Special Topics on Genetics

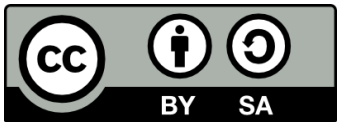
Section 7: Applications of functional genomics

Triantafyllidis A.
School of Biology

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

License

- The offered educational material is subject to Creative Commons licensing.
- For educational material, such as images, that is subject to other form of licensing, the license is explicitly referred to within the presentation.



Funding

- The offered educational material has been developed as part of the educational work of the Instructor.
- The project "Open Academic Courses at Aristotle University of Thessaloniki" has financially supported only the reorganization of the educational material.
- The project is implemented under the Operational Program "Education and Lifelong Learning" and is co-funded by the European Union (European Social Fund) and national resources.

Section Contents

- mRNA analysis in genomics
- Analysis of SNP polymorphisms
- Others applications of NGS machines
- The use of high resolution platforms in proteomics

DNA sequencing (and RNA sequencing)

• The automated sequencer is used for the sequencing of genes (or ESTs or RNA sequences).

• Most studies concerning genome organisms (excluding model species), are mainly descriptive ... they are an essential step towards a better understanding of the genome and the transcriptome of the organisms.

• Consequently ... focus will be on specific genes and their polymorphisms or the creation of microarrays for gene expression analysis.

mRNA analysis in genomics

Concerns

- Yeast: About 70% of the genome codes for proteins and there is one gene per 2 Kb.
- Human: only 5% of the DNA encodes and there is one gene per 60-80 Kb.
- Many genes are expressed at levels less than 50 molecules / cell.
- A typical mammalian cell has 300000 derivatives of 20000 different mRNAs. For 90% of these their amount is less than 100 copies / cell.
- DNA arrays cannot detect mRNAs less than 50 molecules per cell. They only enable a semi-quantitative-comparative analysis of gene expression levels.

mRNA analysis in genomics 1

Conventional ABI Sequencing: If first generation machines are used for the complete sequencing of a cDNA library, many of the cDNAs correspond to the same gene copies.

More than 4 million ESTs from various human cell types have been sequenced, but thousands of ESTs are derived from the same gene that encodes for human albumin.

It is not a cost-effective solution

mRNA analysis in genomics 2

SAGE-serial analysis of gene expression

- The transcripts of different genes differ in their 3‘ untranslated end.
- Only fragments of 15bp are synthesized at the 3’ ends of mRNA and they are joined in a 600 bp fragment which is then sequenced.
- Each piece of 600 bp corresponds to 40 genes fragments (15 x 40).
- The sequence analysis of thousands of such fragments allows a quantitative analysis of mRNAs, which are present in the analysed cell.

Schematic illustration of the method (Fig. 2):

http://www.scq.ubc.ca/painless-gene-expression-profiling-sage-serial-analysis-of-gene-expression/

mRNA analysis in genomics 3

MPSS, massive parallel signature sequence

This technique amplifies specific flags regions of 20 bp using PCR of up to 1 million clones from a cDNA library

- The fragments act as an identification flag of different mRNA molecules,
- Are attached to nylon beads and
- placed in the automatic sequencer which simultaneously sequences them, using specific fluorescent probes.

☞ The sensitivity of the technique is great, and not very expensive.

Schematic illustration of the method:
http://zlgc.seu.edu.cn/jpkc/2010jpkc/jykc2/Content/jxzy/genetics/chapt10/art_library/color_art_library/10_28.jpg

mRNA analysis in genomics 4 (1/7)

- The use of NGS analyses allows full transcriptome analysis – i.e. which genes are expressed in specific tissues, life stage or species.

mRNA analysis in genomics 4 (2/7)

➢ RNA-Seq: a revolutionary tool for transcriptomics

Wang Z., Gerstein M., Snyder M., Nat Rev Genet. 2009 January ; 10(1): 57–63
http://www.nature.com/nrg/journal/v10/n1/full/nrg2484.html

