CAP 5510: Introduction to Bioinformatics

Giri Narasimhan

ECS 254; Phone: x3748
giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS08.html
Microarray/DNA chip technology

- High-throughput method to study gene expression of thousands of genes simultaneously.
- Many applications:
  - Genetic disorders & Mutation/polymorphism detection
  - Study of disease subtypes
  - Drug discovery & toxicology studies
  - Pathogen analysis
  - Differing expressions over time, between tissues, between drugs, across disease states
Polymerase Chain Reaction (PCR)

- For testing, large amount of DNA is needed
  - Identifying individuals for forensic purposes
    - (0.1 microliter of saliva contains enough epithelial cells)
  - Identifying pathogens (viruses and/or bacteria)

- PCR is a technique to amplify the number of copies of a specific region of DNA.

- Useful when exact DNA sequence is unknown
- Need to know “flanking” sequences
- Primers designed from “flanking” sequences
PCR

Flanking Regions with known sequence

Region to be amplified

DNA

Forward Primer

Reverse Primer

Millions of Copies

3/25/08  CAP5510
PCR

Polymerase Chain Reaction (PCR)

DNA

Primer

Heat

Etc.
Schematic outline of a typical PCR cycle

1. **Step 1: Denaturation**
   - Target DNA

2. **Step 2: Primer Annealing**
   - Primers

3. **Step 3: Primer Extension**
   - dNTPs
   - DNA polymerase
POLYMERASE CHAIN REACTION

1. DNA is denatured. Primers attach to each strand. A new DNA strand is synthesized behind primers on each template strand.

2. Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

3. Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

4. Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

5. Continued rounds of amplification swiftly produce large numbers of identical fragments. Each fragment contains the DNA region of interest.
Gel Electrophoresis

- Used to measure the lengths of DNA fragments.
- When voltage is applied to DNA, different size fragments migrate to different distances (smaller ones travel farther).
Gel Pictures
Gel Electrophoresis: Measure sizes of fragments

- The phosphate backbone makes DNA a highly negatively charged molecule.
- DNA can be separated according to its size.
- **Gel:** allow hot 1% solution of purified agarose to cool and solidify/polymerize.
- DNA sample added to wells at the top of a gel and voltage is applied. Larger fragments migrate through the pores slower.
- Varying concentration of agarose makes different pore sizes & results.
- Proteins can be separated in much the same way, only acrylamide is used as the crosslinking agent.
Gel Electrophoresis
Gel Electrophoresis
Sequencing
Why sequencing?

- Useful for further study:
  - Locate gene sequences, regulatory elements
  - Compare sequences to find similarities
  - Identify mutations
  - Use it as a basis for further experiments

Next 4 slides contains material prepared by Dr. Stan Metzenberg. Also see: http://stat-www.berkeley.edu/users/terry/Classes/s260.1998/Week8b/week8b/node9.html
Two methods independently developed in 1974
- Maxam & Gilbert method
- Sanger method: became the standard

Nobel Prize in 1980
Original Sanger Method

- (Labeled) Primer is annealed to template strand of denatured DNA. This primer is specifically constructed so that its 3' end is located next to the DNA sequence of interest. Once the primer is attached to the DNA, the solution is divided into four tubes labeled "G", "A", "T" and "C". Then reagents are added to these samples as follows:
  - “G” tube: ddGTP, DNA polymerase, and all 4 dNTPs
  - “A” tube: ddATP, DNA polymerase, and all 4 dNTPs
  - “T” tube: ddTTP, DNA polymerase, and all 4 dNTPs
  - “C” tube: ddCTP, DNA polymerase, and all 4 dNTPs

- DNA is synthesized, & nucleotides are added to growing chain by the DNA polymerase. Occasionally, a ddNTP is incorporated in place of a dNTP, and the chain is terminated. Then run a gel.

