Bioinformatics Support of Genome Sequencing Projects
Introduction
The Big Picture
Biology reminder
Enzyme for DNA manipulation
DNA cloning
DNA mapping
Sequencing genomes
Alignment of large-scale DNA
Genome sequencing is figuring out the order of DNA nucleotides.

Sequencing the genome is an important step towards understanding it.

The whole genome can't be sequenced all at once because available methods of DNA sequencing can only handle short stretches of DNA at a time.
- Make multiples copies of an existing DNA strain.

- Cut the resulting copies in multiples short overlapping fragments.

- Sequences it using automatic sequencers.

- Link together the sequenced fragments in the correct order to give the master sequences of the chromosomes that make up the genome.

The shotgun approach
THE STEPS OF THE LADDER ARE MADE OF PAIRS OF NITROGEN BASES:

- ADENOSINE = A
- GUANOSINE = G
- CYTIDINE = C
- THYMIDINE = T
Enzymes for DNA manipulation

DNA polymerases are enzymes that synthesize new polynucleotides complementary to an existing DNA or RNA template.
Nuclease degrade DNA molecules by breaking the phosphodiester bonds that link one nucleotide to the next.
Ligase: which join DNA molecules together by synthesizing phosphodiester bonds between nucleotides at the ends of two different molecules, or at the two ends of a single molecule.
Terminal deoxynucleotidyl transferase, obtained from calf thymus tissue, is one example of an end-modification enzyme.

Allow the attachment of radioactive, fluorescent or other types of marker to DNA molecules.
- Vector based on plasmid
- Vector based on virus
- Polymerase Chain Reaction aka PCR
Build a recombinant plasmid.

Use of the natural processes of transformation, enhanced significantly by suspending the cells in calcium chloride before adding the DNA, and then briefly incubating the mixture at 42 °.

Can handle DNA fragments up to 10 kb in size.
Plasmid based
With a $\lambda$ phage up to 18 kb of new DNA can be cloned.

A phage virus
The Polymerase Chain Reaction

PCR is a test-tube reaction and does not involve the use of living cells => way faster.

Need a primer to start the synthesis reaction.

Reaction is done in the 5’ -> 3’ direction.

With the primer we can choose the starting point of the synthesis.
heat the mixture to 94 °C so hydrogen bonds that hold together the two polynucleotides of the double helix are broken.

The temperature is then reduced to 50–60 °C to allow the primers to attach to their annealing positions.

The temperature is raised to 72 °C, the optimum for Taq polymerase.

Fragments up to 40 kb can be amplified.
Genetic and Physical Maps

Genetic mapping is based on the use of genetic techniques to construct maps showing the positions of genes and other sequence features on a genome.

Physical mapping uses molecular biology techniques to examine DNA molecules directly in order to construct maps showing the positions of sequence features, including genes.
(A) The DNA molecule contains a tandemly repeated element made up of many copies of the sequence GATTA.

(B) the DNA molecule contains two copies of a genome-wide repeat. When the sequences are examined, two fragments appear to overlap, but one fragment contains the left-hand part of one repeat and the other fragment has the right-hand part of the second repeat.
- Genes were the first markers to be used.

- Single nucleotide polymorphisms (SNPs)
The first genetic maps, constructed in the early decades of the 20th century for organisms such as the fruit fly.

By 1922 over 50 genes had been mapped onto the four fruit-fly chromosomes, but nine of these were for eye color.

Problem: microbes, such as bacteria and yeast, have very few visual characteristics.

The standard blood groups such as the ABO and immunological proteins such as the human leukocyte antigens (HLA system) were added.

Problem again: in most eukaryotic genomes the genes are widely spaced out with large gaps between them.
A single nucleotide polymorphism (SNP).

Single nucleotide polymorphisms (SNPs)

These are positions in a genome where some individuals have one nucleotide (e.g. a G) and others have a different nucleotide.

In the human genome there are at least 1.42 million SNPs.

SNP detection is rapid and powerful because it is based on oligonucleotide hybridization analysis.
Oligonucleotide hybridization

Completely base-paired hybrid is stable

Single mismatch – hybrid is unstable

Oligonucleotide

CTGGTGCAGTCTTTTATTT

GACCAGCAGTCAGAAATCAA

DNA target

SNP

Mismatches – base pair cannot form

CTGGTGCAGTCTTTTATTT

GACCAGCAGTCAGAAATCAA
The oligonucleotide probe has two end-labels. One of these is a fluorescent dye and the other is a quenching compound. The two ends of the oligonucleotide base-pair to one another, so the fluorescent signal is quenched. When the probe hybridizes to its target DNA, the ends of the molecule become separated, enabling the fluorescent dye to emit its signal.
A map generated by genetic techniques is rarely sufficient for directing the sequencing phase of a genome project because genetic maps have limited accuracy and a low resolution.

