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

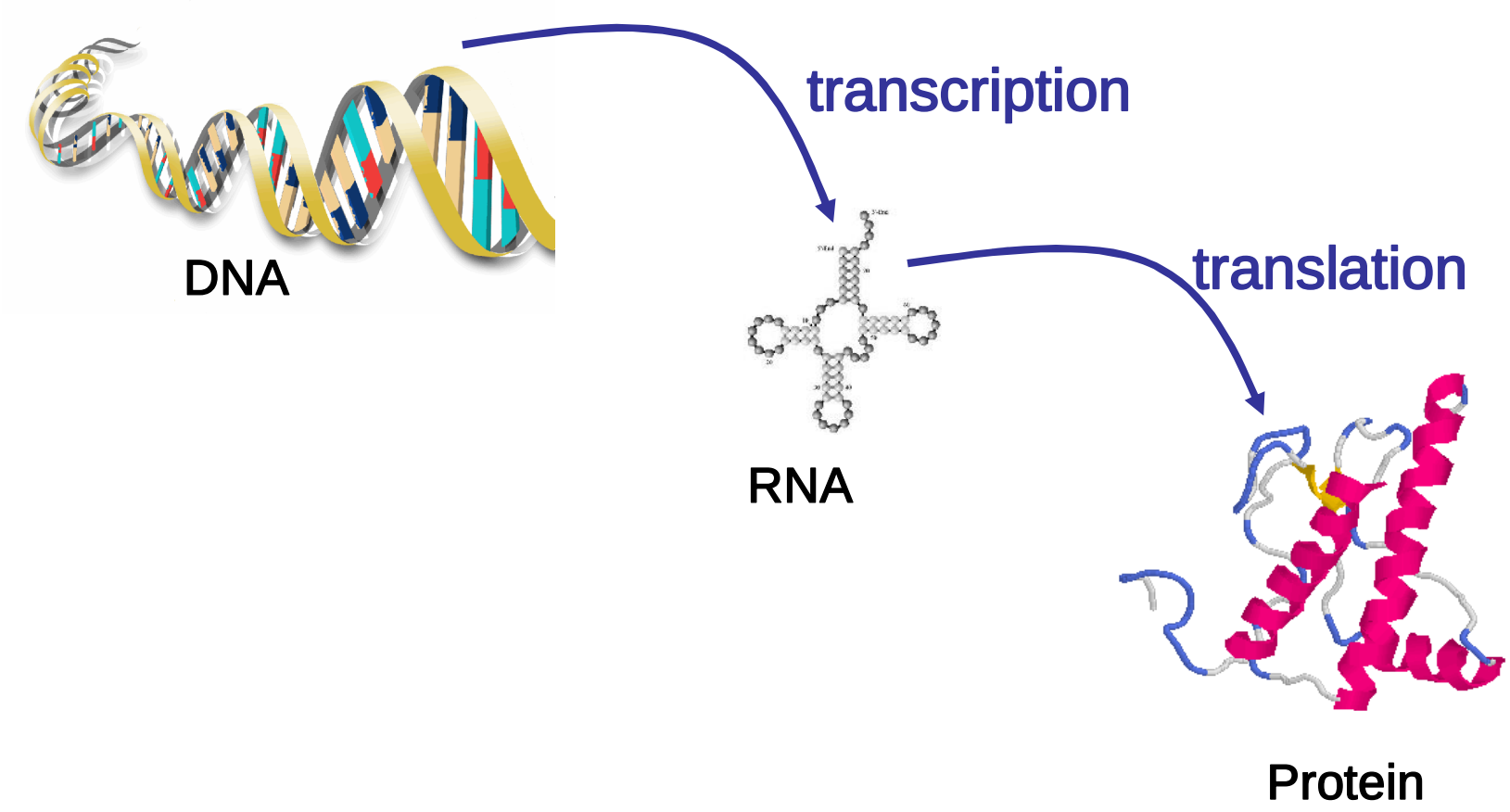
What will be needed to complete this course

- Complete some coursework on real data analysis using tools we'll try in the lectures
- Details to be finalised later this week

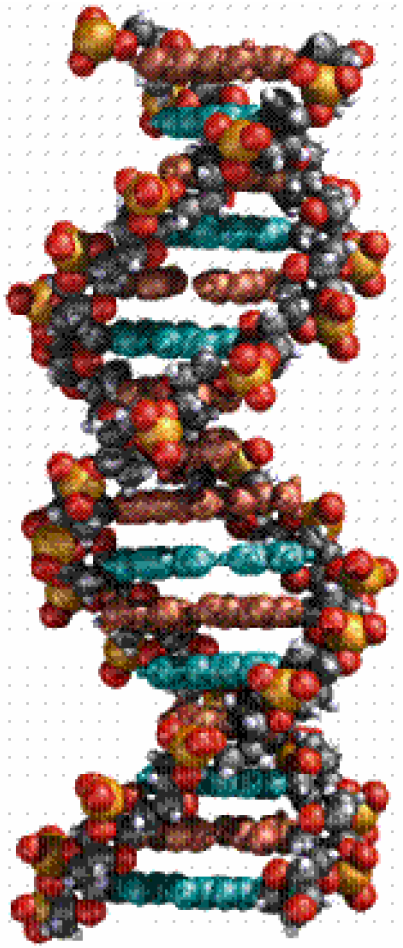
1. All you need to know about biology about this course in 10 – 20 min

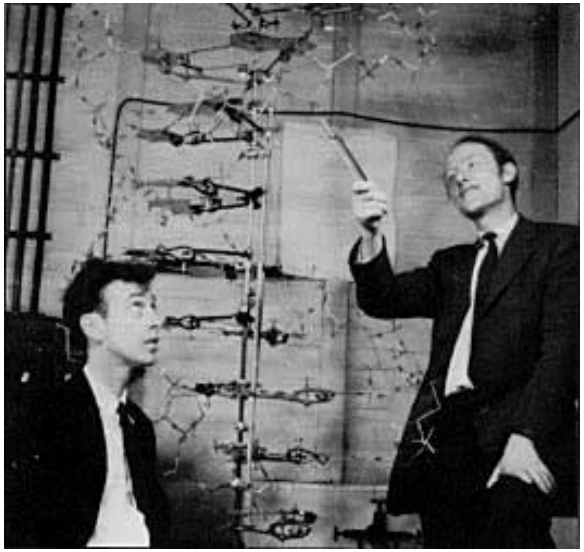
- http://www.ebi.ac.uk/microarray/biology_intro.html
- Genomes and genes