- All sequences in a tube have same prefix and same last nucleotide.
Reactions performed in a single tube containing all four ddNTP's, each labeled with a different color dye.
Both Sanger Methods

Example of sequences seen in gel:

5'-GAATGTCTTTTCTCTAAGTCCTAAG
3'-GGAGACCTTCAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCTTTTCTCTAAGTCCTAAG
3'-GGAGACCTTCAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCTTTTCTCTAAGTCCTAAG
3'-GGAGACCTTCAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCTTTTCTCTAAGTCCTAAG
3'-GGAGACCTTCAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCTTTTCTCTAAGTCCTAAG
3'-GGAGACCTTCAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
Sequencing Gels: Separate vs Single Lanes

GCCAGGTGAGCCTTTTGCA

A C G T

A C G T
FIGURE 13.3 A sample chromatogram, as viewed with the vtrace program (Ewing, 2002). Signal intensities corresponding to fragments ending with A (green), C (blue), G (black), and T (red) are shown out to approximately 722 bases.
Shotgun Sequencing

From http://www.tulane.edu/~biochem/lecture/723/humgen.html
Sequencing

FIGURE 13.1 Shotgun cloning. Genomic DNA sequencing begins with isolated genomic DNA in green at the top of the figure. In the hierarchical clone-based shotgun approach on the left, DNA is sheared and the size is selected for large fragments on the order of 200 Kb, then ligated to a suitable vector, such as a BAC vector shown in blue. Individually isolated clones in turn are sheared independently, generating fragments of approximately 4 Kb, which are then ligated to a small-scale vector, typically a plasmid (red bar) suitable for sequencing reactions. The whole genome shotgun approach bypasses the intermediate large-insert clone and generates large numbers of small fragments, typically 4 Kb and 10 Kb.
Sequencing: Generate Contigs

- Short for “contiguous sequence”. A continuously covered region in the assembly.

A supercontig is formed when an association can be made between two contigs that have no sequence overlap.

This commonly occurs using information obtained from paired plasmid ends. For example, if both ends of a BAC clone are sequenced, then it can be inferred that these two sequences are approximately 150-200 Kbp apart (based on the average size of a BAC). If the sequence from one end is found in a particular sequence contig, and the sequence from the other end is found in a different sequence contig, the two sequence contigs are said to be linked. In general, it is useful to have end sequences from more than one clone to provide evidence for linkage.

[NCBI Genome Glossary]
Shotgun Sequencing

From http://www.tulane.edu/~biochem/lecture/723/humgen.html
Human Genome Project

Play the Sequencing Video:
• Download Windows file from http://www.cs.fiu.edu/~giri/teach/6936/Papers/Sequence.exe
• Then run it on your PC.
Assembly: Simple Example

- ACCGT, CGTGC, TTAC, TACCGT
- Total length = ~10

- ----ACCGT----
- -----CGTGC
- TTAC-----
- TACCGT--
- TTACCGTGC
Assembly: Complications

- Errors in input sequence fragments (~3%)
  - Indels or substitutions
- Contamination by host DNA
- Chimeric fragments (joining of non-contiguous fragments)
- Unknown orientation
- Repeats (long repeats)
  - Fragment contained in a repeat
  - Repeat copies not exact copies
  - Inherently ambiguous assemblies possible
  - Inverted repeats
- Inadequate Coverage
Assembly: Complications

\[ w = \text{AGTATTGGCAATC} \]
\[ z = \text{AATCGATG} \]
\[ u = \text{ATGCAAACCT} \]
\[ x = \text{CCTTTTGG} \]
\[ y = \text{TTGGAATCACT} \]

\begin{align*}
\text{AGTATTGGCAATC} & \quad \text{---AATCGATG---} \\
\text{-------------} & \quad \text{---ATGCAAACCT---} \\
\text{-----TTGGAATCACT-----} & \quad \text{---------CCTTTTGG} \\
\text{AGTATTGGCAATCACTAATCGATGCAAACCTTTTGG} & \\
\end{align*}

**FIGURE 4.20**

A bad solution for an assembly problem, with a multiple alignment whose consensus is a shortest common superstring. This solution has length 36 and is generated by the Greedy algorithm. However, its weakest link is zero.

