intensity of fluorescent light indicates that a high level of mRNA hybridizes to that spot and therefore the gene which is characterized by that spot is very active in that cell or cell population. Conversely, if the spot is dark it means that the gene for that spot is inactive and a moderate amount of intensity means that the gene is somehow active. The level of intensity is measured by a laser beam and the intensity levels are pre-calibrated for the activity level. For further details, visit the websites mentioned earlier.

Cancer Detection Using Microchip

Cancers that are caused by mutations of genes (such as BRCA1 and BRCA2 which are responsible of 60% of breast and ovarian cancer) can be identified by microchip easily. Since a large gene has several possible places where mutations may occur to cause diseases other than cancer, it is a difficult task to pinpoint exactly which mutation is responsible for cancer by trial and elimination.

First, the DNA chip is filled with synthetic single stranded DNA sequences that are found in a normal targeted gene such as BRCA1. To determine whether an individual has a specific BRCA1 cancer mutation, a blood sample is collected from the patient and also from a normal person that does not have any of BRCA1 and BRCA2 mutations. Single stranded DNAs are derived from both of these samples, cut into pieces of appropriate size and labeled with colored dyes, the patient's DNA is dyed green and the normal person's DNA is dyed red. The leveled sets of DNA are then mixed together and allowed to hybridize with the array. If the patient does not have any mutation on the BRCA1 gene (that is it is a normal gene) then both green and red DNA fragments will hybridize in equal proportion to the spots in the microarray. If the patient has cancer, the normal DNA (red) will still hybridize but the green DNA will not hybridize in locations of spots where fragments contain the regions of mutations. The biologist can then further investigate these regions of the gene of the patient for further analysis of risk factors. Micrarrays have been used for assessing risk factors for diseases like cancer, heart disease and diabetics.

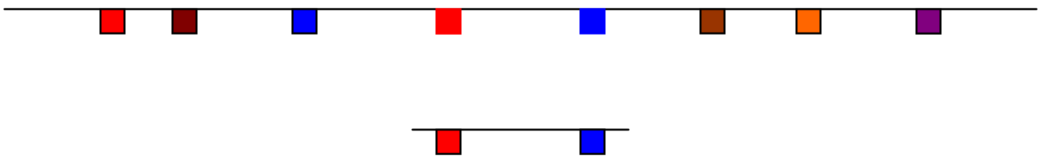
Further use of DNA molecules includes protein microarays, study of biological pathways, and macromolecular interactions.

Mapping

In our lectures on similarity search and multiple alignments, we made the assumption that the DNA or protein sequences are given to us in a data base. We will now discuss how to obtain the molecular sequences. This process is called mapping. There are three types of mapping discussed in the literature: cryptogenic mapping, genetic mapping and physical mapping.

Physical mapping

The goal of the Human Genome project was to obtain a complete physical mapping of the three-billion-nucleotide long DNA transcript of the human genome. Physical mapping attempts to establish the physical locations of specific markers or landmark features such as genes or motifs. A class of landmarks of interest are STSs (*sequence-tagged sites*) or ESTs (*expressed sequence tags*) . An STS is a DNA string of length 200-300 bp whose right and left ends of 20-30 bp each, occur only once in the genome. Thus each STS occurs uniquely in the DNA sequence. An early goal of Human Genome project was to select and map a set of STSs such that any substring in the genome of 100,000 bp or more contains at least one such STS. A more refined goal was to map the ESTs which are STSs that come from genes rather than introns that code certain proteins. The ESTs are derived from mRNA and cDNA. Suppose, we have a map of a chromosome with these markers…that is, we know the order in which these markers appear in the chromosome with a rough resolution of 100,000 bp as shown above. Suppose now a segment S has been sequenced with two STSs that contain a gene. If the map is available, by comparing the two STS tags, we know where the gene appears in the chromosome. This is the underlying idea of producing the *physical maps* of the genome.