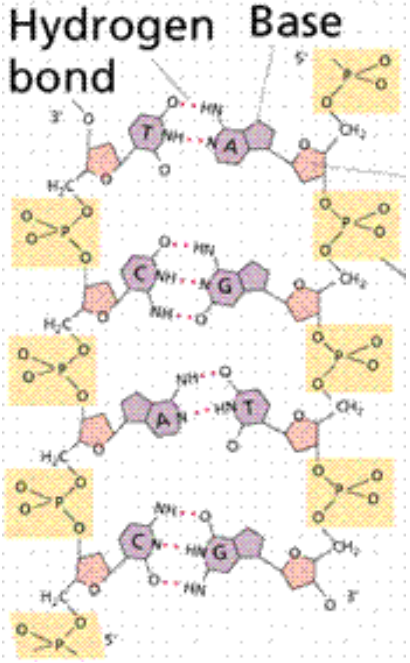
Central dogma of molecular biology



DNA







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**

```
5' C-G-A-T-T-G-C-A-A-C-G-A-T-G-C 3'
   | | | | | | | | | | | | | | |
3' G-C-T-A-A-C-G-T-T-G-C-T-A-C-G 5'
```

DNA - Biology as and information science

```
5' C-G-A-T-T-G-C-A-A-C-G-A-T-G-C 3'
   | | | | | | | | | | | | | | |
3' G-C-T-A-A-C-G-T-T-G-C-T-A-C-G 5'
```

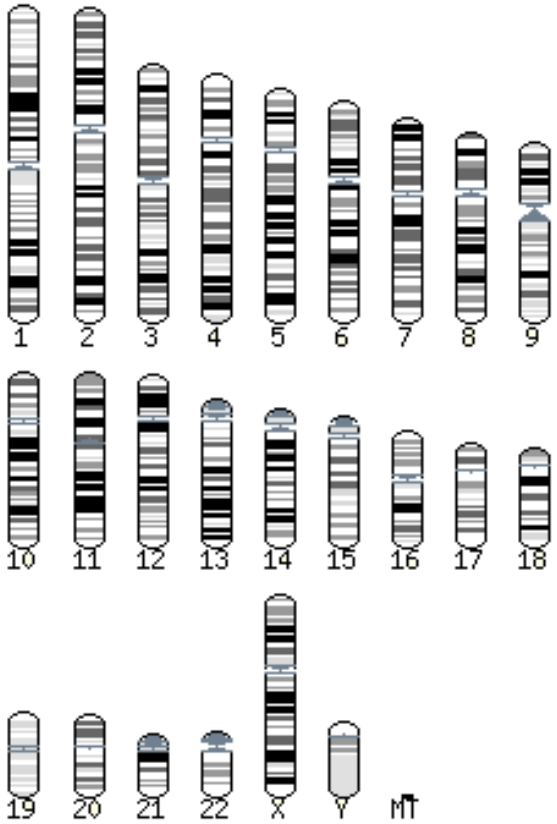
Thus, for many information related purposes, the molecule can be represented as

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 6x10 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) DAN molecule per chromosome



The 23 human chromosomes

Organism	Number or chromosomes	Genome size in base pairs
Bacteria	1	~400,000 - ~10,000,000
Yeast	12	14,000,000
Worm	6	100,000,000
Fly	4	300,000,000
Weed	5	125,000,000
Human	23	3,000,000,000

Human (Homo sapiens) - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.ensembl.org/Homo_sapiens/index.html Go Links

Ensembl Human

Search e!Human: -- Go

e.g. AL138722.15.1.44776, ENSG00000139618

Ensembl release 42 - Dec 2006

HOME · BLAST · BIOMART · SITEMAP HELP

Your Ensembl

- Login or Register
- About User Accounts

Help & Documentation

- Table of Contents
- Helpdesk
- What's New
- About Ensembl
- Downloading data
- Displaying your own data
- Ensembl software

Select a species

- Mammals
  - *Bos taurus* (Cow)
  - *Canis familiaris* (Dog)
  - *Dasypus novemcinctus* (Armadillo)
  - *Echinops telfairi* (Lesser hedgehog tenrec)

Explore the *Homo sapiens* genome

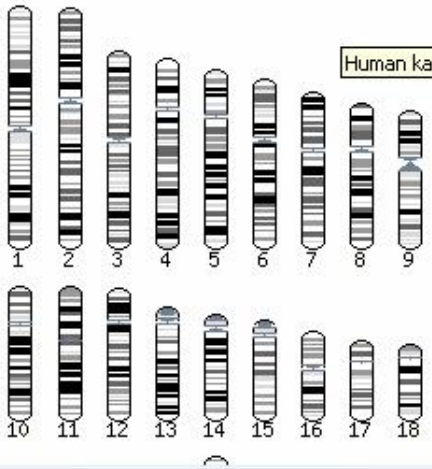
Search Ensembl *Homo sapiens*

Search: Go

e.g. **chromosome X** or **14:10000..200000** or **BRCA2**

Karyotype

Click on a chromosome for a closer view



About the Human genome

Assembly

This release is based on the NCBI 36 assembly of the human genome [November 2005]. The data consists of a reference assembly of the complete genome plus the Celera WGS 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