\begin{align*}
\text{AGTATTGGCAATC} & \quad \text{---------CCTTTTGG} \\
\text{-------------} & \quad \text{---AATCGATG---} \\
\text{-------------} & \quad \text{---TTGGAATCACT} \\
\text{-------------} & \quad \text{---ATGCAAACCT---} \\
\text{AGTATTGGCAATCGATGCAAACCTTTTGGCAATCACT} & \\
\end{align*}

**FIGURE 4.21**

Solution according to the unique Hamiltonian path. This solution has length 37, but exhibits better linkage. Its weakest link is 3.
Assembly: Complications

**FIGURE 4.8**
Target sequence leading to ambiguous assembly because of repeats of the form XXX.

**FIGURE 4.9**
Target sequence leading to ambiguous assembly because of repeats of the form XYXY.

**FIGURE 4.10**
Target sequence with inverted repeat. The region marked $\overline{X}$ is the reverse complement of the region marked $X$. 
Other sequencing methods

- Sanger Method (70Kbp/run)
- Sequencing by Hybridization (SBH)
- Dual end sequencing
- Chromosome Walking (see page 5-6 of Pevzner’s text)
- 454 Sequencing (60Mbp/run)
- Solexa Sequencing (600Mbp/run) [Illumina]
454 Sequencing: New Sequencing Technology

- 454 Life Sciences, Roche
- Fast (20 million bases per 4.5 hour run)
- Low cost (lower than Sanger sequencing)
- Simple (entire bacterial genome in days with one person -- without cloning and colony picking)
- Convenient (complete solution from sample prep to assembly)
- PicoTiterPlate Device
  - Fiber optic plate to transmit the signal from the sequencing reaction
- Process:
  - Library preparation: Generate library for hundreds of sequencing runs
  - Amplify: PCR single DNA fragment immobilized on bead
  - Sequencing: “Sequential” nucleotide incorporation converted to chemilluminoscent signal to be detected by CCD camera.
(a) Fragment, (b) add adaptors, (c) “1 fragment, 1 bead”, (d) emPCR on bead, (e) put beads in PicoTiterPlate and start sequencing: “1 bead, 1 read”, and (f) analyze
FIGURE 8

DNA Library Preparation

- Anneal sstDNA to an excess of DNA Capture Beads

Emulsify beads and PCR reagents in water-in-oil microreactors

Clonal amplification occurs inside microreactors

Break microreactors enrich for DNA-positive beads

emPCR

Sequencing

4.5 HOURS

8 HOURS

7.5 HOURS

gDNA → sstDNA Library
Sequencing

FIGURE 9

DNA Library Preparation  emPCR  Sequencing

4.5 HOURS  8 HOURS  7.5 HOURS

- Well diameter: average of 44μm
- 400,000 reads obtained in parallel
- A single cloned amplified ssDNA bead is deposited per well

Amplified ssDNA library beads  →  Quality filtered bases
Sequencing

**FIGURE 10**

- **DNA Library Preparation**: 4.5 hours
- **emPCR**: 8 hours
- **Sequencing**: 7.5 hours

- 4 bases (TACG) cycled 100 times
- Chemiluminescent signal generation
- Signal processing to determine base sequence and quality score

**DNA Capture Bead**
- Containing millions of copies of a single clonal fragment

**Amplified sstDNA library beads** → **Quality filtered bases**
Solexa Sequencing

1. PREPARE GENOMIC DNA SAMPLE
   - Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE
   - Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION
   - Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Solexa Sequencing

4. Fragments become double-stranded
   - Attached terminus
   - Free terminus

5. Denature the double-stranded molecules
   - Attached

6. Complete amplification
   - Clusters

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
Denaturation leaves single-stranded templates anchored to the substrate.
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.
Solexa Sequencing

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
Solexa Sequencing

10. Image Second Chemistry Cycle
   After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. Sequence Reads Over Multiple Chemistry Cycles
   Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

12. Align Data
   Align data, compare to a reference, and identify sequence differences.
Assembly Software

- Parallel EST alignment engine (http://corba.ebi.ac.uk/EST") with a CORBA interface to alignment database. Can perform ad hoc assemblies. Can act as foundation for CORBA-based EST assembly and editing package. [Parsons, EBI]

- Software using multiple alternative sequence assembly "engines" writing to a common format file [Staden, Cambridge] (http://www.mrc-lmb.cam.ac.uk/pubseq/index.html).