Advantages of these applications

- they are absolutely quantitative
- they can discover isoforms of a gene or even different alleles
- they are relatively cheaper
- prior knowledge of the genome of an organism is not needed
- they can discover / identify unknown genes

mRNA analysis in genomics 4 (3/7)

Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing

Vera J.C. *et al*. , Molecular Ecology (2008) 17, 1636–1647

Glanville fritillary butterfly: *Melitaea cinxia*
Sample = genetically diverse group larvae, nymphs, mature individuals
454 sequencing of the transcriptome produced 608053 ESTs (~110 bp mean size, more than 9000 genes)
Microarrays were designed and when they were tested, significant biological differences between individuals were found regarding gene expression

mRNA analysis in genomics 4 (4/7)

Transcriptome and genome sequencing uncovers functional variation in humans

The first completed analysis of the correlation of the expression of total mRNA and microRNA from lymphoid cells in 462 individuals from the 1000 Genomes project. Significant differences between populations and a great heterogeneity in gene expression was discovered.

Nature 501, 506–511 (26 September 2013)

mRNA analysis in genomics 4 (5/7)

Diversity and dynamics of the *Drosophila* transcriptome

Nature 512, 393–399 (28 August 2014)

29 tissues were analyzed, 24 cell lines and 21 different individuals that were placed in difficult environmental conditions. 300.000 different transcripts for 17564 genes (14692 of these are protein) were produced.

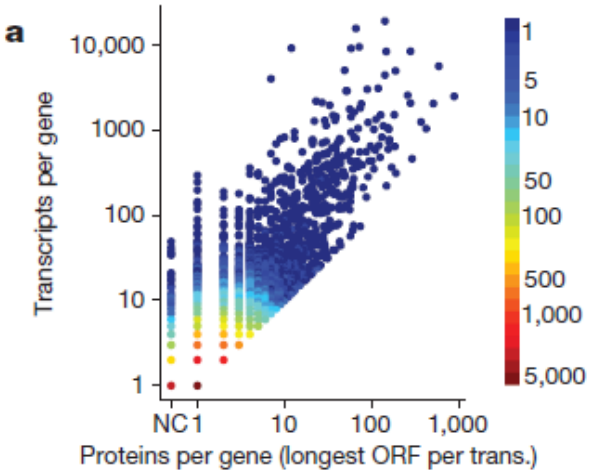
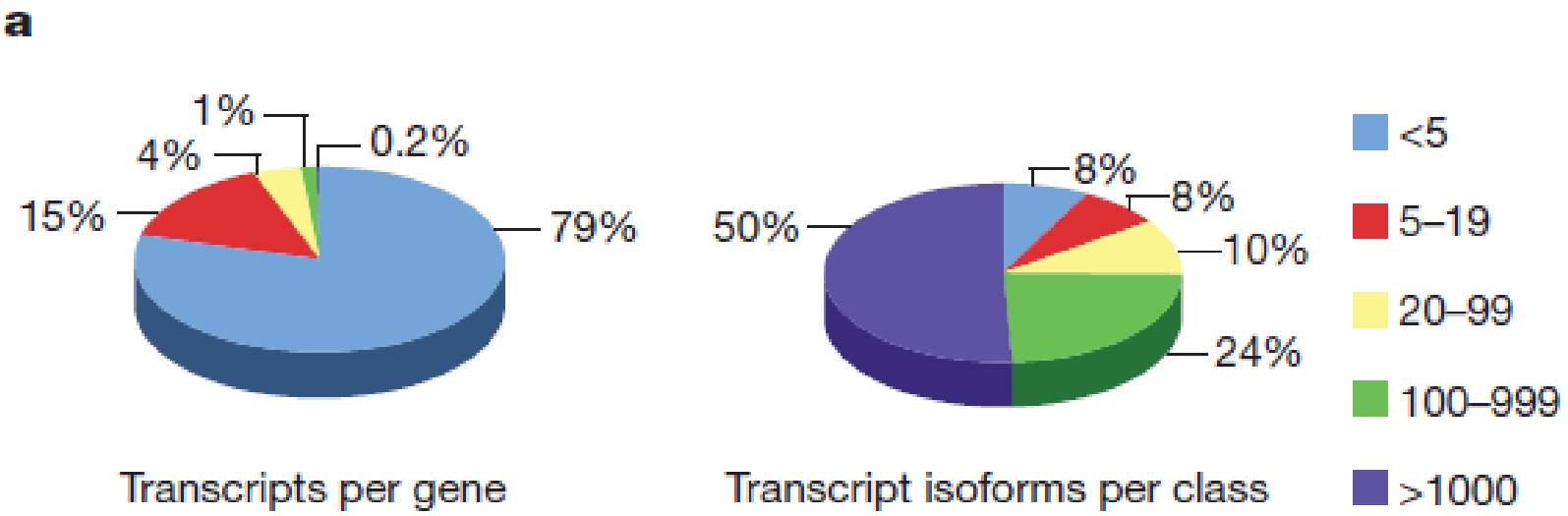