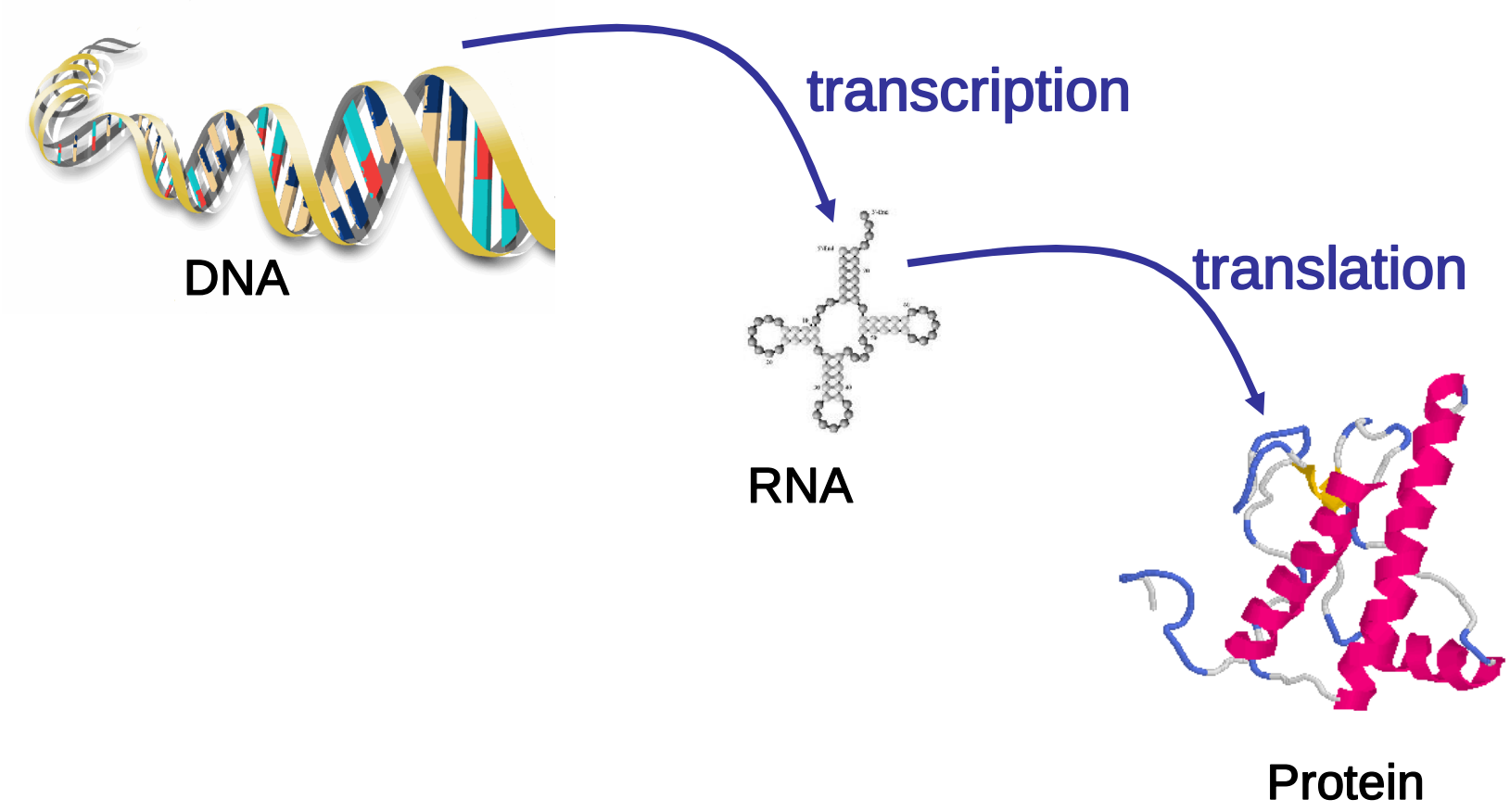
Done Internet

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

Organism	The number of predicted genes	Part of the genome that encodes proteins (exons)
E.Coli (bacteria)	5000	90%
Yeast	6000	70%
Worm	18,000	27%
Fly	14,000	20%
Weed	25,500	20%
Human	25,000	< 5%

Central dogma of molecular biology



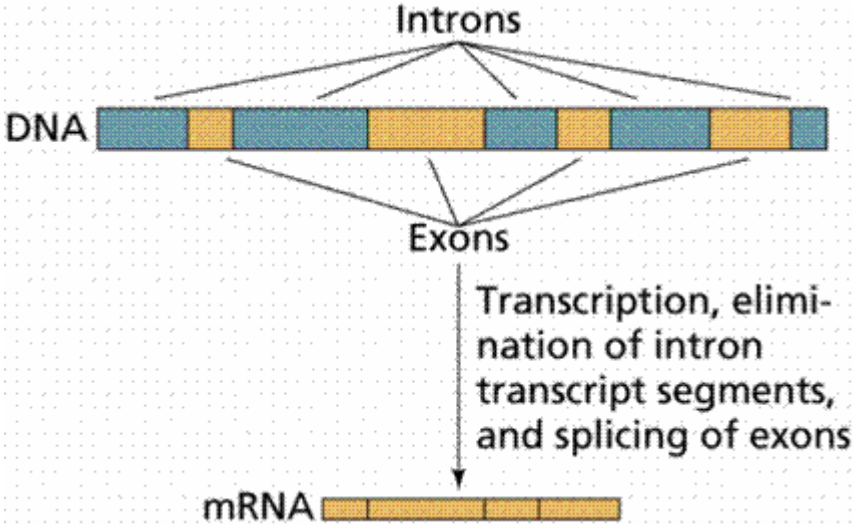
RNA

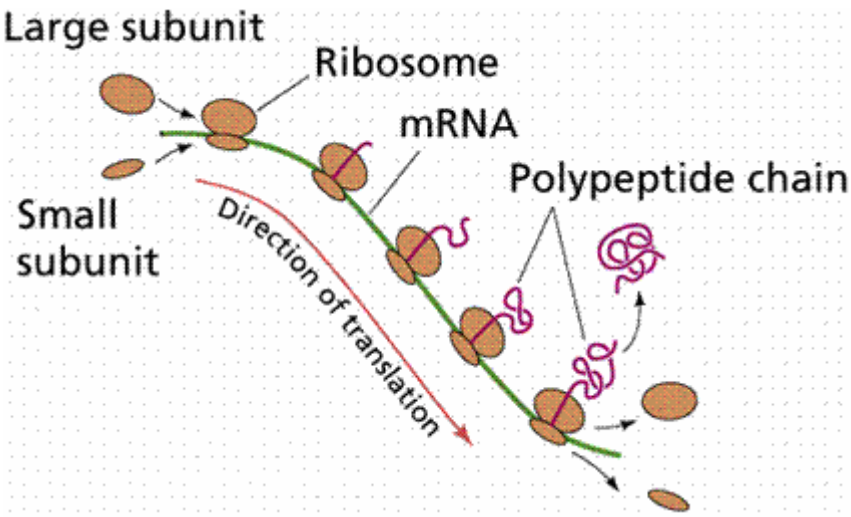
- Like DNA, RNA consists of 4 nucleotides, but instead of the thymine (T), it has an alternative uracil (U)
- RNA is similar to a DNA, but it’s chemical properties are such that it keeps itself single stranded
- RNA is complimentary to a single stranded DNA

```
5' C-G-A-T-T-G-C-A-A-C-G-A-T-G-C 3' DNA
   | | | | | | | | | | | | | | |
3' G-C-U-A-A-C-G-U-U-G-C-U-A-C-G 5' RNA
```