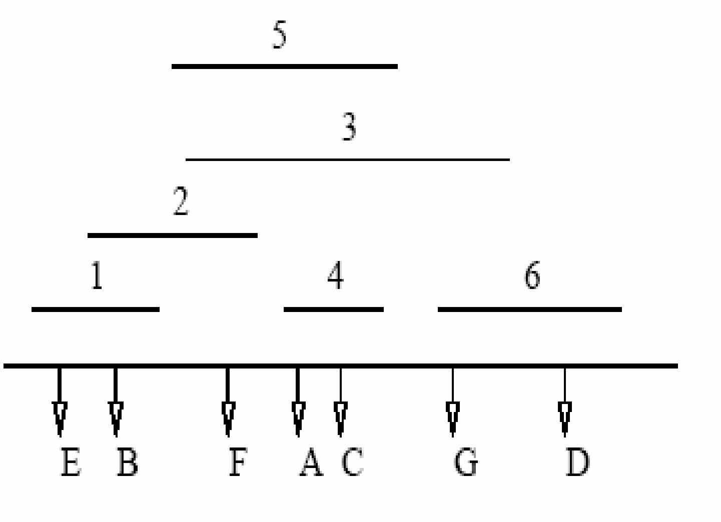


STS-content Mapping

The occurrences of STS can be determined by exact pattern matching algorithms. To be of greater value, the exact locations of the STSs in the genome must also be known.
A *clone library* consists of a set of short DNA segments, called *clones* (created by techniques discussed earlier) that originate from a stretch of a DNA of interest. The relative ordering of the clones or their exact positions in the DNA sequence may be unknown. The clones may overlap and should cover the entire DNA sequence. It will be useful to create *an ordered clone* library which gives theorder as well as locatins of the clones. **The STS-content mapping** simultaneously solves two problems: finds the exact locations of the STSs and order the clones in the clone library.

Given a clone library and a set of STSs on a DNA segment, hybridization or PCR can be used to determine the occurrences of STSs on the clones. To determine which clones in the library contain which STSs, one can attempt hybridizing each STS to each clone. Alternately, PCR can be used to generate this data: the ends of each STS contain unique PCR primer templates which are known. Using this information, one can use the PCR technique to generate large amount of new DNA from the DNA in the clone. The particular STS is located in the clone if and only if the PCR can generate new DNA using the known primers. This experiment can be repeated for each clone/STS pair and the data can be tabulated in the form of a binary matrix as shown below.

If we know which STSs occur in which clone, we can compute the ordering of the STSs. We are assuming that the data is error free: the STSs are unique, the clones actually represent contiguous stretches of the DNA segment (they are not recombinant **chimera**). Under these conditions, if an STS occur in more than one clones, these clones must be occuring consecutively in the ordering. In the following example, let A-G be seven STSs on a DNA segment covered by six clones.



We represent the segment by a matrix *M* which has a row for each STS and a column for each clone. A 'X' is placed in entry *M*[*i*,*j*] if STS in row *i* occurs on clone *j*. The matrix for our example is:

	1	2	3	4	5	6
B	X	X				
C			X	X	X	
D						X
E	X					
F		X	X		X	
G			X			X

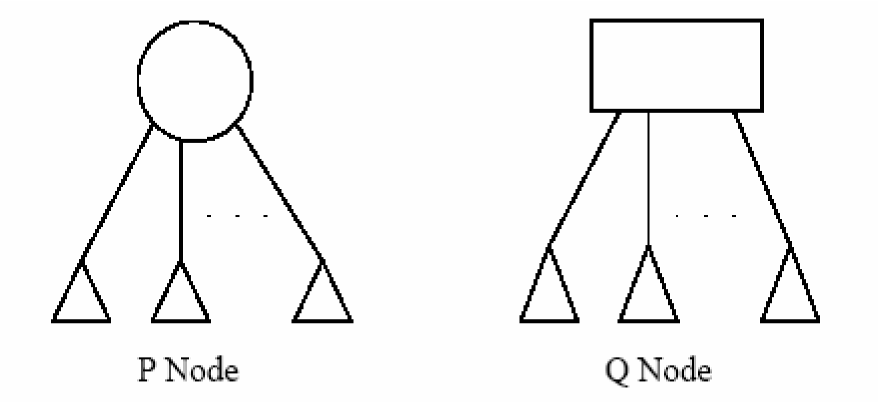
The X's in a single column are STSs on a single clone and therefore must be consecutive. Thus, if we could rearrange the rows such that they occur in the same order as the STSs occur, then in each column all the X's will be next to each other. So, any rearrangement of the rows which puts all the X's in a column in consecutive positions provides a solution to the mapping problem. Thus, one solution is

