HELSINKI UNIVERSITY OF TECHNOLOGY
LABORATORY OF COMPUTER AND INFORMATION SCIENCE

Introduction to Bioinformatics:

Chapter 11: Measuring Expression of Genome Information

Jarkko Salojärvi

Lecture slides by Samuel Kaski

Assignment:

Think of at least one question for which you want to get an answer during this lecture.

Plan

Let’s follow the book pretty closely (that is the idea of this course).

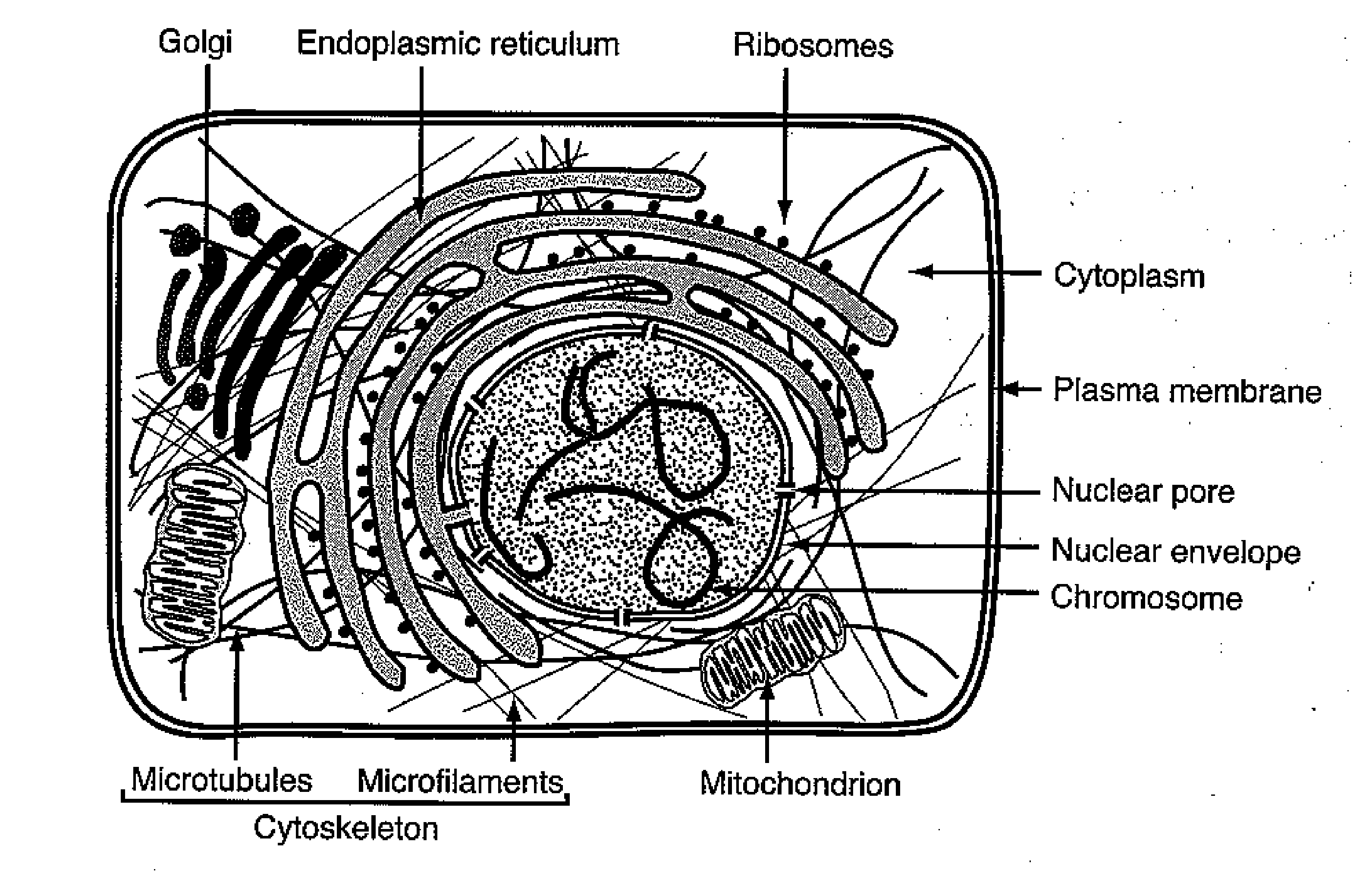
Task of the lectures: Quick overview, views of the lecturer, opportunities to ask/discuss.