Splicing, translation, proteins

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





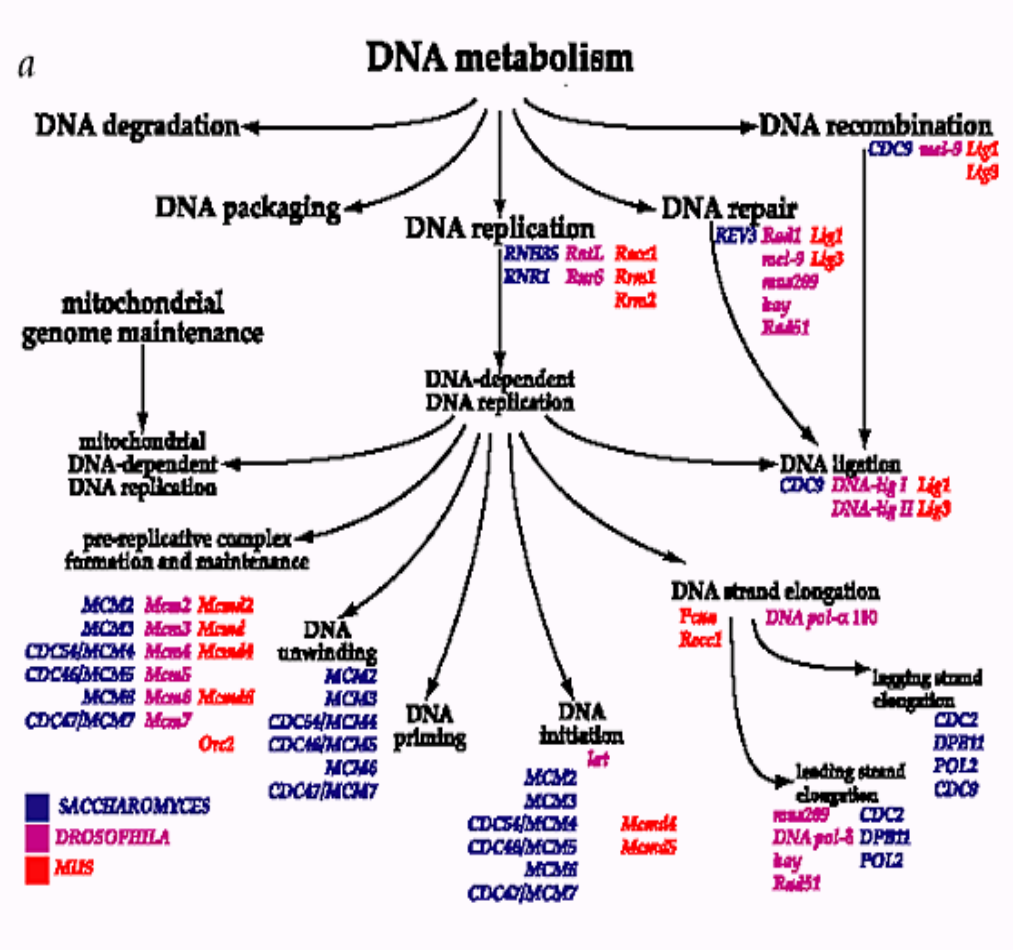
Because of alternative splicing (e.g., exon skipping) and posttranslational modification there are more proteins than genes