	1	2	3	4	5	6
E	X					
B	X	X				
F		X	X		X	
A			X	X	X	
C			X	X	X	
D			X			
G						X

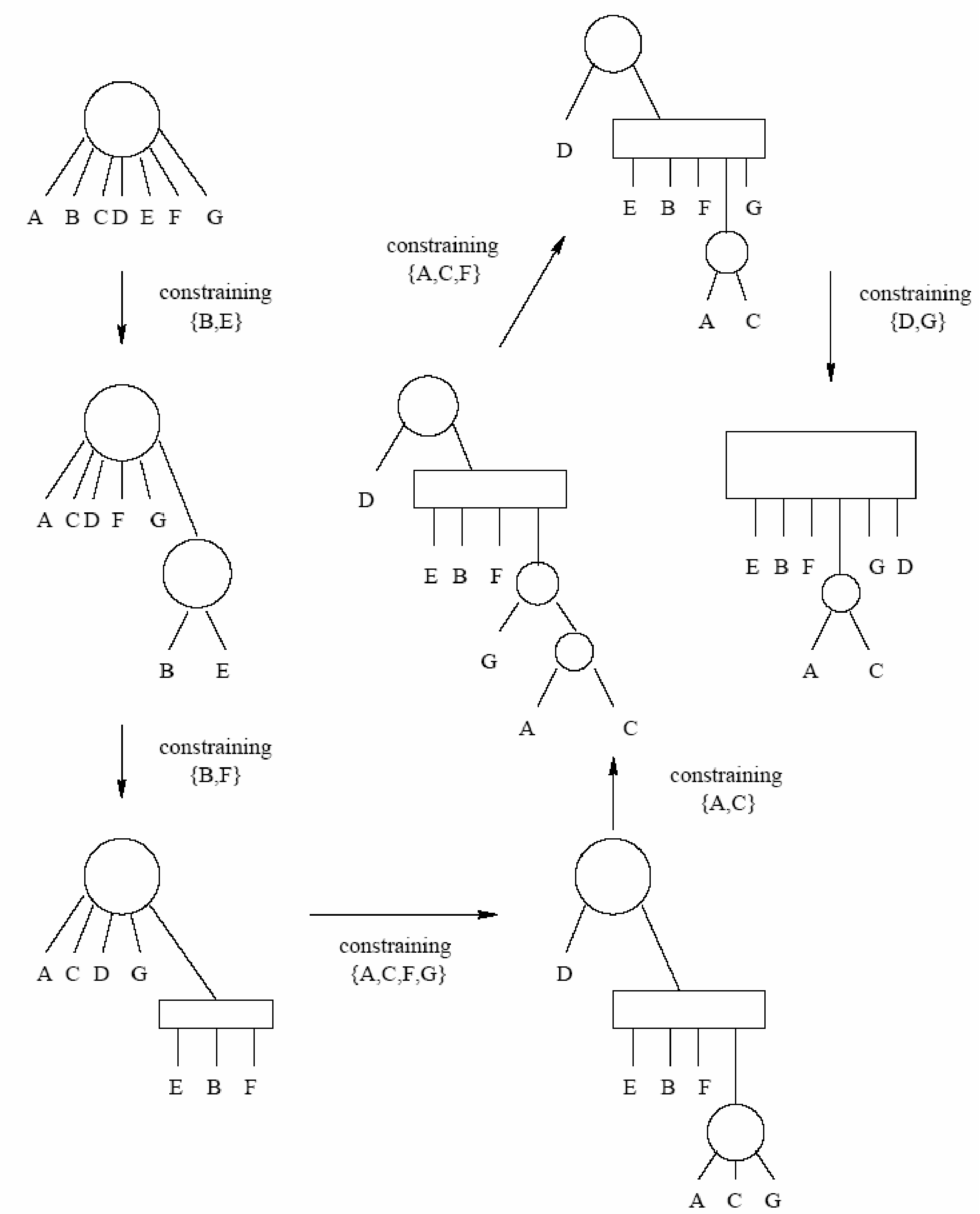
So, the ordering of the STSs is: EBFACGD. Another possible solution is EBFCAGD. The problem that we just presented is the so-called *consecutive ones problem* and the solution of this problem was presented by Booth and Lurker in 1976.