- Phrap (http://bozeman.genome.washington.edu/phrap.docs/phrap.html)

- Assembler (TIGR) for EST and Microbial whole-genome assembly (http://www.tigr.org/softlab/)

- FAK2 and FAKtory (http://www.cs.arizona.edu/people/gene/) [Myers]

- GCG (http://www.gcg.com)

- Falcon [Gryan, Harvard] fast (rascal.med.harvard.edu/gryan/falcon/)


- CAP 2 [Huang] (http://www.tigem.it/ASSEMBLY/capdoc.html)
2D-Gels

(a) Separation in first dimension (by charge)

Protein mixture

Isoelectric focusing (IEF)

Apply first gel to top of second

pH 4.0  pH 10.0

Separation in second dimension (by size)

pH 4.0  pH 10.0

SDS electrophoresis
2D Gel Electrophoresis
First Dimension Methodology of a 2D Gel:
Denatured cell extract layered on a glass tube filled with polyacrylamide saturated with solution of ampholytes, a mixture of polyanionic[(-) charged] and polycationic [(+) charged] molecules. When placed in an electric field, the ampholytes separate and form continuous gradient based on net charge. Highly polyanionic ampholytes will collect at one end of tube, highly polycationic ampholytes will collect at other end. Gradient of ampholytes establishes pH gradient. Charged proteins migrate through gradient until they reach their pI, or isoelectric point, the pH at which the net charge of the protein is zero. This resolves proteins that differ by only one charge.

Entering the Second Dimension:
Proteins that were separated on IEF gel are next separated in the second dimension based on their molecular weights. The IEF gel is extruded from tube and placed lengthwise in alignment with second polyacrylamide gel slab saturated with SDS. When an electric field is imposed, the proteins migrate from IEF gel into SDS slab gel and then separate according to mass. Sequential resolution of proteins by their charge and mass can give excellent separation of cellular proteins. As many as 1000 proteins can be resolved simultaneously.

*Some information was taken from Lodish et al. Molecular Cell Biology.
2D-gels

Comparing Proteomes For Differences in Protein Expression

Comparing Different Sample Types For Changes in Protein Levels
Mass Spectrometry

- **Mass measurements By Time-of-Flight**
  Pulses of light from laser ionizes protein that is absorbed on metal target. Electric field accelerates molecules in sample towards detector. The time to the detector is inversely proportional to the mass of the molecule. Simple conversion to mass gives the molecular weights of proteins and peptides.

- **Using Peptide Masses to Identify Proteins:**
  One powerful use of mass spectrometers is to identify a protein from its peptide mass fingerprint. A peptide mass fingerprint is a compilation of the molecular weights of peptides generated by a specific protease. The molecular weights of the parent protein prior to protease treatment and the subsequent proteolytic fragments are used to search genome databases for any similarly sized protein with identical or similar peptide mass maps. The increasing availability of genome sequences combined with this approach has almost eliminated the need to chemically sequence a protein to determine its amino acid sequence.
Genomics

- Study of all genes in a genome, or comparison of whole genomes.
  - Whole genome sequencing
  - Whole genome annotation & Functional genomics
  - Whole genome comparison
    - PipMaker: uses BLASTZ to compare very long sequences (> 2Mb);
      http://www.cse.psu.edu/pipmaker/
    - Mummer: used for comparing long microbial sequences (uses Suffix trees!)
Genomics

- Study of all genes in a genome
  - Gene Expression
    - Microarray experiments & analysis
      - Probe design (*CODEHOP*)
      - Array image analysis (*CrazyQuant*)
      - Identifying genes with significant changes (*SAM*)
      - Clustering
Comparative Genomics