These two limitations of genetic mapping mean that for most eukaryotes a genetic map must be checked and supplemented by alternative mapping procedures before large-scale DNA sequencing begins.

Let's introduce a physical mapping technique: Restriction mapping.
The objective is to map the *EcoRI* (E) and *BamHI* (B) sites in a linear DNA molecule of 4.9 kb.

We do single and double restrictions on this DNA molecule.

The unresolved issue being the position of one of the three *BamHI* sites

By giving suboptimal condition for a single restriction experiment we can solve this issue.
Does the shotgun approach works with big genome?

1995: the sequence of the 1830 kb genome of the bacterium *Haemophilus influenzae* was published.

Sequence assembly required 30 hours on a computer with 512 Mb of RAM, and resulted in 140 lengthy contiguous sequences.
MUMmer: fast alignment of large-scale DNA

http://www.tigr.org/software/mummer/

MUMmer is able to align rapidly sequences containing millions of nucleotides using suffix tree data structure.

Human chromosome #1: 200 million base pairs
MUMmer highlights the exact differences between two genomes.
MUMmer locates SNP, repeats, tandem repeats, DNA insertion and MUM (maximum unique match)

1999: standard for sequence alignment was dynamic programming or hashing technique (BLAST, FASTA)
dynamic programming: space requirement O(n), time requirement O(n²).
Maximal Unique Sequences

- Sequences in genomes A and B that:
  - occur exactly once in A and in B
  - are not contained in any larger such sequence

- Genome A: tcgatcTAAGATCGA…ACCATCAcgact
- Genome B: gcattaTAAGATCGA…ACCATCAtccag
A trie is a type of tree that has $N$ possible branches from each node, where $N$ is the number of characters in the alphabet. The reason you don't hear much about the use of suffix tries is the simple fact that constructing one requires $O(N^2)$ time and space.
Building trees: $O(N^2)$ algorithm

- Initialize
  - One edge for the entire string $S[1..N]$
- For $i = 2$ to $N$
  - Add suffix $S[i..N]$ to suffix tree
    - Find match point for string $S[i..N]$ in current tree
    - If in “middle” of edge, create new node $w$
    - Add remainder of $S[i..N]$ as edge label to suffix $i$ leaf
- Running Time
  - $O(N-i)$ time to add suffix $S[i..N]$
From suffix trie to suffix tree for the string « BANANAS » using path compression.

In the worst case, a suffix tree can be built with a maximum of 2N nodes, where N is the length of the input text \( \mathcal{O}(N) \).
Building Suffix Trees in O(N) time

- Weiner had first linear time algorithm in 1973
- McCreight developed a more space efficient algorithm in 1976
- Ukkonen developed a simpler to understand variant in 1995
For a given string of text, T. Ukkonen's algorithm starts with an empty tree, then progressively adds each of the N prefixes of T to the suffix tree. For example, when creating the suffix tree for BANANAS, B is inserted into the tree, then BA, then BAN, and so on. When BANANAS is finally inserted, the tree is complete.

Building the Suffix Tree for BANANAS
Rather than storing strings, store a pair of indices (x,y) where x is beginning of string and y is the end.

**SUFFIX TREE**

![Suffix Tree Diagram]

Key: G O O G O L $  
1 2 3 4 5 6 7
Given strings `ababaabs` and `aabaat`:

List of UMs: `aab, abaa, baa`.

Unique matches: internal nodes with two leaf nodes from different genomes.

Bottom-up traversal of the tree

List of UMs: `aab, abaa, baa`.
1999 : MUMmer version 1
2001 : MUMmer version 2
2004 : MUMmer version 3 (open source)

MUMmer 1 use 38 bytes per base-pair. Version 2 uses only 20. Version 3 slightly less.

MUMmer 2 is three times faster than MUMmer 1 and use 60 % less memory.

MUMmer 3 is written in C++. Suffix tree library are in C (Prof. Dr Stefan Kurtz). MUMmer 3 size is about 600 KBytes in tar.gz format.

MUMmer 3.0 can find all 20-basepair or longer exact matches between a pair of 5-megabase genomes in 13.7 seconds, using 78 MB of memory, on a 2.4 GHz Linux desktop computer.
- Two books

- Prof. Dr Stefan Kurtz publications about suffix tree at uni-hamburg.de

- MUMmer home page: http://www.tigr.org/software/mummer/

- Google. Keywords: suffix tree
Thank you for your attention