The PQ-Tree Algrithm

The PQ-Tree algorithm uses the matrix *M* information (assuming data is error free) and uses a data structure called *PQ-tree.* A *P-node* is one whose children can be rearranged in any order and a *Q-node* is one whose children remain consecutive but may be rearranged in reverse order.



The tree is initialized to to be a *P-node* with a child for each row, A through G. At each iteration, a column is selected which places a constraint on some subset of these nodes. For example, we know that B and E must be consecutive, so the algorithm creates a *P-node* with **B** and E as children. The algorithm proceeds column by column until all constraints are added. The steps are illustrated below. Four solutions can be found: EBFACGD and EBFCAGD and their reversals DGCAFBE and DGACFBE. The algorithm has linear run time complexity.



Imperfect Data

In general, there are three kinds of errors that can complicate the STS mapping problem:

1. False Positive: Hybridization or PCR experiments suggest that an STS overlaps a particular clone when it actually does not.
2. False Negative: the converse of 1)

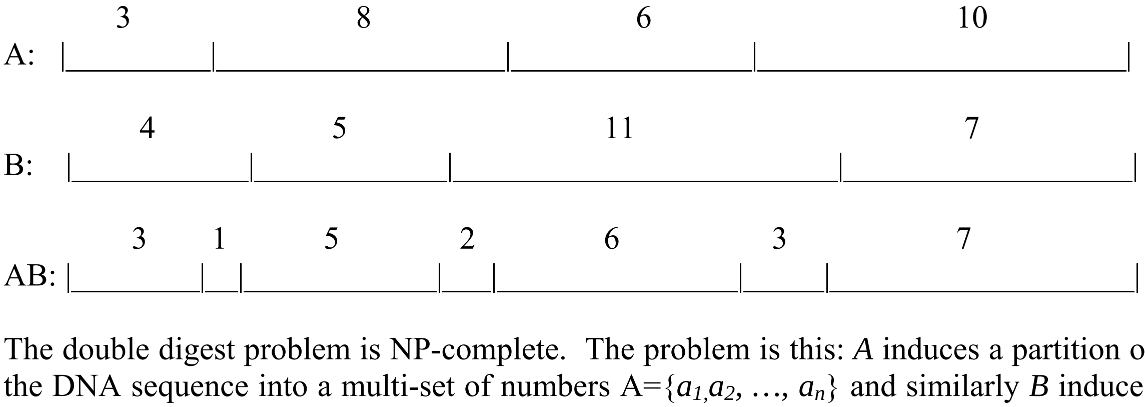
3. Chimeric Clones: This happens when two different DNA fragments coming from widely differing locations join and then behave as a single clone. This phenomenon destroys the consecutive one property. In some clone library, it has been observed that more than 50% of the clones are chimeric and the formation of chimeric clones are more frequent if the fragment sizes increase, like when fragments are inserted in *YAC* (yeast artificial chromosomes) where the length may be 500,000 bases or as high as 2,000,000 bases.

Such imperfections will result in the matrix M that will not have consecutive one property. The problem then becomes NP-hard and can be formulated as a traveling salesman problem.

(Further reading on this topic: Gusfield Sections 16.5, 16.6.)

Physical Mapping: Fingerprinting

Another way of mapping a long DNA sequence is to use **fingerprinting.** The target DNA is broken up into several fragments using restriction enzymes. The fragments (which may be very long) are not yet sequenced but it is possible to obtain a mapping by carefully determining their overlap information.. The restriction site mapping is used for finger printing. The objective is to locate the restriction sites of a given enzyme on the target DNA. The resulting fragments are then measured for lengths by gel electrophoresis. For example , suppose we have two restriction enzymes *A* and *B*, each recognizing an unique restriction site where the sequence is cut. Let us say after applying *A* on the DNA, the resulting fragment lengths ar 3000,6000,8000 and 10000 bp. Applying *B* we get the lengths s 4,5,7 and 11 thousand bp. Now if we apply both *A* and *B* at the same tme, the enzymes will act simultaneously. Let's say we obtain lengths tis time as 1,2,3,3,5,6 and 7 thousand bp. The problem now is: how can we order the segments so that the ordering is consistent with the experimental results. This is called a *double digest problem.* A solution to our example is shown below:



The double digest problem is NP-complete. The problem is this: *A* induces a partition of the DNA sequence into a multi-set of numbers A=$\{a_1,a_2, \ldots, a_n\}$ and similarly *B* induces a partition B=$\{b_1,b_2,\ldots,b_m\}$. From the double digestion, we get a third multi-set C= $\{c_1,c_2,\ldots,c_k\}$, where k, n, m are integers greater than 0, and $k>m$ *and n.* The problem is to find two partitions P_A and P_B such that if we superimpose the plots of P_A and P_B on the integer line, the resulting sub-intervals will be a partition P_C of C. This problem can be mapped on the *set partition problem* in polynomial time, which is known to be NP-complete.