Figure 1: Correlation between the number of genes (colour) with the number of proteins (X-axis) and the number of transcripts (Y axis). 90% of the genes produce a maximum of 10 transcripts and 5 protein isoforms. Only 1% of the genes have both complex splicing pattern, polyadenylation and use of different promoters. The Dscam and para genes produce up to 10000 unique proteins and they are not included in the chart.

mRNA analysis in genomics 4 (6/7)

Diversity and dynamics of the *Drosophila* transcriptome
Nature 512, 393–399 (28 August 2014)

Figure 2: A small number of genes (47, 0.2%) of the brain was shown to produce thousands of different transcripts due to different promoters and different splicing.



mRNA analysis in genomics 4 (7/7)

Diversity and dynamics of the *Drosophila* transcriptome

57 genes (5259 transcripts) were expressed only in extreme conditions.

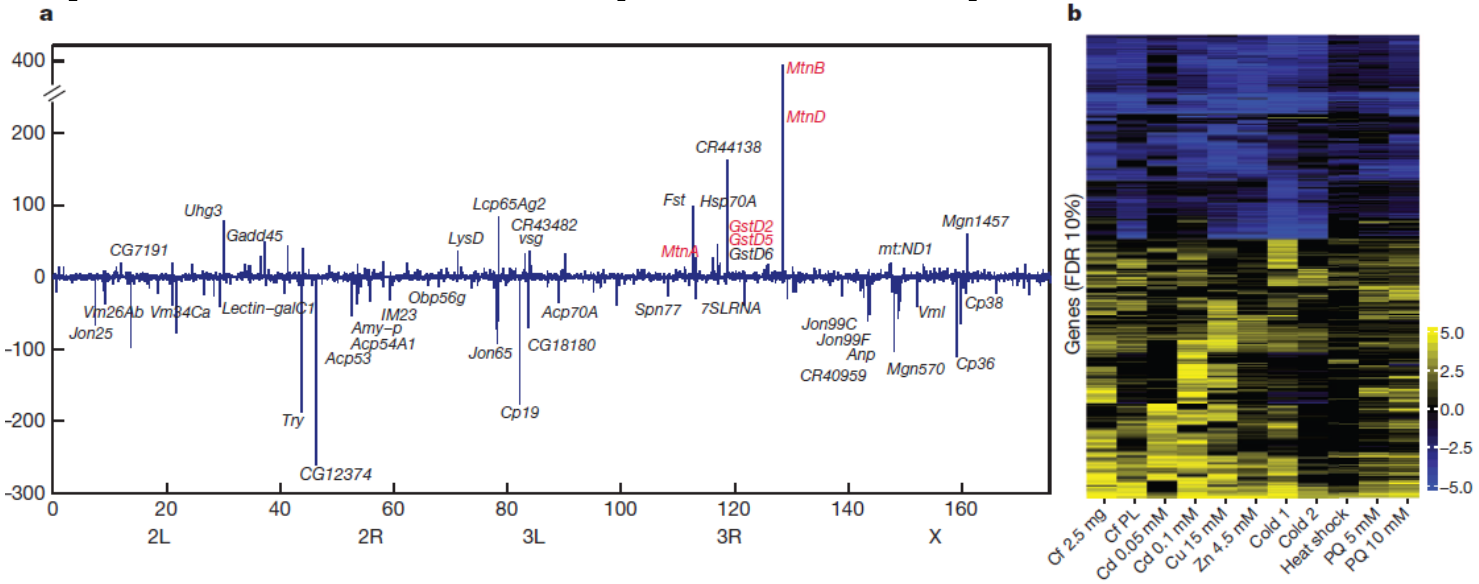


Figure 3: Environmental effects on Drosophila transcriptome. Adults were treated with caffeine (Cf), Cd, Cu, Zn, cold, heat or paraquat (PQ). A) A genome-wide map of genes that are up- or downregulated as a function of Cd treatment. Labelled genes are those that showed > 20% change in expression. Genes highlighted in red are those identified in larvae. B) The total number of genes with 20-fold change in expression.

Analysis of SNP polymorphisms (1/9)

The analysis can be made with oligonucleotide arrays or mass spectrometry (GoldenGate-Illumina).

The optical fiber array analyzes → 1.000.000 SNPs per day

http://res.illumina.com/documents/products/workflows/workflow_goldengate_assay.pdf

Analysis of SNP polymorphisms (2/9)

- Identification of thousands of new genetic markers (such as SNPs and INDELs), at broad levels and detection of polymorphism throughout the genome.

- Conclusions on the evolutionary history of the species, population increases or bottlenecks, migration and admixture of populations.

- It is now possible to study differences in genes between populations with different phenotypes and detect mutations with adaptive advantage and selection phenomena.

