Primer on Genetics and Molecular Biology

Short course: Practical Analysis of DNA Microarray Data
Instructors: Vince Carey & Sandrine Dudoit

KolleKolle, Denmark
October 26-28, 2003
Outline

- Cells, chromosomes, and cell division
- DNA structure and replication
- Proteins
- Central dogma: transcription, translation
- Pathways

*From short course developed with Robert Gentleman, Biostatistics, Harvard.*
A brief history

- **1865** Genes are particulate factors
- **1903** Chromosomes are hereditary units
- **1910** Genes lie on chromosomes
- **1913** Chromosomes contain linear arrays of genes
- **1927** Mutations are physical changes in genes
- **1931** Recombination is caused by crossing over
- **1944** DNA is the genetic material
- **1945** A gene codes for a protein
- **1953** DNA is a double helix
- **1958** DNA replicates semiconservatively
- **1961** Genetic code is triplet
- **1977** DNA can be sequenced
- **1997** Genomes can be sequenced

**Gregor Mendel (1823-1884)**

**Thomas Hunt Morgan (1866-1945)**

**Francis Crick (1916- )**

**James D. Watson (1928- )**
From chromosomes to proteins

Cells
Cells

- **Cells**: the fundamental working units of every living organism.

- **Metazoa**: multicellular organisms. E.g. humans: trillions of cells.

- **Protozoa**: unicellular organisms. E.g. yeast, bacteria.
Cells

• Each cell contains a complete copy of an organism’s genome, or blueprint for all cellular structures and activities.

• 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.
Cell composition

• 90% water.
• Of the remaining molecules, dry weight
  – 50% protein
  – 15% carbohydrate
  – 15% nucleic acid
  – 10% lipid
  – 10% miscellaneous.
• By element: 60% H, 25% O, 10% C, 5% N.

web.mit.edu/esgbio/www/cb/cellbasics.html
The genome

- The genome is distributed along chromosomes, which are made of compressed and entwined DNA.

- A (protein-coding) gene is a segment of chromosomal DNA that directs the synthesis of a protein.
Eukaryotes vs. prokaryotes
Eukaryotes vs. prokaryotes

• **Prokaryotic cells**: lack a distinct, membrane-bound nucleus. E.g. bacteria.

• **Eukaryotic cells**: distinct, membrane-bound nucleus. Larger and more complex in structure than prokaryotic cells. E.g. mammals, yeast.
The eukaryotic cell
The eukaryotic cell

- **Nucleus**: membrane enclosed structure which contains chromosomes, i.e., DNA molecules carrying genes essential to cellular function.
- **Cytoplasm**: the material between the nuclear and cell membranes; includes fluid (cytosol), organelles, and various membranes.
- **Ribosome**: small particle composed of RNAs and proteins that functions in protein synthesis.
The eukaryotic cell

- **Organelle**: a membrane enclosed structure found in the cytoplasm.
- **Vesicle**: small cavity or sac, especially one filled with fluid.
- **Mitochondrion**: organelle found in most eukaryotic cells in which respiration and energy generation occurs.
- **Mitochondrial DNA**: codes for ribosomal RNAs and transfer RNAs used in the mitochondrion; contains only 13 recognizable genes that code for polypeptides.
The eukaryotic cell

- **Centrioles**: either of a pair of cylindrical bodies, composed of microtubules (spindles). Determine cell polarity, used during mitosis and meiosis.
- **Endoplasmic reticulum**: network of membranous vesicles to which ribosomes are often attached.
- **Golgi apparatus**: network of vesicles functioning in the manufacture of proteins.
- **Cilia**: very small hairlike projections found on certain types of cells. Can be used for movement.
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).
Chromosomes

HUMAN CHROMOSOMES

1) Centromere
2) Telomere
3) Chromatid

www.accessexcellence.com/AB/GG/
Of mice and men
Cell divisions

• **Mitosis:** Nuclear division which produces two daughter diploid nuclei identical to the parent nucleus.
  How each cell can be traced back to a single fertilized egg.

