Development of Complex Curricula for Molecular Bionics and Infobionics Programs within a consortial* framework**

Consortium leader

PETER PAZMANY CATHOLIC UNIVERSITY

Consortium members

SEMMELWEIS UNIVERSITY, DIALOG CAMPUS PUBLISHER

The Project has been realised with the support of the European Union and has been co-financed by the European Social Fund ***

**Molekuláris bionika és Infobionika Szakok tananyagának komplex fejlesztése konzorciumi keretben

***A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

INTRODUCTION TO BIOINFORMATICS

(BEVEZETÉS A BIOINFORMATIKÁBA)

CHAPTER 8

Analysis of gene expression

(Génexpressziók elemzése)

Péter Gál

Investing in your future
New Hungary Development Plan

Gene and protein expression data

DNA chip or DNA microarray

There are high-throughput methods in the gene expression analysis that are suitable for detection of the expression of many genes simultaneously.

One of the most widely used methods is the DNA chip or DNA microarray.

It can generate tens of thousand data point in a single experiment.

Besides gene expression analysis we can use the microarray technology to prove the presence or absence certain genes or mutants in a sample (see SNP analysis later).

A DNA microarray or DNA chip contains a series of DNA samples arranged in a grid pattern on a miniature support such as glass chip.

Signal generation is based on the hybridization (i.e. complementary base pairing) of the immobilized DNA molecules with a (usually complex) probe (DNA or RNA).

The probe (typically cDNA) is usually labeled with a fluorophore.

After hybridization the chip is illuminated with laser light and the intensity of the fluorescence light is detected.

If two probes are used with different fluorescent labels, the gene expression differences between the two samples can be detected on the same chip.

There are two major technologies by which DNA microarrays can be prepared:

1.) On-chip photolithographic DNA synthesis. (GeneChip, Affymetrix Inc., Steve Fodor)

An alternative method of on-chip DNA synthesis is the inkjet printing process developed by Agilent. This *in situ* synthesis process prints 60-mer length oligonucleotide probes.

2.) Spotted DNA array. DNA fragment (20-200 bps) are placed on the solid surface using robotic devices that accurately deposit nanoliter quantities of DNA solution.

Both technologies can be used to monitor differential gene expression between samples (e.g. healthy cell vs. disease cell; embryonic cell vs. adult cell; resting cell vs. stimulated cell; etc.)

Spotted DNA arrays: is cheaper and more popular than the other method.

Up to 5000 different DNA samples (features) can be spotted on one square cm.

The sample DNA can be double stranded genomic DNA or more frequently cDNA up to about 400 bp in length. They must be denatured prior to hybridization.

Spotted microarrays can be purchased form a number of companies or can be produced in the laboratory using a suitable robotic facility.

The GeneChips can be purchased exclusively from Affymetrix Inc. California.

The density of the chip can reach up to10^6 samples (single stranded oligonucleotid DNA) per square cm. Note that the number of the different DNA features in a chip is much bigger than the number of the genes in a typical eukaryotic genome.

The size of a typical feature is about 25nt and the number of the DNA molecules in a single spot is about 10^9.

Each gene is represented by 20 different features (non-overlapping sequences) and 20 mismatching controls are included to normalize background hybridization.

Both types of chip can be used for differential gene expression analysis.

In the case of the spotted array the RNA samples from two different cells or tissues are converted into cDNA and labeled by different fluorescent probes. The most frequently used dyes are Cy3 and Cy5. Cy3 fluoresces bright red and Cy5 fluoresces bright green. During hybridization we mix the two cDNA populations. DNA molecules that can be found only in one sample population will emit red or green light on the chip. DNA molecules that can be found in both populations will hybridize to the same spot which consequently will appear yellow.