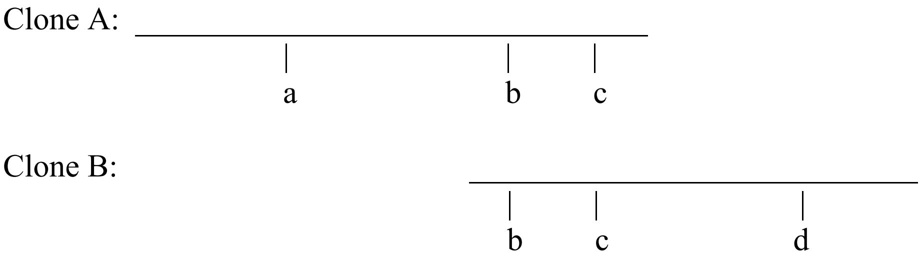
The formulation of the problem can be slightly relaxed by considering what is called a *partial digestion problem.* Here, we apply only one enzyme on multiple copies of DNA and let the experiment run for different amount of times. This yields fragments of different lengths depending on the cooking time or rather how long you allow the DNA strand to be digested. For example, using our *A* restriction enzyme for the above example, we get the following length sequences for partial digest experiment.

3,11,17,27: fragments between left end points and and other sites.
8,14,24:fragments between the first restriction site and all other downstream.
6,16: fragments between the second restriction site and all other downstream
10: the last fragment.

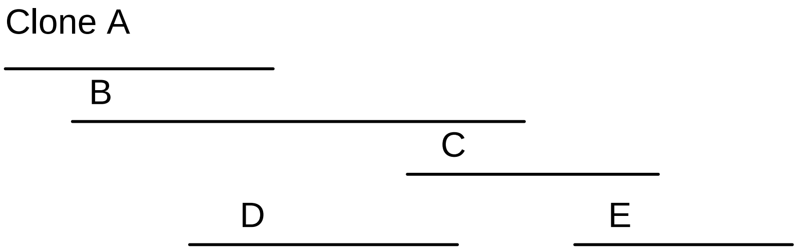
Is the problem NP-complete? It is not known. Most likely it is not since we have a lot of structures here: for any pair of combinations of restriction sites, we have a fragment. Can you find an efficient algorithm?

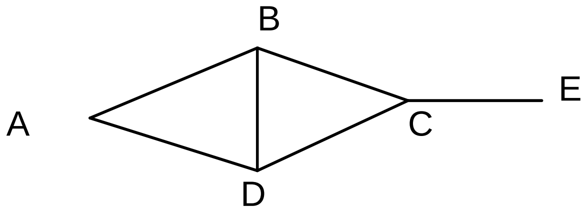
Mapping by Hybridization

The experiment consists of the following: The DNA is first broken up into fragments, each fragment is cloned and a library of thousands of clones made. Then hybridization is applied to the clones, that is, a small oligonucleotides are used as probes to hybridize at complimentary locations in the fragments. We can repeat this experiment with different probes. The set of probes that hybridize to a given fragment becomes the fingerprint of the clone. Two clones that share some fingerprints are have likely to have overlapping regions in the target (unless there are 'repeats"), as illustrated below with respect to probes *b* and *c*.



The computational model for hybridization mapping is an *interval graph* G (*V*,*E*) In this model, the vertices *V* represent clones and an edge in *E* is drawn between vertices if the pair of clones representing the vertices overlaps. An interval graph is shown below.





.

In practice the overlap information is not always certain. We need a second graph $G_t=(V,E_t)$ where E_t represents known plus unknown overlap information (that is some edges are drawn if there is a possibility of overlap while many other edges are left out because the experimental results indicate a high probability of not to have any overlap.). The problem is : Does there exist an interval graph $G_s=(V,E_s)$ such that $E \leq E_s \leq E_t$?

(to be continued)