- Comparison of whole genomes.
  - Whole genome sequencing
  - Whole genome annotation & Functional genomics
  - Whole genome comparison
    - PipMaker, MultiPipMaker, EnteriX: PipMaker uses BLASTZ to compare very long sequences (> 2Mb); http://www.cse.psu.edu/pipmaker/
    - Mummer: used for comparing long microbial sequences (uses Suffix trees!)
    - Many more!
PEDANT useful resource for standard questions in comparative genomics. For e.g., how many known proteins in XXX have known 3-d structures, how many proteins from family YYY are in ZZZ, etc.

COGs Clusters of orthologous groups of proteins.

MBGD Microbial genome database searches for homologs in all microbial genomes.
Study of all **proteins** in a genome, or comparison of whole genomes.

- Whole genome annotation & Functional proteomics
- Whole genome comparison
- Protein Expression: **2D Gel Electrophoresis**
Other Proteomics Tools

From ExPASy/SWISS-PROT:

- **AACompIdent** identify proteins from aa composition
  
  **Input**: aa composition, isoelectric point, mol wt., etc.  
  **Output**: proteins from DB

- **AACompSim** compares proteins aa composition with other proteins

- **MultIdent** uses mol wt., mass fingerprints, etc. to identify proteins

- **PeptIdent** compares experimentally determined mass fingerprints with theoretically determined ones for all proteins

- **FindMod** predicts post-translational modifications based on mass difference between experimental and theoretical mass fingerprints.

- **PeptideMass** theoretical mass fingerprint for a given protein.

- **GlycoMod** predicts oligosaccharide modifications from mass difference

- **TGREASE** calculates hydrophobicity of protein along its length
Gene Networks & Pathways

- Genes & Proteins act in concert and therefore form a complex network of dependencies.
Pathway Example from KEGG

**beta-LACTAM RESISTANCE**

- Penicillins and Cephalosporins Biosynthesis
- **Staphylococcus aureus**

Diagram showing the pathway with key components:
- BlaR1
- BlaI
- BlaZ
- MecR1
- MecI
- MecA

Related reactions and interactions indicated by arrows.
STSSs and ESTs

- **Sequence-Tagged Site**: short, unique sequence
- **Expressed Sequence Tag**: short, unique sequence from a coding region
  - 1991: 609 ESTs [Adams et al.]
  - June 2000: 4.6 million in dbEST
  - *Genome sequencing center at St. Louis produce 20,000 ESTs per week.*
What Are ESTs and How Are They Made?

- Small pieces of DNA sequence (usually 200 - 500 nucleotides) of low quality.
- Extract mRNA from cells, tissues, or organs and sequence either end. Reverse transcribe to get cDNA (5’ EST and 3’EST) and deposit in EST library.
- Used as "tags" or markers for that gene.
- Can be used to identify similar genes from other organisms (Complications: variations among organisms, variations in genome size, presence or absence of introns).
- 5’ ESTs tend to be more useful (cross-species conservation), 3’ EST often in UTR.
DNA Markers

- Uniquely identifiable DNA segments.
- Short, <500 nucleotides.
- Layout of these markers give a map of genome.
- Markers may be polymorphic (variations among individuals). Polymorphism gives rise to alleles.
- Found by PCR assays.
Polymorphisms

- **Length polymorphisms**
  - Variable # of tandem repeats (VNTR)
  - Microsatellites or short tandem repeats
  - Restriction fragment length polymorphism (RFLP) caused by changes in restriction sites.

- **Single nucleotide polymorphism (SNP)**
  - Average once every ~100 bases in humans
  - Usually biallelic
  - dbSNP database of SNPs (over 100,000 SNPs)
  - ESTs are a good source of SNPs
SNPs

- SNPs often act as “disease markers”, and provide “genetic predisposition”.
- SNPs may explain differences in drug response of individuals.
- **Association study**: study SNP patterns in diseased individuals and compare against SNP patterns in normal individuals.
- Many diseases associated with SNP profile.