Content:

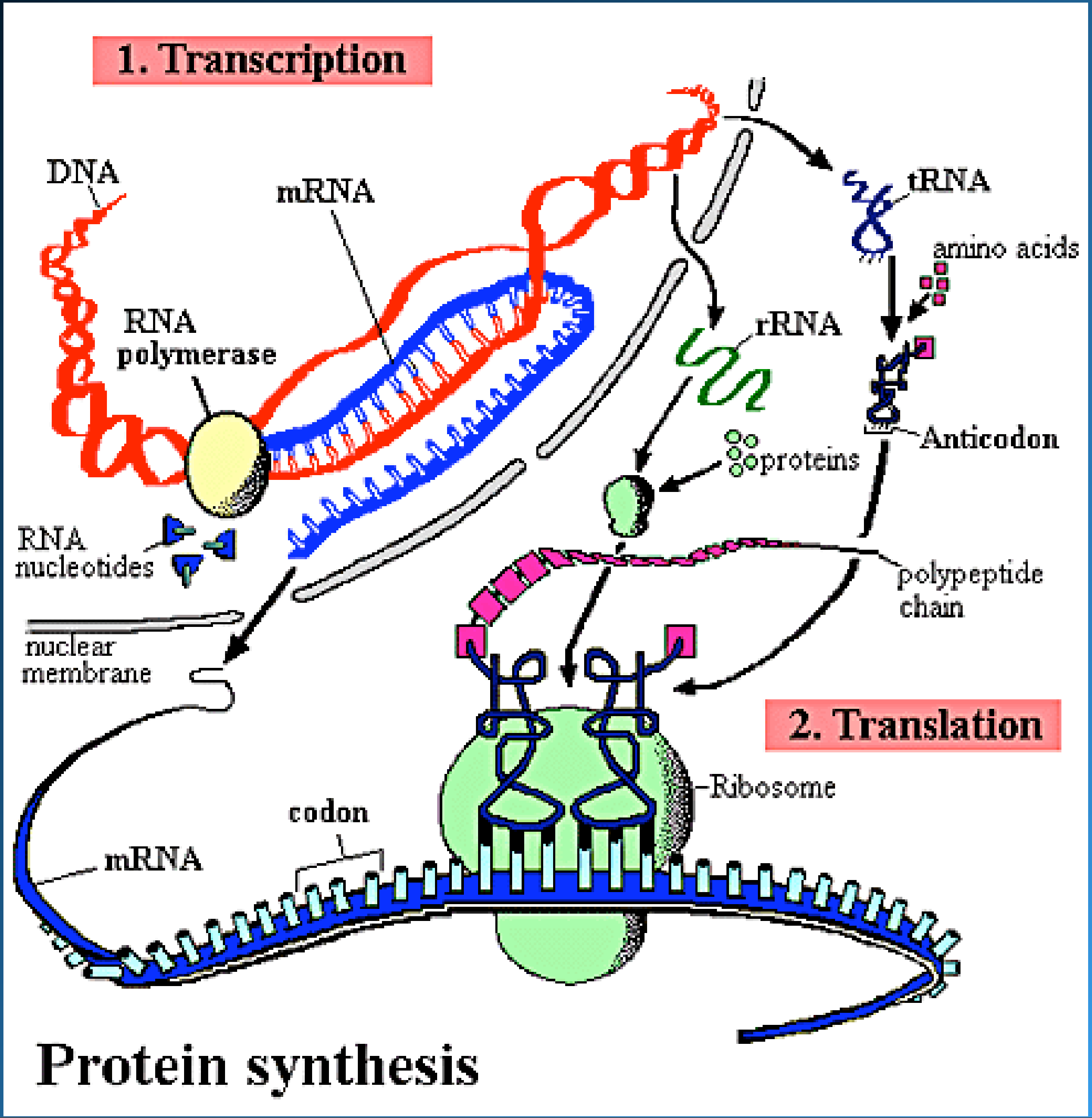
- very brief recap of necessary biology
- what can be measured
- how to measure (focus on a couple of examples)
- data analysis (only the very beginning)