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



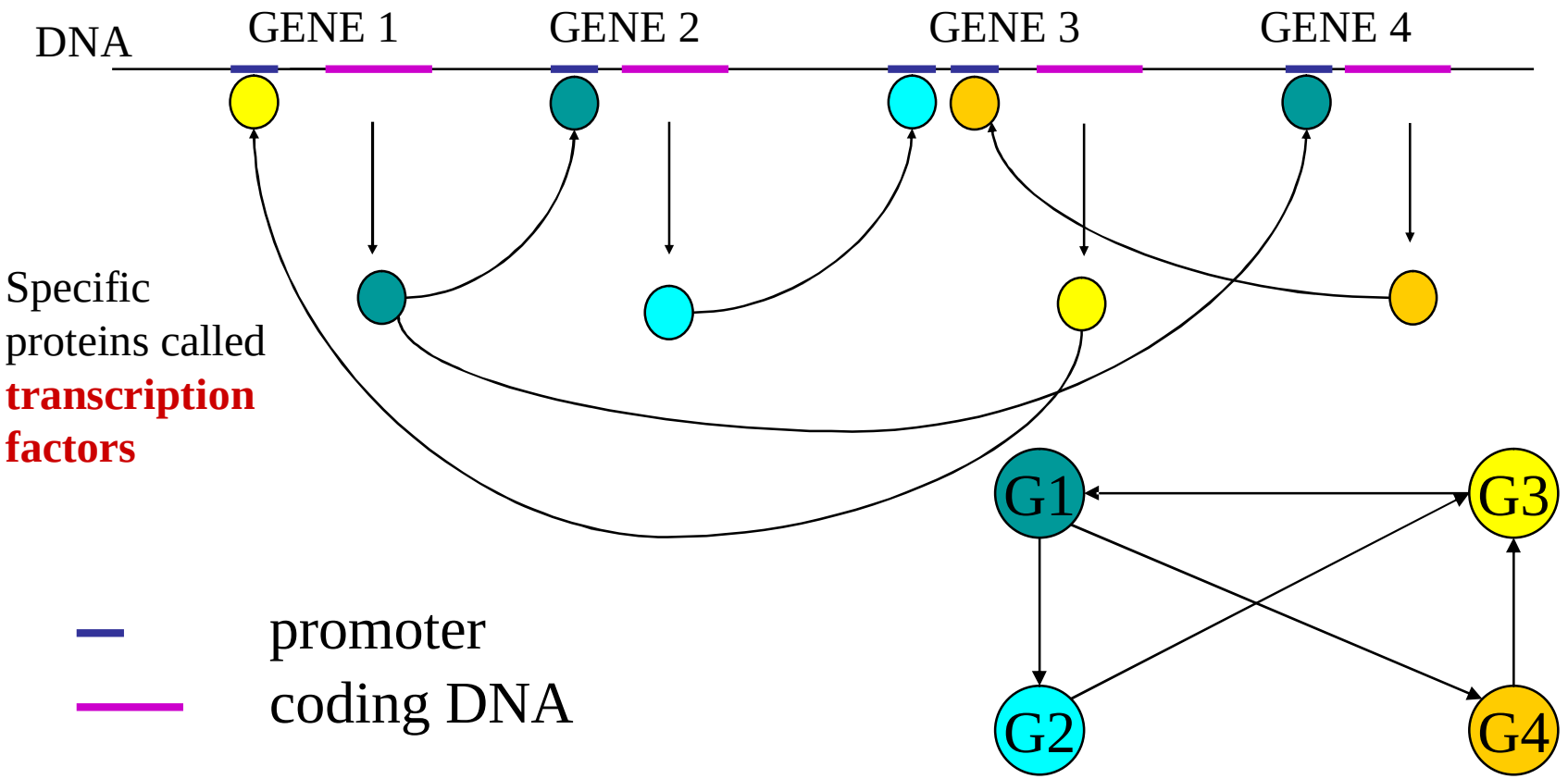
Gene expression

- 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

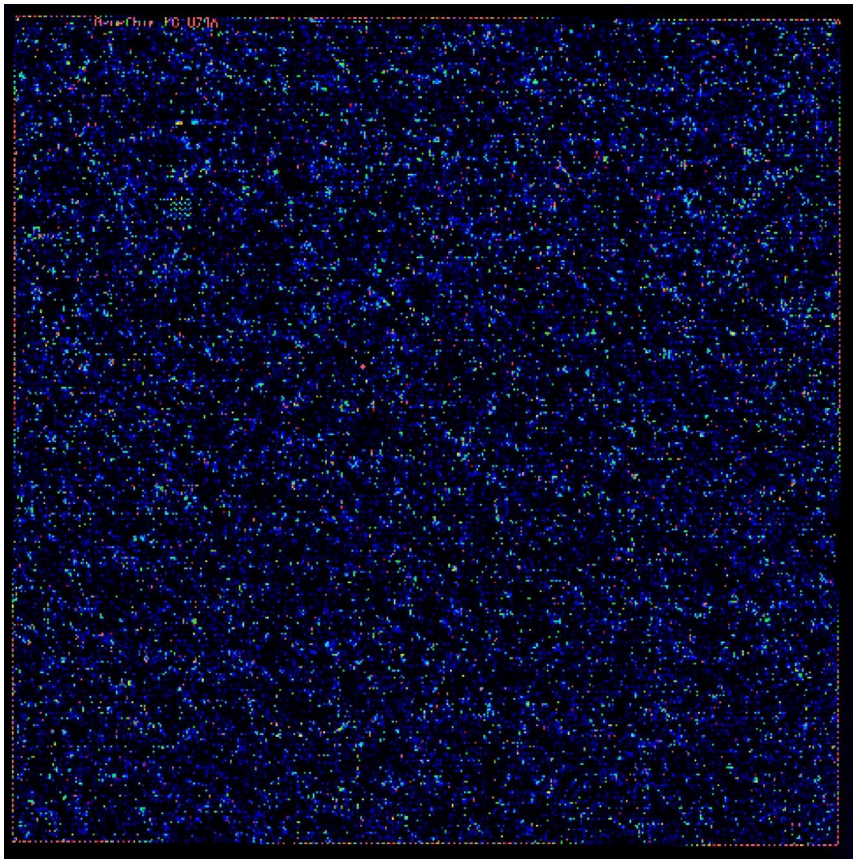
Some questions for the golden age of genomics

- 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?

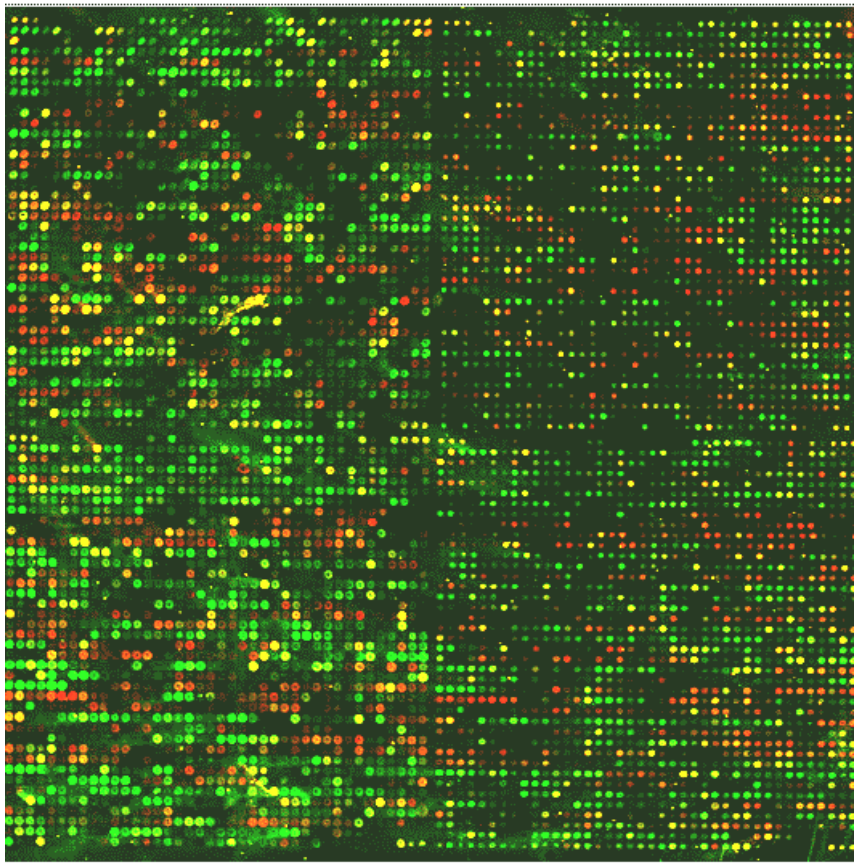
Genes are regulated (switched on or off)
Gene regulation networks – outrageously simplified