• **Meiosis:** Two successive nuclear divisions which produce four daughter haploid nuclei, different from the 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
Meiosis

1st cell division of meiosis

1. Prophase 1
2. Metaphase 1
3. Anaphase 1
4. Telophase 1

2nd cell division of meiosis

1. Prophase 2
2. Metaphase 2
3. Anaphase 2
4. Telophase 2

Sandrine Du

www.accessexcellence.com/AB/GG/
Meiosis vs. mitosis

**Meiosis**
- **Meiotic division 1**
  - DNA replication
  - Homologous chromosomes at the same level on equatorial plate
- **Meiotic division 2**
  - Cell division

**Mitosis**
- DNA replication
- Homologous chromosome line up individually at the equatorial plate
- Cell division
Dividing cell

Image of cell at metaphase from fluorescent-light microscope.

Recombination
Recombination

One cell

Crossing-over

Four cells
(egg or sperm)

Parental
Recombinant
Recombinant
Parental
Recombination
Chromosomes and DNA
DNA structure

“We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.”

DNA structure

• 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;
  – one of four nitrogen bases:
    • purines: adenine (A) and guanine (G),
    • pyrimidines: cytosine (C) and thymine (T).
DNA structure

• Base-pairing occurs according to the following rule:
  – C pairs with G,
  – A pairs with T.

• The two chains are held together by hydrogen bonds between nitrogen bases.
DNA structure

http://academy.d20.co.edu/kadets/lundberg/dnapic.html
DNA structure
DNA structure

Four nucleotide bases:
- purines: A, G
- pyrimidines: T, C

A pairs with T, 2 H bonds
C pairs with G, 3 H bonds

Nucleotide bases

Adenine (A)

Purines

Guanine (G)

Pyrimidines

Thymine (T) (DNA)

Cytosine (C)

Uracil (U) (RNA)
Nucleotide base pairing

G-C pair

A-T pair

www.accessexcellence.com/AB/GG/
DNA structure

- Polynucleotide chains are directional molecules, with slightly different structures marking the two ends of the chains, the so-called 3' end and 5' end.
- The 3' and 5' notation refers to the numbering of carbon atoms in the sugar ring.
- The 3' end carries a sugar group and the 5' end carries a phosphate group.
- The two complementary strands of DNA are antiparallel (i.e., 5' end to 3' end directions for each strand are opposite)
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.
“It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.”

DNA replication

Three possible models

Original DNA double helix

Conservative replication

Dispersive replication

Semiconservative replication
DNA replication

Semiconservative replication

Original DNA Helix

DNA helixes after one round of replication
DNA replication

• In the replication of a double-stranded or duplex DNA molecule, both parental (i.e. original) DNA strands are copied.
• The parental DNA strand that is copied to form a new strand is called a template.
• When copying is finished, the two new duplexes each consist of one of the original strands plus its complementary copy - semiconservative replication.
DNA replication

Base pairing provides the mechanism for DNA replication.
DNA replication

• Many enzymes are required to unwind the double helix and to synthesize a new strand of DNA.

• The unwound helix, with each strand being synthesized into a new double helix, is called the replication fork.

• DNA synthesis occurs in the 5' → 3' direction.
DNA replication

Collaboration of Proteins at the Replication Fork
DNA replication
DNA replication

![Diagram of DNA replication showing synthesis of leading and lagging strands.](image)
DNA replication

Replication fork grows...

Most recently synthesized DNA

...and grows Okazaki fragments
DNA replication

**Figure 13.1** Overview: DNA synthesis occurs by adding nucleotides to the 3'-OH end of the growing chain, so that the new chain is synthesized in the 5'-3' direction. The precursor for DNA synthesis is a nucleoside triphosphate, which loses the terminal two phosphate groups in the reaction.
DNA replication
DNA replication

**Figure 13.8** Synthesis of Okazaki fragments requires priming, extension, removal of RNA, gap filling, and nick ligation.

