

An Illustration of Genome Sequencing

Sequencing Genomes

1. **Sequencing short DNA fragments:** a sequencer generates an image called *chromatogram* (see p298) and then the *base-calling software* like PHRED or ABI-BASE CALLER determines the sequences automatically. Then length of a DNA fragment is usually between 500b and 1,000b. The error rate of a fragment is less than 2% and more errors occur at its ends. (Why?)
2. **Shotgun strategy** collects many short DNA fragments. The total length of short DNA fragments is multiple (6 to 10) of the length of the *target sequence*. The resulting set of frags is called *shotgun data*.
3. **A sequence assembly program** determines a longer sequence by assembling frags based on sequence overlaps.

Sequencing Genomes

continued

Sequencing genomes requires handling a large number of frags and repeats of various lengths. Thus the precoss is very complicated and many tools and strategies are used for genome sequencing.

- **Lab. Information Management System (LIMS)** is used for managing raw data and frags.
- **Mapping the genome** is necessary for large scale genome sequencing.
- **The hierarchical shotgun approach** has been used for many large scale genome sequencing. It generates an overlapping intermediate-sized clones like BAC, selects tiling path of these clones, and perform shotgun sequencing each of selected BACs.
- **The whole genome shotgun approach** is also used even for very large genomes, but needs mapping information.
- There are alternative approaches to genome sequencing.

Software for Genome Sequencing on Solar and IBM SP

We installed a phrap package on Solar and IBM SP (www.phrap.org). See also the manual on our course page.

- **phred** reads in a chromatogram determines base calls and output a numeric value for each base call, which denote the confidence of the base call. Thus there are two files, say **anyname-data** for shotgun data and **anyname-data.qual** for quality information.
- **phrap** reads in the two files and assembles contigs based on overlaps between contigs.
- **consed** reads in contig files and graphically show the assembly result (how fragments overlap). In addition, it allows users to perform re-assemble contigs with some controls.
- There are additional tools such as **RepeatMasker**, **phrapview**, **swat** and **crossmatch**.

NCBI Genome Resources

- Look at the *NCBI Genomic Sequence Assembly and Annotation* site.
- **Genome information** is available from the Entrez Genomes page. Each entry is linked to a taxonomy browser. A genome map is available through **See the Genome** link at the organism's homepage.
- **Map viewer** shows protein coding regions, listings for coding regions for protein and RNA, and other information. Four levels of information are available: the organism's home page, the graphical view, the detailed map for each chromosome, and the sequence view.
- **ORF finder** is a tool for locating open reading frames in a DNA sequence. Sequence pieces longer than a threshold between stop codons are candidates for ORFs.
- **LocusLink** is a curated database of genetic in several eukaryotic organisms that give rise to known phenotypes. A list of known traits is available.
- **Homologene** is a database of pairwise orthologs either from literature or genome-context similarity relationship.

TIGR Genome Resources and EnsEMBL

- TIGR maintains the comprehensive microbial resources that can be searched, *omniome*.
- TIGR also maintains genome specific databases including ESTs.
- EnsEMBL maintains the data for the human genome. There are three ways to search EnsEMBL: by BLAST search, by ID, and by a chromosome map viewer.
- There are many organism specific databases!
Try the link on our course home page and more on the web.

Annotating and Analyzing Whole Genome Sequences

In addition to the determination of genome sequence, there are many tasks for a genome project including ORF finding, gene finding, the prediction of proteins and their functions.

Managing all tasks is quite complex, but unfortunately there is little research efforts on the architectural development of such system except MAGPIE and DAS.

Comparing Genomic DNA

Comparing very long DNA sequences or a large number of sequences is a very challenging task and there are limited resources for this purpose.

- **CROSS_MATCH** is a part of **PHRAP** package.
`www.phrap.org` (intalled on our systems).
- **PIPMAKER** relies on **BLASTZ**.
`bio.cse.psu.edu/pipmaker` (intalled on our systems, but memory problem for larger seqs).
- **MUMMER** uses Suffix tree to find local matches and local matches are combined using an algorithm for longest increasing sequences.
It is available from TIGR (it will be installed).
- **WABA** at UCSC (intalled on our systems).

Functional Genomics Using Sequences

- ESTs are partial sequences from cDNA which are DNA strands built from mRNA. We can measure rough expression levels under particular conditions or in particular tissues.
- GenBank provides a database of ESTs as well as assembled ESTs called *UniGene*.
- *Serial Analysis of Gene Expressions (SAGE)* is a technique to measure transcription levels of genes using short DNA tags.
Look at NCBI Web site.

Functional Genomics Using DNA Microarray

- Immobilized oligomers in a DNA microarray serve as probes to detect unique, complementary DNA or RNA molecules. Using fluorescent probes are used to measure expression levels.
- The microarray technology can be used to measure gene expressions on a genome scale.
- `www.ncbi.nlm.nih.gov/geo`
`genome-www5.stanford.edu/MicroArray/SMD`
- Bioinformatics is important for:
 - the array experiment design (selection of probes and primer design),
 - visualization of the experimental result (see p312), and
 - the analysis of result such as clustering and classification.

Proteomics

As not all mRNA are translated into proteins, measuring protein concentration and activity in the cell is important but still difficult.

- The 2D gel electrophoresis is a tool for proteomics. One dimension is PH gradient (isoelectric focusing) and the other is molecular weight.
- The analysis is image (see page 318) is similar to the analysis of microarray, but connecting images to sequences is not trivial.
- Look at the Expert Protein Analysis System (ExPASy, www.expasy.ch/tools).
- A comprehensive analysis is not easy. One way to combine the results from Microarray and 2D-PAGE analysis with sequence information, and with further sequence analysis.

Biochemical Pathways

- An enzyme is a protein that catalyzes chemical reaction on the input material (substrate). The product of a chemical reaction can be input to another chemical reaction. The representation of this chain of reactions is a metabolic pathways (see p322).
- The enzyme reactions are hierarchically classified and named using the *Enzyme Commission (EC) Numbers*. For example, EC number 1.1.1.1 refers to a alcohol dehydrogenase. Web site?
- There are databases of metabolic pathways:
WIT (What Is There?, wit.mcs.anl.gov/WIT2), KEGG (Kyoto Encyclopedia of Genes and Genomes, www.genome.ad.jp/kegg), Ecocyc (www.ecocyc.org), PathDB.

Modeling Kinetics and Physiology

This type of research is to get the big picture of cellular function and physiology, One way is to model biochemical kinetics and physiology in terms of chemical concentration and associated rate equation. Look at

- Gepasi www.gepasi.org
- XPP www.math.pitt.edu/~bard/xpp/xpp.html
- the virtual cell portal at www.nrcam.uchc.edu
- Peter Ortoleva at IU