Background

Recap: The cell



From genes to proteins



Dirty and noisy real-world measurements, yacc...

Why bother? Why not tackle only well-defined non-noisy problems?

Well, because the world is dirty and noisy... and besides, the more ill-defined a problem is the more interesting it is. Creativity needs to be used both in defining the problem and in solving it!

Noise requires some understanding of the measurement process, and a statistical approach.

Key: model the uncertainties = statistical modeling

What to measure?

Various “omics” have been coined for the various things to be measured.

From OMICS to systems biology. Vidal & Furlong, Nature Reviews Genetics, year xx.

Different levels of understanding cell function

- Genome (sequence)
- Transcription (gene activity); “functional genomics”
- Proteins
- Metabolism
- “Systems biology”
- Phenotype

Functional genomics level

Key questions:

- Which genes are active? Or more specifically:
- How are different conditions different?

Here condition = tissue, treatment, phase of cell cycle, different individual

Simplest answer is given by differential expression: Difference of transcription levels

Examples where differential expression is interesting

- During development: Pattern of activity in a set of genes regulates differentiation of tissue types during development of embryos
- Cancer vs normal tissue
- Effects of drugs
- Differences between organisms

Note: The *development* of differential expression patterns during time would often be the most interesting thing, but often it cannot be measured (for instance in cancer) or would be too costly.

Gene vs protein expression

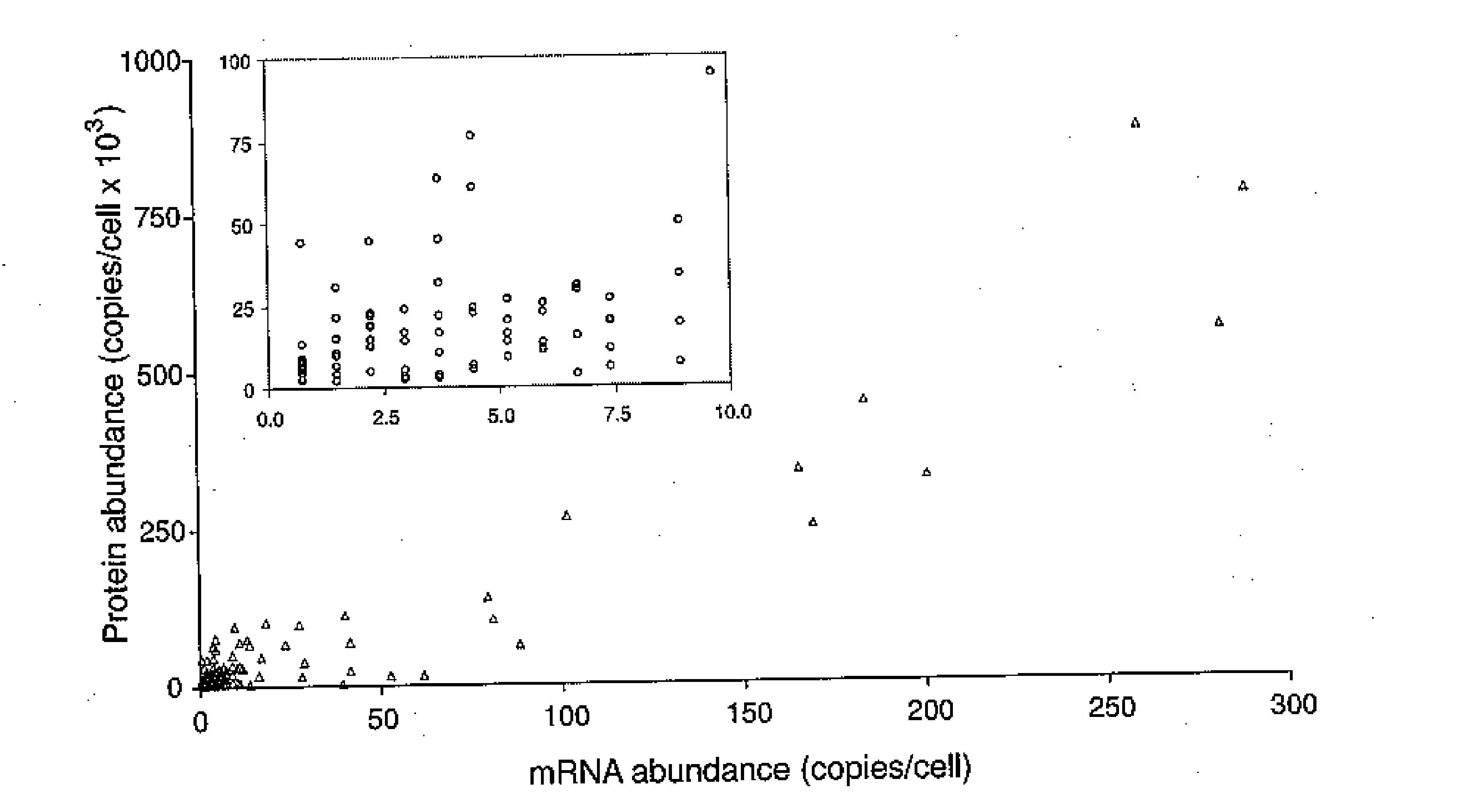
Proteins are the main players in cell function but it is harder to measure them directly on a massive scale.