- **Primase** synthesizes RNA
- **DNA polymerase III** extends RNA primer into Okazaki fragment
- Next Okazaki fragment is synthesized
- **DNA polymerase I** uses nick translation to replace RNA primer with DNA
- **Ligase** seals the nick
Enzymes in DNA replication

1. **Topoisomerase**: removes supercoils and initiates duplex unwinding.
2. **Helicase**: unwinds duplex.
3. **DNA polymerase**: synthesizes the new DNA strand; also performs proofreading.
4. **Primase**: attaches small RNA primer to single-stranded DNA to act as a substitute 3'OH for DNA polymerase to begin synthesizing from.
5. **Ligase**: catalyzes the formation of phosphodiester bonds.
6. **Single-stranded binding proteins**: maintain the stability of the replication fork.
DNA polymerase

- There are different types of polymerases, DNA polymerase III is used for synthesizing the new strand.
- DNA polymerase is a holoenzyme, i.e., an aggregate of several different protein subunits.
- DNA polymerase proceeds along the template and recruits free dNTPs (deoxynucleotide triphosphate) to hydrogen bond with their appropriate complementary dNTP on the template.
- The energy stored in the triphosphate is used to form the covalent bonds.
- DNA polymerase uses a short DNA fragment or primer with a 3'OH group onto which it can attach a dNTP.
β-subunit of DNA polymerase III holoenzyme forms a ring that completely surrounds a DNA duplex.
Proteins

http://www.biochem.szote.u-szeged.hu/astrojan/protein1.htm
Proteins

- **Proteins**: large molecules composed of one or more chains of amino acids, polypeptides.
- **Amino acids**: class of 20 different organic compounds containing a basic amino group (-NH₂) and an acidic carboxyl group (-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
Amino acids with hydrophobic side groups

- Valine (val)
- Leucine (leu)
- Isoleucine (ile)
- Methionine (met)
- Phenylalanine (phe)

Amino acids with hydrophilic side groups

- Asparagine (asn)
- Glutamic acid (glu)
- Glutamine (gln)
- Histidine (his)
- Lysine (lys)
- Arginine (arg)

Amino acids that are in between

- Glycine (gly)
- Alanine (ala)
- Serine (ser)
- Threonine (thr)
- Tyrosine (tyr)
- Tryptophan (trp)
Amino acids

FAMILIES OF AMINO ACIDS
The common amino acids are grouped according to whether their side chains are:
- Acidic
- Basic
- Uncharged polar uncharged
- Neutral
These 20 amino acids are given both their trivial and common chemical names.
Their amino acid is abbreviated as:

BASIC SIDE CHAINS
- Lysine
- Arginine
- Histidine
These are positively charged residues because their side chains are amines (NH₂). These groups have a relatively weak affinity for an H⁺ and are only partly positive at neutral pH.

ACIDIC SIDE CHAINS
- Aspartic acid
- Glutamic acid
These are negatively charged residues. Their side chains are carboxyl groups (COOH), which have strong affinities for H⁺ ions.

NONPOLAR SIDE CHAINS
- Alanine
- Valine
These are uncharged and their side chains are nonpolar hydrocarbons.

THE AMINO ACID
The general formula of an amino acid is:

Optical Isomers
The stereoisomers of amino acids are enantiomers.

PEPTIDE BONDS
Amino acids are covalently joined together by an amide linkage, called a peptide bond.

Peptide bond
- The amide linkage can be formed from an amine and a carboxyl group.
- Peptide bonds are covalent bonds that hold the amino acids together.

Optical Isomers:
- Enantiomers are mirror images of each other.
- Amino acids can exist in both L- and D- configurations.

Proteins consist almost entirely of amino acids.

Peptide bond
- The C-N bond is the Site of nitrogen fixation.
- Peptide bonds can form between two amino acids, forming a peptide bond.

Although the amide bond is not charged at neutral pH, it can be polar.

Proteins are long polymers of amino acids, held together by peptide bonds, and they are always written with the N terminus toward the left. The sequence of this polypeptide is written as Gln-Glu-Ser.
Proteins

Primary protein structure is sequence of a chain of amino acids

Amino Acids

Amino Acid

www.accessexcellence.com/AB/GG/
Proteins

Primary protein structure
- is sequence of a chain of amino acids

Secondary protein structure
- occurs when the sequence of amino acids are linked by hydrogen bonds

Tertiary protein structure
- occurs when certain attractions are present between alpha helices and pleated sheets.