Analysis of SNP polymorphisms(3/9)

Chips for studies of diseases

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Klein et al. Science 2005

Test for the existence of linkage in > 100000 SNPs with the disease (Test of 50000 SNPs per chip).
2 SNPs were identified
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1512523

Analysis of SNP polymorphisms(4/9)

Genetics of rheumatoid arthritis contributes to biology and drug discovery

Genome-wide association study : meta-analysis >100.000 subjects of European and Asian ancestries (29880 rheumatoid arthritis cases and 73758 controls), by evaluating ~10 million SNPs. 42 novel risk loci were discovered, bringing the total to 101. An *in silico* method was used for the identification of 98 candidate genes, at these 101 risk loci. These genes are the targets of approved therapies for rheumatoid arthritis.
http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12873.html

Analysis of SNP polymorphisms(5/9)

Biological insights from 108 schizophrenia-associated genetic loci
GWAS in 36.989 patients and 113075 controls. 108 (66% novel) loci were identified. Associations were enriched among genes expressed in brain, in tissues that have important roles in immunity, use of over 1000 genomic data, >500 K SNPs.
http://www.nature.com/nature/journal/v511/n7510/full/nature13595.html

Nature 511, 421–427 (24 July 2014)

Analysis of SNP polymorphisms(6/9)

Discovery and saturation analysis of cancer genes across 21 tumour types

Although a few cancer genes are mutated in a high proportion of tumours of a given type (>20%), most are mutated at intermediate frequencies (2–20%). Somatic point mutations in exome sequences from 4742 human cancers and their matched normal-tissue samples across 21 cancer types were analysed. It was found that **GWAS** analysis can identify nearly all known cancer genes in these tumour types. Also, 33 genes that were not previously known to be significantly mutated in cancer were identified.

http://www.nature.com/nature/journal/v505/n7484/full/nature12912.html

Analysis of SNP polymorphisms(7/9)

