'Central Dogma' of Molecular Biology

THE CENTRAL DOGMA
DNA $\overset{\text{transcription}}{-\longrightarrow}$ **RNA** $\overset{\text{translation}}{-\longrightarrow}$ protein $\longrightarrow$ phenotype

DNA: 'deoxyribonucleic acid', fundamental biological molecule coding instructions (genetic sequence) for proteins; consists of two (oriented) strands of **base-pairs** A↔T and G↔C wound in *double helix* in cell nuclei.

mRNA: 'messenger' ribonucleic acid, molecule which carries a transcribed sequence mapped from one strand of a part of DNA to cell organelles (ribosomes) where proteins are manufactured.

Genetic Code: the DNA molecule consists of a chained sequence of 4 nucleotides (*'bases'*) (A=adenine, G=guanine, C=Cytosine, T=thymine, oriented in the way they connect), grouped into triplets called **codons** (64-symbol alphabet) which correspond in a unique way to 20 amino acids which chain into proteins called 'polypeptides'.

Transcription: unique symbol-by-symbol mapping from portion of DNA sequence with alphabet A,T,C,G to (complementary) RNA sequence with alphabet U, A, G, C which occurs biologically when DNA strands are separated enzymatically (by 'DNA polymerase')

Glossary of Additional Terms

DNA sequence $\longrightarrow$ RNA sequence $\longrightarrow$ molecular structure $\longrightarrow$ function

Exon: DNA segment which after transcription to RNA codes directly to peptide units of a polypeptide, ie which 'is expressed in protein' (200 base-pairs on average, in human genome)

Intron: DNA segment which is not directly expressed for protein, involved in regulation,'splicing' and other functions (not all known) each 2000 bp's on average, in human genome)

Translation: process by which non-coding segments are defined ('splicing') and removed, and the rest is coded symbol-by-symbol to amino acids to form proteins

Gene: DNA segment units consisting of start- and stop-codons with exons and introns between which together code for biologically meaningful protein(s). 30,000–40,000 Genes in human genome. Typical gene has on average between 5 and 8 exons, 8000 base-pairs.

Genome: Total human complement of DNA $\approx 3.2 \times 10^9$ bp. Amount estimated to code for protein between 2 and 5% (depends on whether only segments directly coding are counted)

More Terms & Background Concepts

Splicing: enzyme- (and intron-) controlled snipping of DNA segment to be transcribed to RNA. 'Alternative splicings' mean that the same Gene could code for multiple proteins depending on where snips are made.

Markers: locations along DNA sequence at which known available enzymes can be made to snip, and at which the codon for individual DNA strands can therefore be easily recovered in the laboratory

Reverse Transcription: natural procedure in which relatively unstable (m)RNA codes back symbol-by-symbol to 'complementary DNA" (**cDNA**) sequence which is more stable. Doing this in lab with 'labelled' components ('fluorescent dyes') prepares a sample so that bound cDNA will be detectable by suitable laser light.

Hybridization: spontaneous attachment (chemical binding) of exactly complementary DNA to DNA segments except where geometrically obstructed, e.g. by folding.

Cross-hybridization: DNA segments attaching to non-complementary segments, e.g. sequence partially shared. **NB**: Human DNA sequence is 'repeat-rich' ($> 50\%$ repeated), including large duplicated segments (50–500kb) with 'high sequence identity' (98–99.9% symbol-matches)

DNA Sequencing & Alignment

Key point is that the distribution of DNA bases (A,T,G,C) is very nonuniform: there is *context* related to regions (Gene, non-coding region, Exon, Intron, etc., R-bands = 'early replicating' regions in DNA synthesis, G-bands = 'late replicating' regions, etc.)

There is both **short-term memory** (certain bases following others within regions with greater frequencies) and **long-term memory** (e.g., persistence of 'region', and at translation stage, interaction of distant amino acids in 'protein-folding').

Key Problems Requiring Sequence Models:

1. Search & modelling of genes within sequence
2. Probabilistic algorithms for alignment
3. Probabilistic measures of similarity

Main Class of Probabilistic Models: **Hidden Markov**

'Context' modelled differently in DNA and RNA.

Key References & Topics

0. High school biology text or: Campbell, Reece, Mitchell **Biology** (currently used as BIOL 105 text).

1. Hamadeh, H. and Afshari, C. (2000) *Gene chips and functional genomics.* Amer. Scientist **88**, 508–515.

2. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) **Biological Sequence Analysis.** Cambridge Univ. Press.

GENERAL TOPICS OF SEMINAR

- Hidden Markov Models (could begin with basic papers of Baum et al. 1967 or use 2. above)
- Sequence-searching and matching (Smith & Waterman, Karlin & Altschul or many others)
- Phylogenetic aspects of sequence-similarity (eg., Aldous *Statist. Sci.* paper)
- Hierarchical clustering (Eisen et al. or many other papers, in Microarray area; or Hartigan, Hubert & Arabie or others in Clustering Methodology)
- Miscellaneous issues (normalization, multiple comparisons, experimental design, and modeling relating to sources of variation in microarray spot intensities)

(cDNA) Microarray Experimental Steps

(1) 'Attach ' single-strand DNA segments for each of $i = 1, \ldots, N$ (eg 1.e4) Genes robotically to $N \times p$ distinct tiny 'spots' on treated membrane or glass slide (eg $p = 30$).

(2) Extract tissue samples from one or two sources (eg healthy and cancerous organ cells); purify mRNA from each.

(3) Reverse-transcribe and copy ('clone') mRNA from tissue samples to cDNA in solutions where fluorescent dyes Cye3 and 5 will be chemically taken up in the phosphate/sugar backbone of newly formed cDNA.

(4) Apply cloned/amplified cDNA solution(s) labelled with fluorescent dye to (subset of) the $N \times p$ array of spots on the slide. For example each of the p columns might represent a distinct tissue-sample (eg from distinct individual or cell-line). In two-dye comparative setting, **both** cDNA preparations are applied to the same (columns of) spots.

(5) 'Wash' the slide so that only hybridized (tightly chemically bound) cDNA-to-fixed DNA will remain attached on the slide.

Microarray Experimental Steps, cont'd

(6) Image-processing consists of laser-light shone onto the slide which causes fluorescently labelled cDNA components to register red or green (according to Cye 3 or 5) dye. Greater length of bound segments 'should' correspond to greater intensities.

(7) Preprocessing involves 'subtracting out' background intensities, and in some cases, forming local contrasts to remove local spatial effects on registered intensities.

(8) Resulting (preprocessed) data consists of $N \times p$ matrix of intensities registered in 1 or two channels. Such data is sometimes replicated across additional slides and/or tissue-sources (cell-lines).