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

**CELL TYPES**

There are over 200 types of cells in the human body. These are categorized into a few main types, such as:

- **Epithelia**
- **Connective tissue**
- **Muscle**
- **Blood**
- **Nervous tissue**
- **Germs**

Each type of cell has a specific function within the body. For example, epithelial cells line the surfaces of the body, connective tissue supports and connects other structures, and muscle cells allow movement.

**Epithelia**

Epithelial cells form a protective layer over the body. They are specialized for different functions, such as secreting hormones, absorbing nutrients, or excreting wastes.

**Connective tissue**

Connective tissue is the most abundant type of tissue in the body. It includes bones, cartilage, and fascia, which provide support and structure.

**Muscle**

Muscle tissue allows movement. There are three main types:

- **Skeletal muscle** is attached to bones and allows voluntary movement.
- **Cardiac muscle** is found in the heart and is responsible for heart contractions.
- **Smooth muscle** is found in the walls of organs and blood vessels, allowing controlled movements.

**Blood**

Blood is a connective tissue that circulates throughout the body. It contains red and white blood cells, platelets, and plasma.

**Nervous tissue**

Nervous tissue is responsible for communication in the body. It includes neurons, which are the cells that transmit signals, and neuroglia, which support the neurons.

**Germs**

Germs are single-celled organisms that can cause disease. They can be classified into several types, such as bacteria, viruses, and fungi.

*Source: Sandrine Dudoit © Copyright 2003, all rights reserved*
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 how much each gene is expressed.
- On average, 40% of our genes are expressed at any given time.
Central dogma

DNA → transcription → RNA → translation → protein
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
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

Start codon: initiation of translation (AUG, Met).
Stop codons: termination of translation.

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.
Protein synthesis

1. Transcription

DNA → RNA polymerase → mRNA

RNA nucleotides

nuclear membrane

2. Translation

mRNA → Ribosome → polypeptide chain

Anticodon → rRNA → amino acids → proteins

Protein synthesis
Transcription

• Analogous to DNA replication: several steps and many enzymes.

• **RNA polymerase** synthesizes an RNA strand complementary to one of the two DNA strands.

• The RNA polymerase recruits **rNTPs** (ribonucleotide triphosphate) in the same way that DNA polymerase recruits **dNTPs** (deoxynucleotide triphosphate).

• However, synthesis is **single stranded** and only proceeds in the 5' to 3' direction of mRNA (no Okazaki fragments).
Transcription

• The strand being transcribed is called the template or antisense strand; it contains anticodons.

• The other strand is called the sense or coding strand; it contains codons.

• The RNA strand newly synthesized from and complementary to the template contains the same information as the coding strand.
Transcription

5’ ...A T G G C C T G G A C T T C A... 3’  
3’ ...T A C C G G A C C T G A A G T... 5’  

Sense strand of DNA  
Antisense strand of DNA

Transcription of antisense strand

5’ ...A U G G C C U G G A C U U C A... 3’  
mRNA

Translation of mRNA

Met — Ala — Trp — Thr — Ser — Peptide
Transcription

- **Promoter.** Unidirectional sequence upstream of the coding region (i.e., at 5' end on sense strand) that tells the RNA polymerase both *where* to start and on *which strand* to continue synthesis. E.g. TATA box.
- **Terminator.** Regulatory DNA region signaling end of transcription, at 3' end.
- **Transcription factor.** A protein needed to initiate the transcription of a gene, binds either to specific DNA sequences (e.g. promoters) or to other transcription factors.
Transcription

Figure 9.2 Overview: a transcription unit is a sequence of DNA transcribed into a single RNA, starting at the promoter and ending at the terminator.
Exons and introns

• Genes comprise only about 2% of the human genome.
• The rest consists of non-coding regions
  – chromosomal structural integrity,
  – cell division (e.g. centromere)
  – regulatory regions: regulating when, where, and in what quantity proteins are made.
• The terms exon and intron refer to coding (translated into a protein) and non-coding DNA, respectively.
Exons and introns

www.accessexcellence.com/AB/GG/
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% of the genes are subject to alternative splicing.
Alternative splicing

Primary isoform

Cryptic exon

Exon extension (5' or 3')

Exon skipping

Exon truncation
Immunoglobulin

• B cells produce antibody molecules called immunoglobulins (Ig) which fall in five broad classes: IgG, IgA, IgM, IgD, IgE.

