Introduction to genome biology and DNA microarray experiments

Sandrine Dudoit and Robert Gentleman

Statistics and Genomics - Lecture 1, Part I
Department of Biostatistics
Harvard School of Public Health
January 23-25, 2002
Outline of lecture 1

Part I:
• Introduction to genome biology;
• Introduction to microarray experiments.

Part II:
• Image analysis (cDNA microarrays);
• Normalization (cDNA microarrays);
• Experimental design.
Introduction to genome biology
The human genome

- The **cell** is the fundamental working unit of every living organism.
- Humans: trillions of cells (metazoa); other organisms like yeast: one cell (protozoa).
- Cells are of many different types (e.g. blood, skin, nerve cells), but all can be traced back to a single cell, the fertilized egg.
The human genome

• The genome, or blueprint for all cellular structures and activities in our body, is encoded in DNA molecules.

• Each cell contains a complete copy of the organism's genome.
The human genome

- The human genome is distributed along 23 pairs of chromosomes:
  - 22 autosomal pairs;
  - the sex chromosome pair, XX for females and XY for males.
- In each pair, one chromosome is paternally inherited, the other maternally inherited (cf. meiosis).
The human genome

• Chromosomes are made of compressed and entwined DNA.

• A (protein-coding) gene is a segment of chromosomal DNA that directs the synthesis of a protein.
The eukaryotic cell
Chromosomes

HUMAN CHROMOSOMES

1  2  3  4  5
X  X  X  X  X

6  7  8  9  10  11  12
X  X  X  X  X  X  X

13  14  15  16  17  18
X  X  X  X  X  X

19  20  21  22  23
X  X  X  X  X

Chromatid
Telomere
Centromere
Chromosomes and DNA
Cell divisions

- **Mitosis.** One nuclear division produces two daughter **diploid** nuclei identical to the parent nucleus.

- **Meiosis.** Two successive nuclear divisions produces four daughter **haploid** nuclei, different from original cell.

Leads to the formation of gametes (egg/sperm).
Mitosis

- **Prophase**: Chromatin condenses into chromosomes. Nuclear envelope disappears.

- **Metaphase**: Chromosomes align at the equatorial plate.

- **Anaphase**: Sister chromatids separate. Centromeres divide.

- **Telophase**: Chromatin expands. Cytoplasm divides. Two daughter cells
Recombination

Gametes

Crossing-over and recombination during meiosis
DNA

• A deoxyribonucleic acid or DNA molecule is a double-stranded polymer composed of four basic molecular units called nucleotides.
• Each nucleotide comprises a phosphate group, a deoxyribose sugar, and one of four nitrogen bases: adenine (A), guanine (G), cytosine (C), and thymine (T).
• The two chains are held together by hydrogen bonds between nitrogen bases.
• Base-pairing occurs according to the following rule: G pairs with C, and A pairs with T.
DNA

The diagram illustrates the structure of DNA, showing the sugar-phosphate backbone, base pairs, and hydrogen bonds between complementary bases. The diagram includes nucleotides and the double helix configuration of DNA.
Genetic and physical maps

Genetic map

Physical map

Chromosome

DNA sequence

Sequences of base pairs mapping
Genetic and physical maps

- **Physical distance**: number of base pairs (bp).
- **Genetic distance**: expected number of crossovers between two loci, per chromatid, per meiosis. Measured in Morgans (M) or centiMorgans (cM).
- 1cM ~ 1 million bp (1Mb).
The human genome in numbers

• 23 pairs of chromosomes;
• 2 meters of DNA;
• 3,000,000,000 bp;
• 35 M (males 27M, females 44M);
• 30,000-40,000 genes.
Proteins

• **Proteins**: large molecules composed of one or more chains of amino acids.

• **Amino acids**: class of 20 different organic compounds containing a basic amino group (\(-\text{NH}_2\)) and an acidic carboxyl group (\(-\text{COOH}\)).

• The order of the amino acids is determined by the base sequence of nucleotides in the gene coding for the protein.

• E.g. hormones, enzymes, antibodies.
Amino acids

**Families of Amino Acids**

The common amino acids are grouped according to whether their side chains are acidic, basic, uncharged-polar, or nonpolar.

These 20 amino acids are given both in textbook and online examinations.

**Basic Side Chains**

- **Lysine**
  - $\text{Arg}$ or $\text{Rb}$
  - The group is very basic because its positive charge is stabilized by resonance.

**Lysine**

**Acidic Side Chains**

- **Aspartic Acid**
  - $\text{Asp}$ or $\text{D}$
  - Although the amino group is not charged at neutral pH, it is polar.

**The Amino Acid**

The general formula of a amino acid is:

$\text{NH}_2\text{R}+\text{COOH}$

- $\text{R}$ is a nonpolar amine group of amino acids.
- It forms one of 22 different side chains.
- At pH 7 both the amino and carboxyl groups are ionized.