Transcription can be measured.

+ control at the transcript level (splicing etc) is taken into account

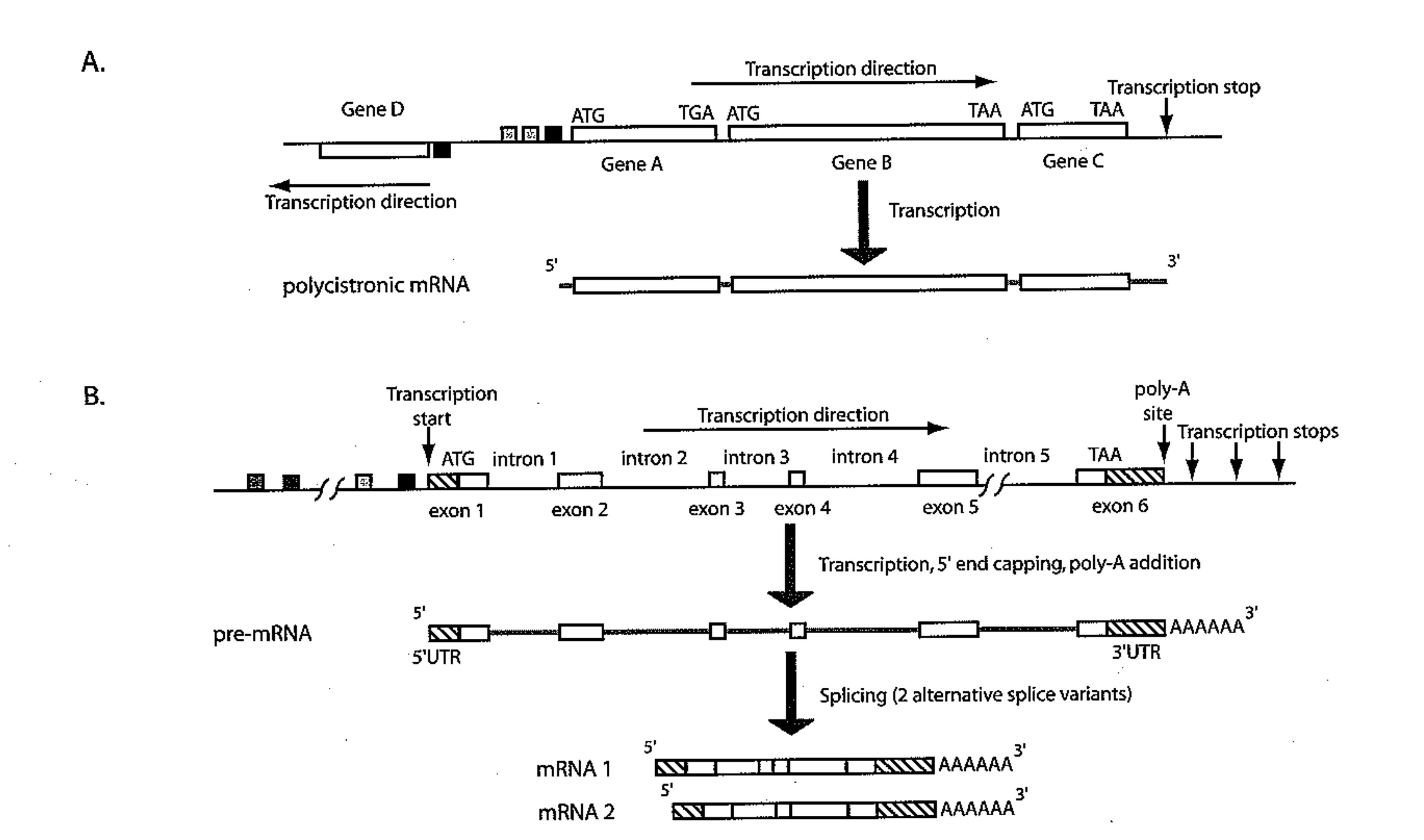
- regulation at the translational level is not
- modifications of the proteins after translation, and differences in degradation speed are not

Correlation of protein and mRNA abundances



How to measure?

Details of transcription



Measuring transcript levels

“Closed” vs. “open” architectures

Closed: Need to have prior knowledge to define “probes” of what to look for

- spotted microarrays
- oligonucleotide chips

Open: Do not need probes

- TOGA (TOtal Gene expression Analysis)
- SAGE (Serial Analysis of Gene Expression)

TOtal Gene expression Analysis (TOGA)

