Introduction to Microarray Data Analysis and Gene Networks

Alvis Brazma

European Bioinformatics Institute
A brief outline of this course

• What is gene expression, why it’s important
• Microarrays and how they measure expression
• Steps in microarray data analysis
• Try some basic analysis of real microarray data
• A bit of theory about microarray data analysis
• Gene networks, what are they
• Methods or describing gene networks
• How microarrays can help to understand them
• Some more fancy stuff about gene networks
What will be needed to complete this course

- Complete some coursework on real data analysis using tools we’ll try in the lectures
- Details to be finalised later this week
1. All you need to know about biology about this course in 10 – 20 min

- http://www.ebi.ac.uk/microarray/biology_intro.html
- Genomes and genes
Central dogma of molecular biology

DNA → transcription → RNA → translation → Protein
Four different nucleotides: adenosine, guanine, cytosine, and thymine. They are usually referred to as bases and denoted by their initial letters, A, C, G, and T.

\[
\text{5'} \quad \text{C-G-A-T-T-G-C-A-A-C-G-A-T-G-C} \quad \text{3'} \\
\quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad | \\
\text{3'} \quad \text{G-C-T-A-A-C-G-T-T-G-C-T-A-C-G} \quad \text{5'}
\]
Thus, for many information related purposes, the molecule can be represented as

**CGATTCAACGATGC**

The maximal amount of information that can be encoded in such a molecule is therefore 2 bits times the length of the sequence. Noting that the distance between nucleotide pairs in a DNA is about 0.34 nm, we can calculate that the linear information storage density in DNA is about \(6 \times 10^8\) bits/cm, which is approximately **75 GB or 12.5 CD-Roms per cm**.
Genomes, chromosomes

Genome is a set of DNA molecules. Each chromosome contains (long) DNA molecule per chromosome

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of chromosomes</th>
<th>Genome size in base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td>1</td>
<td>~400,000 - ~10,000,000</td>
</tr>
<tr>
<td>Yeast</td>
<td>12</td>
<td>14,000,000</td>
</tr>
<tr>
<td>Worm</td>
<td>6</td>
<td>100,000,000</td>
</tr>
<tr>
<td>Fly</td>
<td>4</td>
<td>300,000,000</td>
</tr>
<tr>
<td>Weed</td>
<td>5</td>
<td>125,000,000</td>
</tr>
<tr>
<td>Human</td>
<td>23</td>
<td>3,000,000,000</td>
</tr>
</tbody>
</table>

The 23 human chromosomes
Explore the *Homo sapiens* genome

Search Ensembl *Homo sapiens*

Search:

- e.g. chromosome X or 1:1000000:299999 or BRCA2

Karyotype

Click on a chromosome for a closer view

About the Human genome

Assembly

This release is based on the NCBI36 assembly of the human genome (November 2005). The data consists of a reference assembly of the complete genome plus the Celera Y05 and a number of alternative assemblies of individual haplotypic chromosomes or regions.

Full list of assemblies

The International Human Genome Sequencing Consortium have published their scientific analysis of the finished human genome.

- Nature 431, 931 - 945 (21 October 2004)
- WT Sanger Institute Press Release

Annotation
Genes and gene products, proteins

For purposes of this course a gene is a continuous stretch of a genomic DNA molecule, from which a complex molecular machinery can read information (encoded as a string of A, T, G, and C) and make a particular type of a protein or a few different proteins.

<table>
<thead>
<tr>
<th>Organism</th>
<th>The number of predicted genes</th>
<th>Part of the genome that encodes proteins (exons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.Coli (bacteria)</td>
<td>5000</td>
<td>90%</td>
</tr>
<tr>
<td>Yeast</td>
<td>6000</td>
<td>70%</td>
</tr>
<tr>
<td>Worm</td>
<td>18,000</td>
<td>27%</td>
</tr>
<tr>
<td>Fly</td>
<td>14,000</td>
<td>20%</td>
</tr>
<tr>
<td>Weed</td>
<td>25,500</td>
<td>20%</td>
</tr>
<tr>
<td>Human</td>
<td>25,000</td>
<td>&lt; 5%</td>
</tr>
</tbody>
</table>
Central dogma of molecular biology

DNA → transcription → RNA → translation → Protein
RNA

• Like DNA, RNA consists of 4 nucleotides, but instead of the thymine (T), it has an alternative uracil (U)
• RNA is similar to a DNA, but it’s chemical properties are such that it keeps itself single stranded
• RNA is complimentary to a single stranded DNA


| | | | | | | | | | | | | | |

Splicing, translation, proteins

When as according to the ‘central dogma’ genes are transcribed into RNA, there may be ‘interruptions’ called introns.

Because of alternative splicing (e.g., exon skipping) and posttranslational modification there are more proteins than genes.
Proteins, their function

Proteins are chains of 20 different types of aminoacids, and they have complex structures determined by their sequence. The structures in turn determine their functions.
What are gene products doing?

Gene ontology

- **Molecular Function** — elemental activity or task
- **Biological Process** — broad objective or goal
- **Cellular Component** — location or complex
Gene expression

- A human organism has over 250 different cell types (e.g., muscle, skin, bone, neuron), most of which have identical genomes, yet they look different and do different jobs.
- It is believed that less than 20% of the genes are ‘expressed’ (i.e., making RNA) in a typical cell type.
- Apparently the differences in gene expression is what makes the cells different.
Some questions for the golden age of genomics

• How gene expression differs in different cell types?
• How gene expression differs in a normal and diseased (e.g., cancerous) cell?
• How gene expression changes when a cell is treated by a drug?
• How gene expression changes when the organism develops and cells are differentiating?
• How gene expression is regulated – which genes regulate which and how?
Genes are regulated (switched on or off)
Gene regulation networks – outrageously simplified

Specific proteins called transcription factors

promoter
coding DNA
2. Microarrays – a tool for finding which genes have their products being produced (expressed)

Type 1 - single channel (expensive)

Type 2 - dual channel (cheaper)
How do microarrays work

- They exploit the DNA-RNA complementarity principle
- A single stranded DNA complementary to each gene are attached on the slide in a known location
Oligonucleotide chips

GeneChip Probe Array

Hybridized Probe Cell
- Single stranded, labeled RNA target
- Oligonucleotide probe

Image of Hybridized Probe Array
- 1.28 cm²

24 µm

- Millions of copies of a specific oligonucleotide probe
- >200,000 different complementary probes

Compliments of D. Gerhold
How do microarrays work

condition 1

mRNA  cDNA

→

hybridise to microarray

condition 2
A microarray experiment

• Normally it will be more than one array per ‘experiment’
  – More than 2 conditions can be compared
  – The same condition can be used on array many times (replicate experiments) to find out what is the ‘noise level’ or natural gene expression variability within the same experiment
A microarray experiment

Sample
RNA extract
labelled nucleic acid
hybridisation
Microarray
genes
Array design
Protocol

Gene expression data matrix
normalization
integration
Steps in microarray data processing

Array scans → Spots → Quantitations → Genes → Samples
ArrayExpress

ArrayExpress is a public repository for microarray data, which is aimed at storing MIAME-compliant data in accordance with MGED recommendations. The ArrayExpress Data Warehouse stores gene-indexed expression profiles from a curated subset of experiments in the repository.

Experiments

- Search term(s): RNAi
- Query
- Advanced query interface
- Submitter/reviewer login

Expression Profiles

- Genes: Clnmc2
- Species: Homo sapiens
- Query

Microarray Informatics at the EBI

- How to link to ArrayExpress
- How to submit data to ArrayExpress
- ArrayExpress interface tutorial
- Documentation and online help