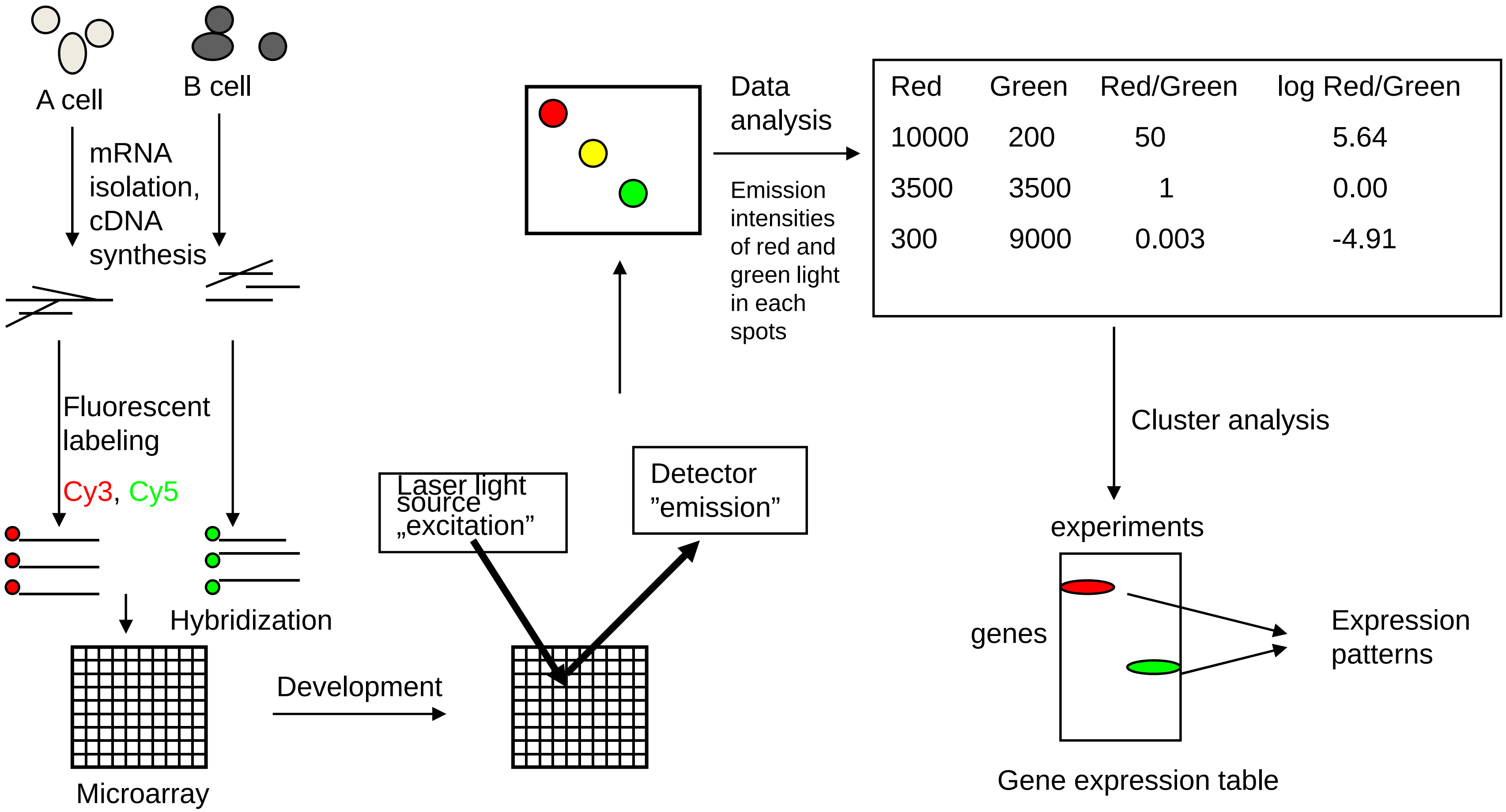
The raw data of a microarray experiment are arranged into a matrix called the gene expression table.

In this matrix the rows correspond to different genes and the columns to different experiments.

The gene expression table indicates relative expression levels.

From the gene expression patterns we can deduce the possible coordinated expression of different genes participating in the same biological process (e.g. the enzymes of a metabolitic route, or the components of multiprotein/multienzyme complexes).

A typical microarray experiment



Red	Green	Red/Green	log Red/Green
10000	200	50	5.64
3500	3500	1	0.00
300	9000	0.003	-4.91

In the case of the Affymetrix GeneChips the dual labeling technique is not used.

Instead, two identical chips are used for the two uniformly labeled cDNA probes.

The signal intensities measured on the corresponding positions of the two chips are compared.

It should be noted that the emergence of the next generation sequencing (NGS) techniques seriously challenged the DNA array method. Using the NGS we can directly sequence and quantify the DNA or RNA composition in a sample at an affordable price.

Application of DNA microarrays

1.) Analysing of individual genotypes of tissues or organisms

We can correlate the genotype of an organism with a phenotypic feature, e.g. susceptibility to certain disease.

2.) Investigation of cellular states and developmental processes

We can follow the expression of different genes in different developmental status of an organism. We can also monitor the gene expression profiles of a bacteria under different growth conditions (e.g. aerobic vs. anaerobic).

3.) Diagnosis of genetic diseases

Detection of mutations that are involved in the pathogenesis of inheritable diseases.

4.) Genetic warning signs

Many diseases are not determined by a single mutation, but a number of mutations (SNPs) in different genes make the individual susceptible to certain diseases (e.g. different types of cancer). A person aware of enhanced risk can adjust his lifestile to avoid the development of the disease.

5.) Specialized diagnosis and treatment of disease

Certain types of leukemia cannot be distinguished by morphological signs of the leukocytes, bur their gene expression patterns are different. The microarray analysis can help in selecting the optimal treatment.

6.) Personalized medicine

Genetic factors of the individual determine whether a certain drug effectively treat a certain disease or it is ineffective or it has serious side effects. Using microarray data we can select the optimal drug regimen for each individual.