Chips for studies of diseases – The future

- New arrays analyze 900000 SNPs and 950000 Copy Number Variations (CNVs) !!!
- Creation of specific arrays per disease (cancer, arteriosclerosis, arthritis etc with ~ 500 genes with minimum requirements for initial RNA quantities (20 cells could be enough)
- Arrays for microRNAs related to cancer
- Reproducibility still remains a problem

Analysis of SNP polymorphisms(8/9)

Genome-Wide survey of SNP variation uncovers the genetic structure of cattle breeds

The Bovine HapMap Consortium, http://www.sciencemag.org/content/324/5926/528.full

- Study of 37470 SNPs in 497 bovines from 19 geographically and genetically distinct breeds
- The data show, that there has been a recent decline in population because of possible bottleneck phenomena

Genome-wide SNP and haplotype analysis reveal a rich history underlying dog domestication *Nature* **464**, 898-902(8 April 2010)
http://www.nature.com/nature/journal/v464/n7290/fig_tab/nature08837_F1.html

- Study of 48000 SNPs in 912 dogs from 85 breeds
- There was very good correlation between genetic and phenotypic data

Analysis of SNP polymorphisms(9/9)

Applications on dog breeds – characteristics

Characteristics such as the spiral tail, the droopy ears and the sociability are associated with specific SNPs identified in specific sites on chromosomes of the dog (*Vaysse et al., 2011*) http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002316

Others applications of NGS machines (1/7)

- Study of epigenetic phenomena, ie. changes in characteristics which are not related to direct heritable mutations of DNA, but with other genetic modifications (such as methylation of DNA, or modification of histone proteins).
- With new NGS methodologies (BS-seq and Me-DIP) regions of DNA which are methylated are identified and methylation maps are generated.
- Antibodies for specific proteins can be used, which package the DNA or are attached on this as transcription factors, and subsequently these regions are sequenced exclusively (CHIP-seq).

http://csls-text3.c.u-tokyo.ac.jp/large_fig/fig25_08.html

Others applications of NGS machines (2/7)

Study of epigenetic phenomena

Charting a dynamic DNA methylation landscape of the human genome Ziller *et al*. 2013 Nature 500

- Study of CpGs methylation
- Analysis of 42 sequencing datasets of whole genomes in 30 different human cells or tissues
- Dynamic regulation only in 21.8% of the autosomal CpGs

http://www.nature.com/nature/journal/v500/n7463/full/nature12433.html

Others applications of NGS machines (3/7)

ENCODE/modENCODE

All the above NGS methodologies have been used in the context of ENCODE and modENCODE programs, in which an attempt is made to identify all the regulatory elements related to gene expression based on analyses of the transcriptome, the chromatin organization and the binding sites of thousands of regulatory elements.

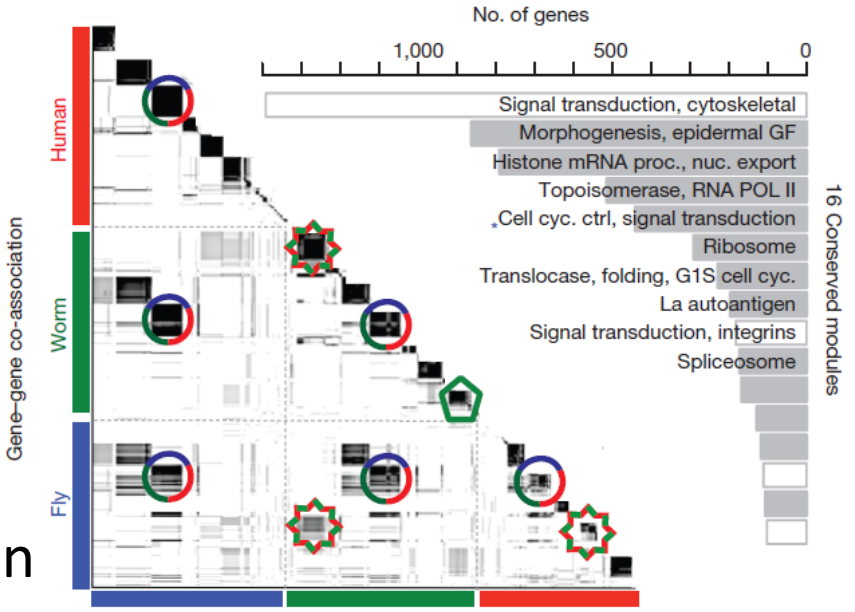
1600 new datasets, 3300 in total
http://www.nature.com/nature/journal/v512/n7515/full/512374a.html