**Optical Isomers**

- The molecule (R) is asymmetric, which allows for two mirror images (or stereo) forms, known as:
  - **L**
  - **D**

**Protein Consistency Stability of Amino Acids**

**Peptide Bonds**

Amino acids are covalently linked together by an amide linkage, called a peptide bond.

- **Peptide Bond**
  - The four atoms of the amide group form a sp2 hybridized carbon.
  - There is no rotation around the C-N bond.

**Proteins**

- Proteins are long polymers of amino acids linked by peptide bonds, and they are always written with the N terminus on the left.
- The sequence of this polypeptide is histidine-cysteine-alanine.

**Peptide bonds**

These two single bonds allow rotation, so long chains of amino acids are very flexible.

**Uncharged Polar Side Chains**

- **Serine**
  - $\text{Ser}$ or $\text{S}$
  - The $\text{CH}_2$ group is polar.

**Phospho bonds**

- Phospho bonds can form between two cysteine side chains in proteins.
Proteins

Primary protein structure is sequence of a chain of amino acids

Amino Acids

Amino group: $\text{NH}_2$

Acidic carboxyl group: $\text{COOH}$

R group
Proteins

**Primary protein structure**
- is sequence of a chain of amino acids
- Amino Acids

**Secondary protein structure**
- occurs when the sequence of amino acids are linked by hydrogen bonds
- Pleated sheet
- Alpha helix

**Tertiary protein structure**
- occurs when certain attractions are present between alpha helices and pleated sheets
- Pleated sheet
- Alpha helix

**Quaternary protein structure**
- is a protein consisting of more than one amino acid chain
Cell types
Differential expression

• Each cell contains a complete copy of the organism's genome.
• Cells are of many different types and states
  E.g. blood, nerve, and skin cells, dividing cells, cancerous cells, etc.
• What makes the cells different?
• **Differential gene expression**, i.e., **when, where**, and in **what quantity** each gene is expressed.
• On average, 40% of our genes are expressed at any given time.
Central dogma

The expression of the genetic information stored in the DNA molecule occurs in two stages:

– (i) transcription, during which DNA is transcribed into mRNA;

– (ii) translation, during which mRNA is translated to produce a protein.

DNA $\rightarrow$ mRNA $\rightarrow$ protein

Other important aspects of regulation: methylation, alternative splicing, etc.
Central dogma

Replication
DNA duplicates

Transcription
RNA synthesis

Translation
Protein synthesis

The Central Dogma of Molecular Biology
RNA

• A ribonucleic acid or RNA molecule is a nucleic acid similar to DNA, but
  – single-stranded;
  – ribose sugar rather than deoxyribose sugar;
  – uracil (U) replaces thymine (T) as one of the bases.
• RNA plays an important role in protein synthesis and other chemical activities of the cell.
• Several classes of RNA molecules, including messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and other small RNAs.
The genetic code

- **DNA**: sequence of *four* different nucleotides.
- **Proteins**: sequence of *twenty* different amino acids.
- The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the **genetic code**, which relates nucleotide triplets or **codons** to **amino acids**.
The genetic code

Mapping between codons and amino acids is many-to-one: 64 codons but only 20 a.a..

Third base in codon is often redundant, e.g., stop codons.
Exons and introns

- Genes comprise only about 2% of the human genome; the rest consists of non-coding regions, whose functions may include providing chromosomal structural integrity and regulating when, where, and in what quantity proteins are made (regulatory regions).
- The terms **exon** and **intron** refer to coding (translated into a protein) and non-coding DNA, respectively.
Exons and introns
Splicing
Alternative splicing

- There are more than 1,000,000 different human antibodies. How is this possible with only ~30,000 genes?
- **Alternative splicing** refers to the different ways of combining a gene’s exons. This can produce different forms of a protein for the same gene,
- Alternative pre-mRNA splicing is an important mechanism for regulating gene expression in higher eukaryotes.
- E.g. in humans, it is estimated that approximately 30% genes are subject to alternative splicing.
Alternative splicing
Immunoglobulin

• B cells produce antibody molecules called immunoglobulins (Ig) which fall in five broad classes.

• Diversity of Ig molecules
  – DNA sequence: recombination, mutation.
  – mRNA sequence: alternative splicing.
  – Protein structure: post-translational proteolysis, glycosylation.
Functional genomics

• The various genome projects have yielded the complete DNA sequences of many organisms.
  E.g. human, mouse, yeast, fruitfly, etc.
  Human: 3 billion base-pairs, 30-40 thousand genes.
• Challenge: go from sequence to function, i.e., define the role of each gene and understand how the genome functions as a whole.
Pathways

• The complete genome sequence doesn’t tell us much about how the organism functions as a biological system.

• We need to study how different gene products function to produce various components.

• Most important activities are not the result of a single molecule but depend on the coordinated effects of multiple molecules.
**TFG-β pathway**

- TGF-β (transforming growth factor beta) plays an essential role in the control of development and morphogenesis in multicellular organisms.
- This is done through SMADS, a family of signal transducers and transcriptional activators.
Pathways

- [http://www.grt.kyushu-u.ac.jp/spad/](http://www.grt.kyushu-u.ac.jp/spad/)
- There are many open questions regarding the relationship between expression level and pathways.
- It is not clear whether expression level data will be informative.
DNA microarrays
DNA microarrays

**DNA microarrays** rely on the *hybridization* properties of nucleic acids to monitor DNA or RNA abundance on a genomic scale in different types of cells.