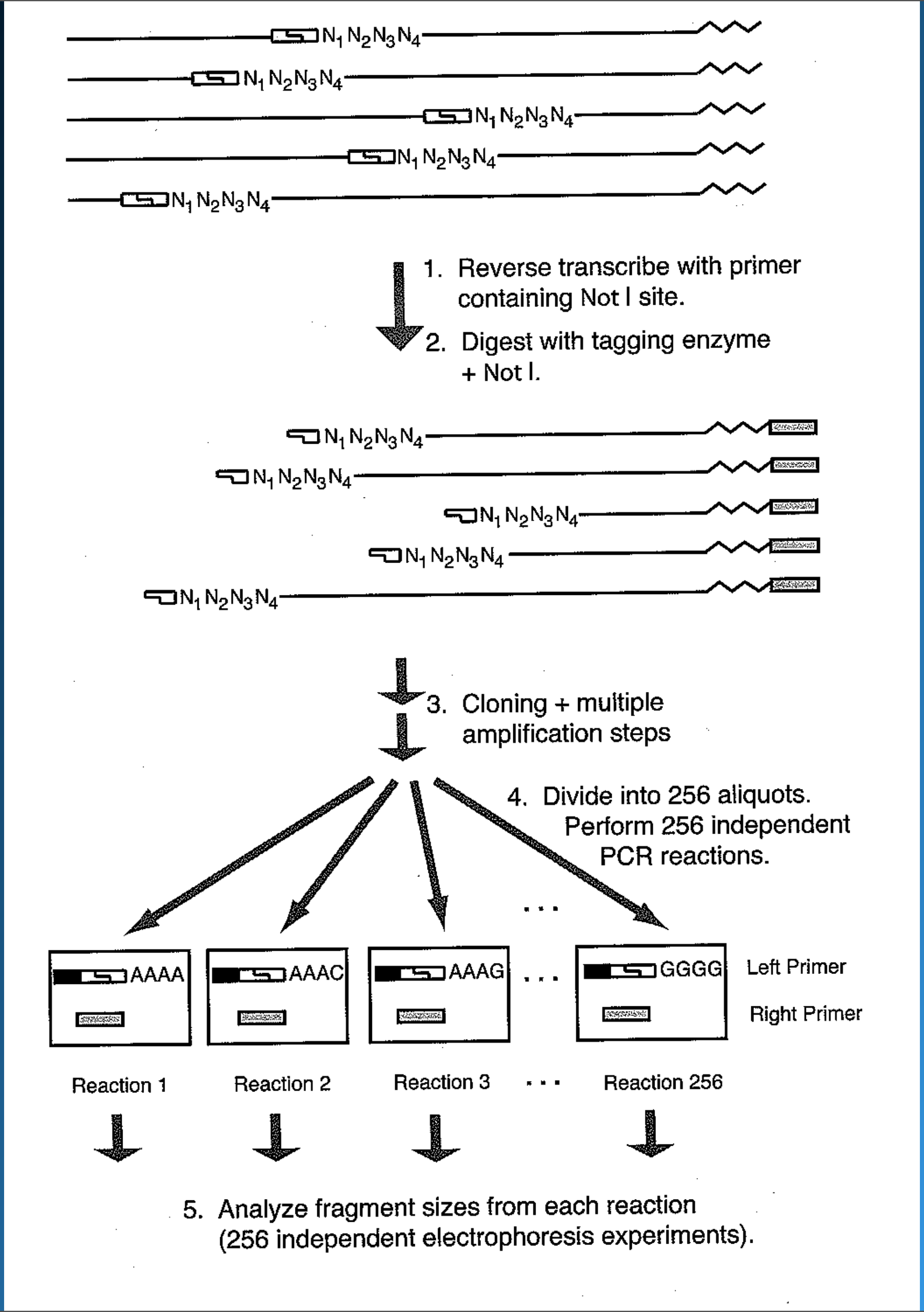
Overall idea: Divide a pile of unknown mRNA samples, with a fixed algorithm, into a large set of smaller piles such that with reasonable accuracy each pile contains only one kind of mRNA.

Algorithm:

- search for the last occurrence of CCGG
- divide into 256 subpiles based on the four next nucleotides
- divide each subpile into subsubpiles based on the length of the sequence from CCGG to the end