Others applications of NGS machines (4/7)

Figure 4: ENCODE/modENCODE

Transcriptome analysis (Gerstein et al. 2014, RNA seq data) of the three species, revealed the existence of 16 groups (intraspecific or even inter-specific) of genes that exhibit correlation in their expression and are related to development.

A model was created that can predict the gene expression levels using data on chromatin structure in the promoter regions.



Others applications of NGS machines (5/7)

Barcoding-metagenomics – Biodiversity analysis

Second-generation environmental sequencing unmasks marine metazoan biodiversity Fonseca *et al*. 2010

- Metagenetics analysis
- Sequencing with 454 Roche FLX
- Assessment of meiobenthic richness
- Use of 18S nuclear small subunit(nSSU)
- 305702 sequences were produced
- The majority of the samples belong to the Nematoda

http://www.nature.com/ncomms/journal/v1/n7/fig_tab/ncomms1095_F2.html

Others applications of NGS machines (6/7)

Biodiversity analysis

Applications of next generation sequencing in molecular ecology of non-model organisms Ekblom and Galindo 2011

- Sequencing of genomes of organisms that are not models
- Application of NGS technologies in ecological and population genetics studies
- The workflow starting from the sample collection through to the application of NGS technologies in molecular ecology is shown at the link below:

http://www.nature.com/hdy/journal/v107/n1/fig_tab/hdy2010152f1.html

Others applications of NGS machines (7/7)

The use of an NGS machine, costs at least 10.000-20.000 euro, an amount that is not negligible.

Before choosing the research approach researchers have to take into consideration the costs and the data analysis capabilities.

The data volume generated by new technologies is a real challenge in terms of their storage and analysis. Also which genomic platform will be used depends on the biological question and the specifications of each machine.

We therefore conclude that the most basic prerequisite for the success of each experiment is the right design.

The use of high resolution platforms in proteomics (1/11)

Proteomics uses high resolution technologies for the:

- Description of molecular and biochemical pathways in a cell.
- Recording and understanding of the interactions of all proteins of an organism in certain cell phases and conditions.

Questions about the function of proteins and the corresponding genes are multiplying. We do not know the function of thousands of genes in *E. coli*, *S. cerevisiae* and human.

The use of high resolution platforms in proteomics (2/11)

Proteome analysis constitutes a greater challenge than genome analysis for two reasons:

- The expression levels of a protein in a cell present an enormous range: from one molecule up to 10^6 copies per cell. Since there is nothing similar to PCR for proteins, protein analysis tools should be able to operate on differences of six classes size.
- The proteins (in complex mixtures) have specific characteristics that need to be analyzed.

The use of high resolution platforms in proteomics (3/11)

The particular characteristics of the proteins to be analyzed are as follows:

- the gene which codes for the protein,
- its expression levels in various cell types (in different stages of development, or in response to changes in physiology)
- Possible modifications (phosphorylation, glycosylation…),
- interactions with other proteins-macromolecules-micromolecules,
- detection within the cell,
- activation ways,
- half time life,
- three dimensional structure,
- structure-function relationship.

The use of high resolution platforms in proteomics (4/11)

The recognition-identification of proteins in complex mixtures (1)

When the genome of an organism is fully sequenced:

- Cells of a specific type are isolated.
- Followed by cell lysis.
- Proteins are isolated and fragmented with trypsin into smaller peptides.
- Peptides are fractionated on reverse phase columns based on their hydrophobicity and their molecular weight.
- Groups of peptides from the column are promoted and analyzed by mass spectrometer.
- The peptides can be cut into smaller ones and further information obtained on the mass of sub-peptides from a MS / MS spectrometer.