7.) Target selection for drug development

During disease the expression level of certain protein can change considerably. These changes can be detected by differential gene expression analysis. The proteins identified by this way could serve as target molecules for subsequent drug development process.

8.) Pathogen resistance

Comparative genomics of bacterial strains in order to find the genes that confer antibiotic resistance.

Note that DNA microarray measures the gene expression at mRNA level. mRNA level however do not accurately reflect protein levels. Not all mRNA molecules will be translated into protein molecules. Moreover, reverse transcription and PCR amplification can change the original composition of the transcriptome.

Taken together we can say that microarray technology is suitable for semi quantitative measurement of gene expression but it inadequate for determination of absolute expression levels.

Microarray databases

The microarray databases contain microarray gene expression data in a suitable form for analysis and interpretation. MIAME (Minimum Information About a Microarray Experiment) is a standard describing the content and format of the information to be recorded in the experiment , and deposited.

Examples of useful, peer reviewed, publicly-available microarray databases are:

Gene Expression Omnibus (GEO) from NCBI http://www.ncbi.nlm.nih.gov/geo/

ArrayExpress from EBI

http://www.ebi.ac.uk/arrayexpress/

Expression databases

The most straightforward way to measure gene expression on the transcription level is to sequence the mRNA population of a given tissue or organism at a given time.

Now this approach could work by using next generation sequencing methods.

The more traditional solution to this problem however is the construction of EST (expressed sequence tag) libraries.

ESTs are generated by rapid, single-path sequencing of random clones from cDNA libraries. An EST represents a short, usually terminal, fragment (200-300bp) of the original cDNA but it unambiguously identifies the gene.

Since ESTs are derived from cDNAs the composition of an EST library depends on the tissue of origin, state of development, condition of growth, etc.

If EST data are available , the abundance of different mRNAs can be estimated by determining the representation of each sequence in the database.

EST data can be very useful to identify genes in genomic DNA, to map their position within the genome and to prepare large clone sets for spotted DNA microarray.

There are many databases containing ESTs (e.g. NCBI Gene Bank dbEST).

At present (August 2010) dbEST collection contains almost 67 million entries.There are 13 species that are represented by more than 1 million ESTs, led by:

Species	# of entries
Homo sapiens (human)	8,300,000
Mus musculus (mouse)	4,850,000
Zea mays (maize)	2,020,000
Sus scrofa (pig)	1,620,000
Bos taurus (cattle)	1,600,000
Arabidopsis thaliana (thale cress)	1,530,000
Danio rerio (zebrafish)	1,480,000
Glycine max (soybean)	1,460,000

Serial analysis of gene expression (SAGE) is another method of gene expression analysis.

In this technique short cDNA sequence tags are generated, however they are not sequenced individually.

The short tags are ligated together to form longer DNA molecules (concatemers) which are sequenced.

The occurrence of each tag is counted.

Since 50-100 tags can be counted for each sequencing reaction, this technique is more efficient, than sequencing the individual ESTs.

Single-nucleotide polymorphism (SNP)

A single nucleotide polymorphism (SNP) (pronounced „snip”) is a small genetic change or variation between individuals limited to a single base in the DNA sequence (point mutation).

A SNP could be any type of point mutation i.e. substitution, insertion or deletion.

SNPs can be found in every segment of the genome, including the protein coding exons, the non-coding intron sequences, the regulator elements and the intergenic regions.

A SNP in the non-coding region of DNA can influence the phenotype of an individual organism.

The average occurrence of SNPs in the human genome is 1 SNP per 5000 base pairs.

Every individual has its own SNP pattern that is made up of many different genetic variations.

Although SNPs can be found everywhere in the genome the density of SNPs is higher in the genes than in other parts of the genome.

A SNP in the coding region can change the protein sequence, e.g. missense or frame shift mutations.

Sikle-cell anaemia is a disease caused by a SNP: A→T mutation in the β-blobin gene resulting in Glu→Val change.

Application of SNPs:

1.) SNPs and diseases

Only a few SNPs are directly responsible for disease development, however many SNPs are linked to diseases because they are located near a gene found to be associated with a certain disease.

The SNP pattern of an individual can predict its propensity for certain diseases.

Association study: comparison of the SNP pattern of healthy individuals with that of sick individuals in order to identify the SNPs associated with a certain disease.

2.) SNPs and personalized medicine

SNPs may be associated with the absorbance and clearance of therapeutic agents.

A drug which is useful for one patient can be ineffective to the other or it even can cause severe unwanted immune reaction (drug allergy).

Analysing a patient's SNP profile can help in the selection of the most appropriate drug for an individual in advance of treatment.

Development of personalized medicine would allow pharmaceutical companies to bring new drugs to market that are effective only in patients with certain genetic background (SNP pattern).