• Diversity of Ig molecules
  – Somatic recombination prior to transcription (VDJ recombination).
  – Alternative splicing.
  – Post-translational proteolysis, glycosylation.
Translation

• **Ribosome:**
  – cellular factory responsible for protein synthesis;
  – a large subunit and a small subunit;
  – structural RNA and about 80 different proteins.

• **transfer RNA (tRNA):**
  – adaptor molecule, between mRNA and protein;
  – specific **anticodon** and **acceptor site**;
  – specific **charger protein**, can only bind to that particular tRNA and attach the correct amino acid to the acceptor site.
Translation

• Initiation
  – Start codon AUG, which codes for methionine, Met.
  – Not every protein necessarily starts with methionine. Often this first amino acid will be removed in post-translational processing of the protein.

• Termination:
  – stop codon (UAA, UAG, UGA),
  – ribosome breaks into its large and small subunits, releasing the new protein and the mRNA.
Translation

Initiation
30S subunit on mRNA binding site is joined by 50S subunit and aminoacyl-tRNA binds

Elongation
Ribosome moves along mRNA and length of protein chain extends by transfer from peptidyl-tRNA to aminoacyl-tRNA

Termination
Polypeptide chain is released from tRNA, and ribosome dissociates from mRNA

www.oup.co.uk/best.textbooks/biochemistry/genesvii
tRNA

- The tRNA has an **anticodon** on its mRNA-binding end that is complementary to the codon on the mRNA.
- Each tRNA only binds the appropriate amino acid for its anticodon.
Post-translational processing

• Folding.
• Cleavage by a proteolytic (protein-cutting) enzyme.
• Alteration of amino acid residues
  – phosphorylation, e.g. of a tyrosine residue.
  – glycosylation, carbohydrates covalently attached to asparagine residue.
  – methylation, e.g. of arginine.
• Lipid conjugation.
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 interact to produce various components.

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

- Transforming Growth Factor beta, TGF-β, plays an essential role in the control of development and morphogenesis in multicellular organisms.

- The basic pathway provides a simple route for signals to pass from the extracellular environment to the nucleus, involving only four types of molecules.
TGF-β pathway

Four types of molecules
• TGF-β
• TGF-β type I receptors
• TGF-β type II receptors
• SMADS, a family of signal transducers and transcriptional activators.
TGF-β pathway
TGF-\(\beta\) pathway

• Extracellular TGF–\(\beta\) ligands transmit their signals to the cell's interior by binding to type II receptors, which form heterodimers with type I receptors.

• The receptors in turn activate the SMAD transcription factors.
TGF-\(\beta\) pathway

- Phosphorylated and receptor-activated SMADs (R-SMADs) form heterodimers with common SMADs (co-SMADs) and translocate to the nucleus.

- In the nucleus, SMADs activate or inhibit the transcription of target genes, in collaboration with other factors.
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 gene expression levels (e.g. mRNA levels) and pathways.

• It is not clear to what extent microarray gene expression data will be informative.
References

• **Access Excellence**

• **Genes VII**
  [www.oup.co.uk/best.textbooks/biochemistry/genesvii/](http://www.oup.co.uk/best.textbooks/biochemistry/genesvii/)

• **Human Genome Project Education Resources**
  [www.ornl.gov/hgmiis/education/education.html](http://www.ornl.gov/hgmiis/education/education.html)

• **Kimball’s Biology Pages**
  [www.ultranet.com/~jkimball/BiologyPages/](http://www.ultranet.com/~jkimball/BiologyPages/)

• **MIT Biology Hypertextbook**

• [http://academy.d20.co.edu/kadets/lundberg/dnapic.html](http://academy.d20.co.edu/kadets/lundberg/dnapic.html)