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

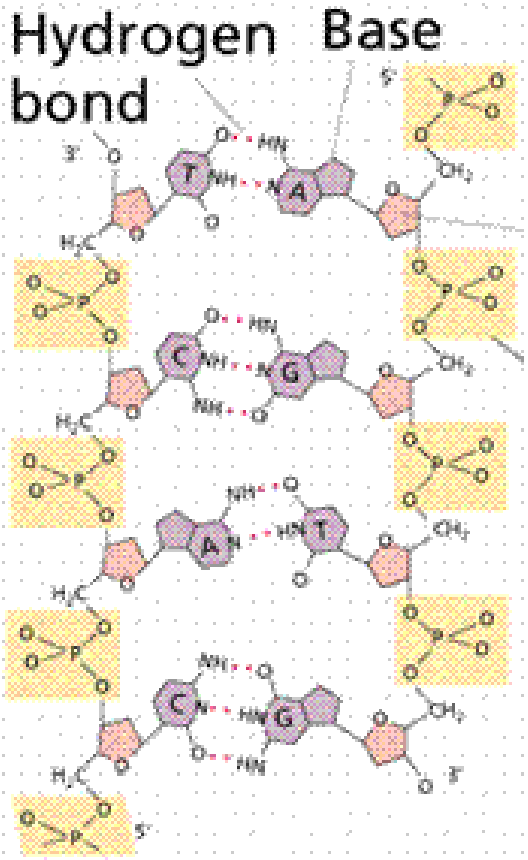


Type 1 - single channel (expensive)



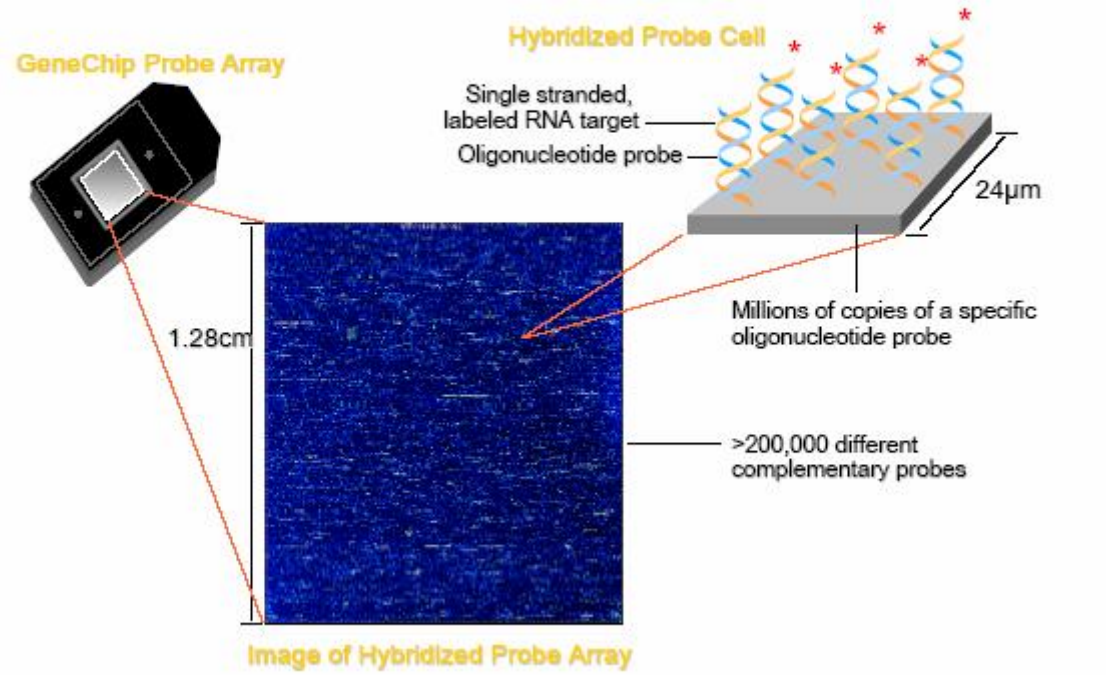
Type 2 - dual channel (cheaper)