The ancestor of microarrays: the Northern blot.
Nucleic acid hybridization

(a) DNA is denatured by heating
(b) Renaturation on cooling
(c) Hybridization

Nucleic Acid Hybridization
Gene expression assays

The main types of gene expression assays:

- Serial analysis of gene expression (SAGE);
- Short oligonucleotide arrays (Affymetrix);
- Long oligonucleotide arrays (Agilent Inkjet);
- Fibre optic arrays (Illumina);
- cDNA arrays (Brown/Botstein).
Applications of microarrays

- Measuring transcript abundance (cDNA arrays);
- Genotyping;
- Estimating DNA copy number (CGH);
- Determining identity by descent (GMS);
- Measuring mRNA decay rates;
- Identifying protein binding sites;
- Determining sub-cellular localization of gene products;
- …
Applications of microarrays

• **Cancer research:** Molecular characterization of tumors on a genomic scale
  → more reliable diagnosis and effective treatment of cancer.

• **Immunology:** Study of host genomic responses to bacterial infections; reversing immunity.

• ...
cDNA microarray experiment

Prepare cDNA target

"Normal" Tumor

RT / PCR Label with Fluorescent Dyes

Combine Equal Amounts

Hybridize target to microarray

Microarray Technology

Prepare Microarray

SCAN
Mixture of neuron cells (Control) Label with Green Fluorescent Dye

Certain type of Neuron cell Label with red Fluorescent Dye

cDNA

Denature

Denature

Combine equal amounts

Hybridization

Denature cDNA Microarray

RIKEN Mouse EST (sequenced gene)
The process

**Building the chip:**
- MASSIVE PCR
- PCR PURIFICATION AND PREPARATION
- PRINTING

**RNA preparation:**
- CELL CULTURE AND HARVEST
- RNA ISOLATION
- cDNA PRODUCTION

**Hybing the chip:**
- ARRAY HYBRIDIZATION
- PROBE LABELING
- DATA ANALYSIS

**Post-processing:**
The arrayer

Ngai Lab arrayer, UC Berkeley

Print-tip head
Print-tips collect cDNA from wells

384 well plate
Contains cDNA probes

Glass Slide
Array of bound cDNA probes
4x4 blocks = 16 print-tip groups

Print-tip group 7

Print-tips collect cDNA from wells

384 well plate
Contains cDNA probes

Glass Slide
Array of bound cDNA probes
4x4 blocks = 16 print-tip groups

Print-tip group 7

Print-tip group 1

cDNA clones
Spotted in duplicate

4x4 blocks = 16 print-tip groups

Print-tip group 7

Print-tip group 1

cDNA clones
Spotted in duplicate

Print-tip group 1

cDNA clones
Spotted in duplicate
Sample preparation
Hybridization

Hybridize for 5-12 hours

Binding of cDNA target samples to cDNA probes on the slide
Hybridization chamber

- Humidity
- Temperature
- Formamide
  (Lowers the Temp)
Scanning

Detector

PMT

Image

Duplicate spots

Cy5: 635nm
Cy3: 532nm
RGB overlay of Cy3 and Cy5 images
Raw data

• Human cDNA arrays
  – ~43K spots;
  – 16–bit TIFFs: ~ 20Mb per channel;
  – ~ 2,000 x 5,500 pixels per image;
  – Spot separation: ~ 136um;
  – For a “typical” array:
    Mean = 43, med = 32, SD = 26 pixels per spots
WWW resources

• Complete guide to “microarraying”
  http://cmgm.stanford.edu/pbrown/mguide/
  http://www.microarrays.org
  – Parts and assembly instructions for printer and scanner;
  – Protocols for sample prep;
  – Software;
  – Forum, etc.

• Animation:
  http://www.bio.davidson.edu/courses/genomics/chip/chip.html
Integration of biological data

• Expression, sequence, structure, annotation.
• Integration will depend on our using a common language and will rely on database methodology as well as statistical analyses.
• This area is largely unexplored.
Testing

Biological verification and interpretation

Microarray experiment

Image analysis

Normalization

Experimental design

Biological question

Statistics and Microarrays

Estimation

Testing

Clustering

Discrimination

Biological verification and interpretation
Statistical computing

Everywhere …

- for statistical design and analysis:
  pre-processing, estimation, pattern discovery and recognition, etc.

- for integration with biological information resources
  (in-house and external databases).
Road map

• Lecture 1, Part II: cDNA arrays

  – Pre-processing: Image analysis;

  – Pre-processing: Normalization;

  – Experimental design.
Road map

• Lecture 2: Differential expression.

• Lecture 3: Applications of HMMs to sequence analysis.

• Lecture 4: Affymetrix chips.

• Lecture 5: Classification.