The use of high resolution platforms in proteomics (5/11)

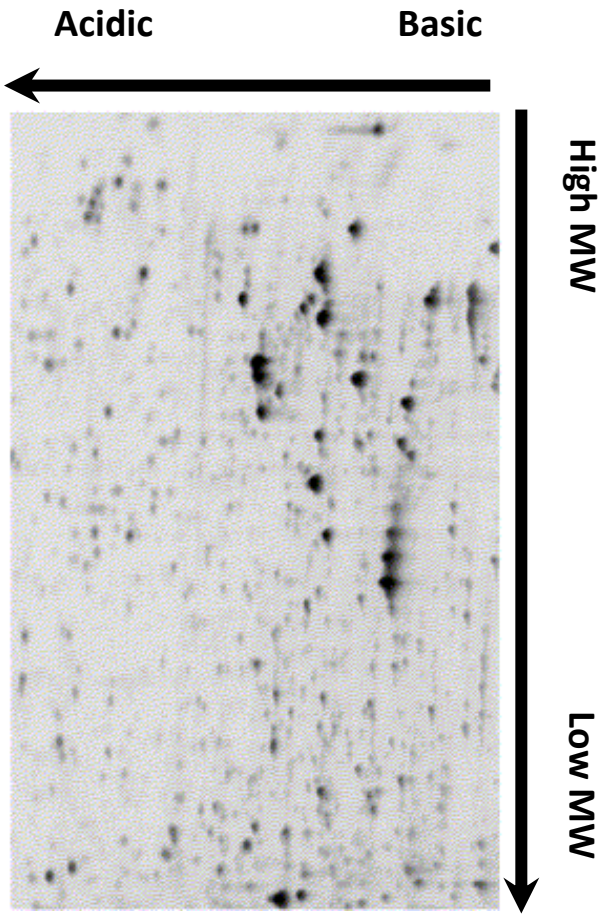
The recognition-identification of proteins in complex mixtures (2)

2-dimensional gel electrophoresis, 2DGE

- Proteins are initially separated in the first dimension according to their isoelectric point and then according to their molecular weight.
- Gels are stained with Coomassie Blue and the proteins appear as dots at specific positions - These positions may be from 200 to 10000.
- The proteins of interest, are eluted from the gel, cut with trypsin and then their analysis by mass spectrometry follows.

The use of high resolution platforms in proteomics (6/11)

- **Figure 5: 2-dimensional gel electrophoresis**



Disadvantage: low sensitivity for proteins in low C

The use of high resolution platforms in proteomics (7/11)

The recognition-identification of proteins in complex mixtures (3)

HPLC-high performance liquid chromatography

- Separation of the liquid phase (proteins) in a column under high pressure
- Advantages of HPLC

  1) speed

  2) sensitivity &

  3) ease of automation and connection with MS

http://www.lamondlab.com/MSResource/separations/affinityIsolation.php

The use of high resolution platforms in proteomics (8/11)

- In each yeast cell there are theoretically 6000 proteins and 35.000 peptides derived from digestion with trypsin.
- Most of these peptides have characteristic mass values deposited in databases.
- By comparing the values collected by the mass spectrometer with the database it is possible to identify hundreds of proteins.

The use of high resolution platforms in proteomics (9/11)

Proteomic analysis of *M. genitalium*

Which genes are expressed in the growth and stationary phase?

- Using 2DGE electrophoresis 427 proteins positions were found (from 480 genes) in cells at a growth phase
- Analysis of 201 proteins was followed, using digestion with trypsin, mass spectrometer analysis and comparison with known proteins
- 158 known proteins were discovered (33% of the proteome) and 17 unknown
- The remaining positions corresponded to broken fragments from other proteins, the same protein isoforms or posttranslational modifications
- In the stationary phase there was a 42% reduction in the number of proteins produced
- Some new proteins appeared, while others greatly reduced their levels

The main drawback is the sensitivity

The use of high resolution platforms in proteomics (10/11)

A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis Nature (2013).494, 266-270

- A strategy is followed based on the high performance peptide synthesis and mass spectrometry to generate a nearly complete reference map (97% of the predicted proteins) of the proteome of *Saccharomyces cerevisiae.*
- Protein measurements in 78 strains .
- There is a complex relationship between independent loci that affect the levels of proteins.