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 know location

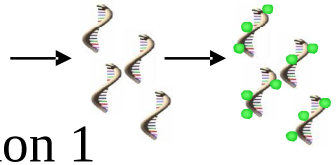
Oligonucleotide chips



Compliments of D. Gerhold

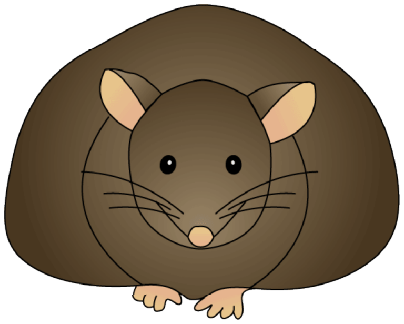
How do microarrays work

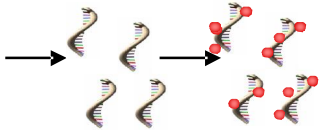
condition 1



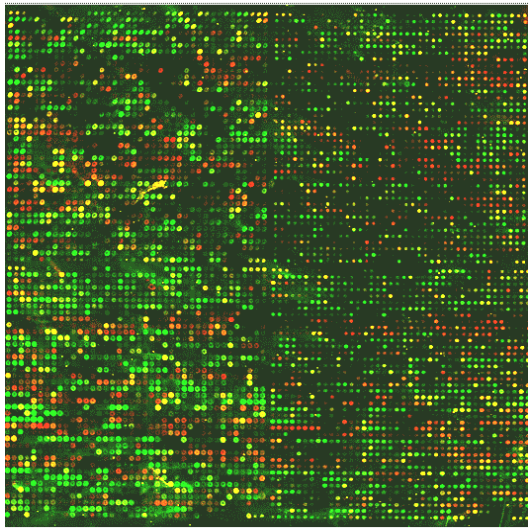
mRNA cDNA

hybridise to microarray



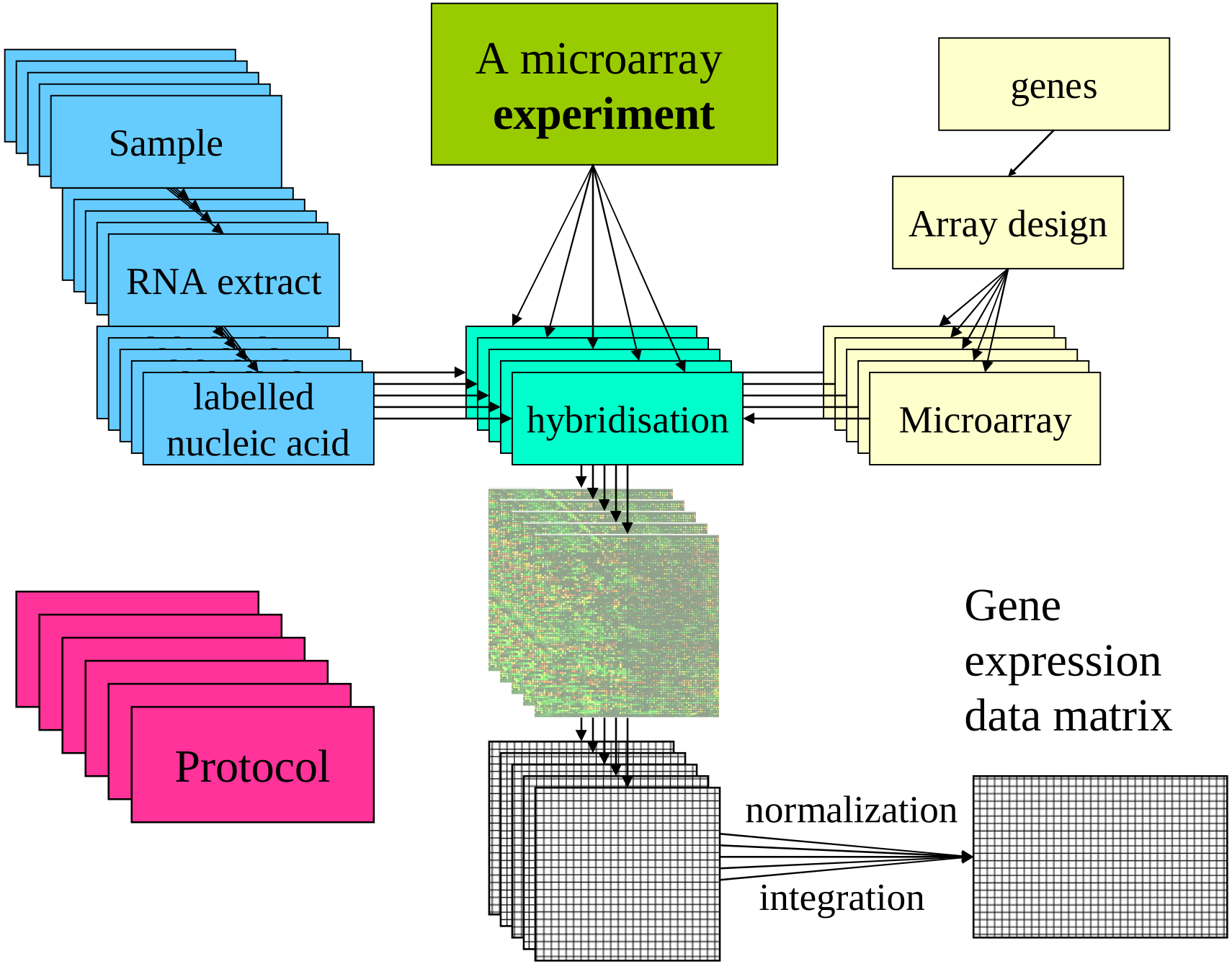


condition 2

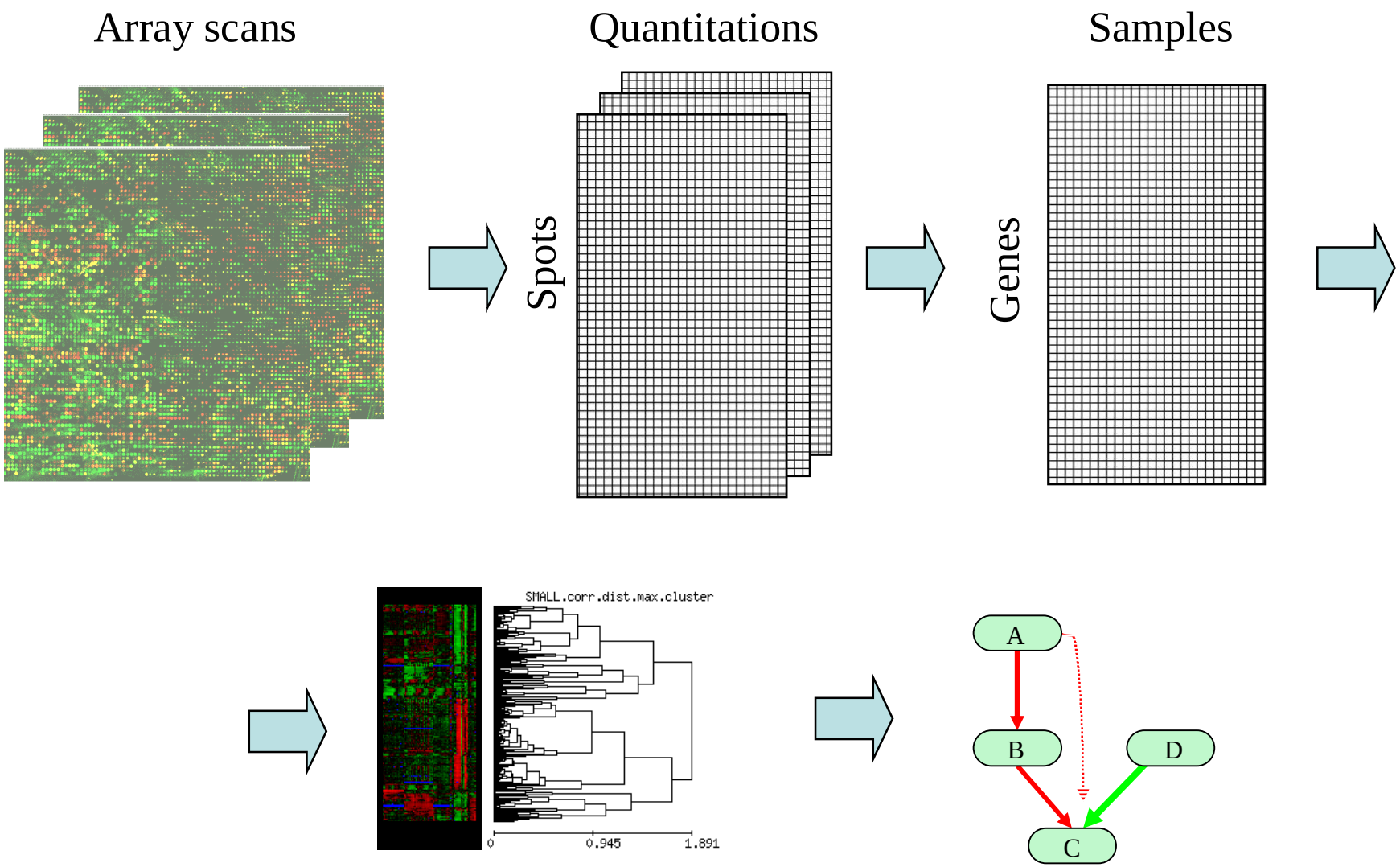


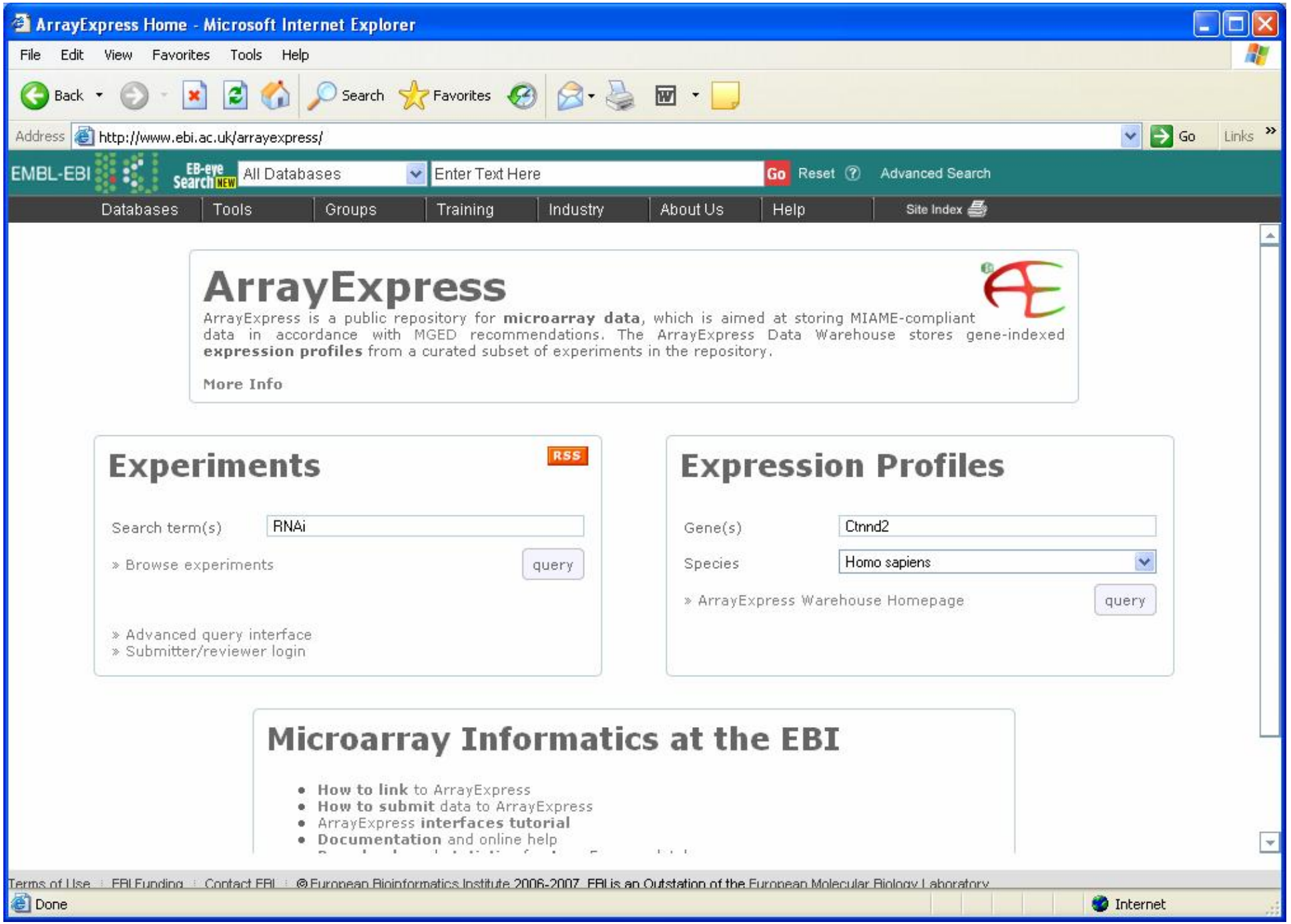
A microarray experiment

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


A microarray
experiment
genes
Array design
Sample
RNA extract
labelled
nucleic acid
hybridisation
Microarray
Protocol
Gene
expression
data matrix
normalization
integration

Steps in microarray data processing




ArrayExpress Home - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ebi.ac.uk/arrayexpress/
Go Links
EMBL-EBI
EB-eye Search NEW
All Databases
Enter Text Here
Go Reset Advanced Search
Databases Tools Groups Training Industry About Us Help Site Index
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.
More Info
Experiments
RSS
Search term(s)
RNAi
» Browse experiments
query
» Advanced query interface
» Submitter/reviewer login
Expression Profiles
Gene(s)
Ctnnd2
Species
Homo sapiens
» ArrayExpress Warehouse Homepage
query
Microarray Informatics at the EBI
How to link to ArrayExpress
How to submit data to ArrayExpress
ArrayExpress interfaces tutorial
Documentation and online help
Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2006-2007. EBI is an Outstation of the European Molecular Biology Laboratory
Done
Internet