Variation and genetic control of protein abundance in humans Nature 499, 79–82 (04 July 2013)

- Mass spectrometry was used to determine the levels of 5953 genes of proteins in lymphoblastoid cell lines from 95 different individuals.

The use of high resolution platforms in proteomics (11/11)

FIRST DRAFT OF HUMAN PROTEOME

Analysis of 30 different tissues from adults and embryos.

Identification of 293.000 unique proteins, that correspond to 84% of identified genes (including proteins corresponding to ~ 2500 genes that had not been identified so far).

~50% of peptides were not databased.

Peptides from 2350 genes are house keeping and are expressed in all tissues and in high level s (>75% of total protein).

Identification of protein derivatives even from 149 supposed pseudogenes and from 9 ncRNAs.

Allows the revision of thousands of entries about new transcription start sites, and new coding exons.

Kim et al. 2014 Nature 509, 575-581,
http://www.nature.com/nature/journal/v509/n7502/abs/nature13302.html

Exercise

18.17 When cells are exposed to short-term high temperature (heat shock), they modify all transcribed genes within a protective response.
a. What steps will you follow to characterize the changes in yeast transcriptome after a heat shock?
b. Assume that the analyses of the transcriptome lead to the identification of a group of genes whose transcripts levels are increased after a thermal shock. How could you experimentally determine which of these genes are essential in a protective response after thermal shock?

Exercise - Solution

18.17

a. You can use either microarray, where you compare the expression in normal cells and cells after thermal shock or you can use an RNA-seq analysis with an NGS machine.

b. You need to make directed mutagenesis to cause a knock out on genes, one by one, and check which ones are needed for survival.

Internet links with information about sequences and genomes

http://www.ncbi.nlm.nih.gov/genome NCBI - Includes images of chromosomes, maps and genes with links for further information at NCBI
http://www.genome.gov National Human Genome Research Institute Includes information about the organisms in which sequencing of their genome is carried out
http://www.ensembl.org EBI/Sanger Center - Access to DNA and proteins sequences
http://genome.ucsc.edu University of Santa Krouz- includes information about the human genome
http://www.ncbi.nlm.nih.gov/Omim Mendelian Inheritance in human Include information about the human genes and diseases

http://www.youtube.com/playlist?list=PLF09DBAA3E24C5068

As for the future, your
task is not to foresee, but to enable it.”

Antoine de Saint-Exupéry
The Wisdom of the Sands

Note of use of third party works

Figure 1: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12962.html, by Brown JB et al., CC-BY-NC-SA-3.0, http://creativecommons.org/licenses/by-nc-sa/3.0/.
Figure 2: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12962.html, by Brown JB et al., CC-BY-NC-SA-3.0, http://creativecommons.org/licenses/by-nc-sa/3.0/.
Figure 3: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature12962.html, by Brown JB et al., CC-BY-NC-SA-3.0, http://creativecommons.org/licenses/by-nc-sa/3.0/.
Figure 4: http://www.nature.com/nature/journal/v512/n7515/full/nature13424.html, by Gerstein et al. 2014, CC-BY-NC-SA-3.0, http://creativecommons.org/licenses/by-nc-sa/3.0/.

Reference note

Licensing note

End of Section

Processing: Minoudi Styliani
Thessaloniki, Winter Semester 2014-2015

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Notes Preservation

Any reproduction or adaptation of the material should include:

- the Reference Note
- the Licence Note
- the Notes Preservation
- Note of use of third party works

accompanied with their hyperlinks.