+ No need to define the set of sought mRNAs *a priori.*

- Does not give out the mRNA sequence



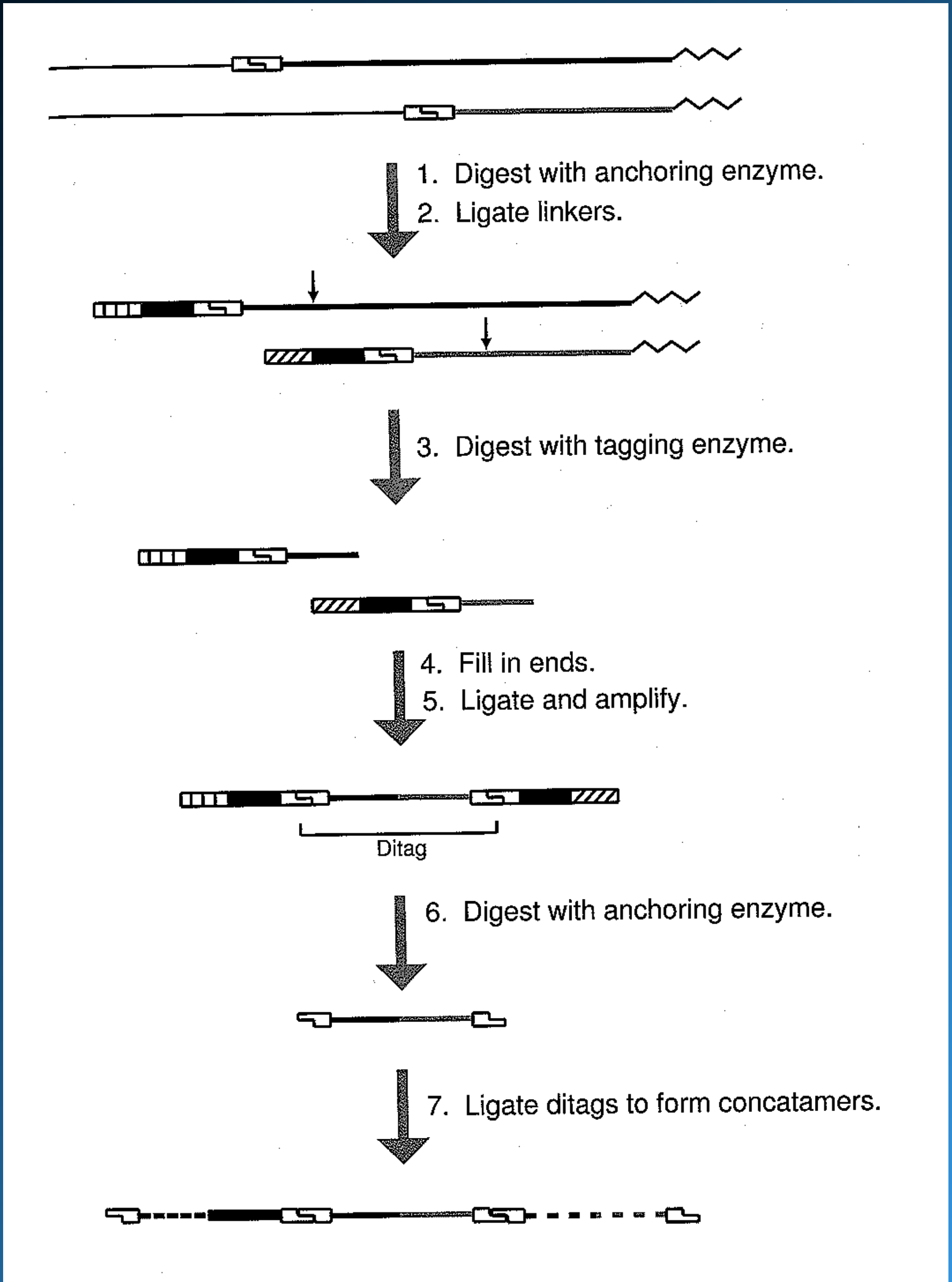
1. Reverse transcribe with primer containing Not I site.
2. Digest with tagging enzyme + Not I.
3. Cloning + multiple amplification steps
4. Divide into 256 aliquots. Perform 256 independent PCR reactions.
5. Analyze fragment sizes from each reaction (256 independent electrophoresis experiments).

Serial Analysis of Gene Expression (SAGE)

Overall idea: Pick 14 nt long sequences from each mRNA, resulting in sequences that are unique to the mRNAs with reasonable accuracy. Then compute the abundance of each 14 nt long sequence.

Algorithm: Search for the last CCGG in each mRNA (and for the last GATC but let's skip that). Find the 14-mer starting from that CCGG. Compute the abundance of the 14-mers.

Difference from TOGA: TOGA used PCRs and electrophoresis gels. SAGE uses sequencing. Both PCRs and sequencing machines are ubiquitous, but the gels are harder to analyze.



1. Digest with anchoring enzyme.
2. Ligate linkers.
3. Digest with tagging enzyme.
4. Fill in ends.
5. Ligate and amplify.
6. Digest with anchoring enzyme.
7. Ligate ditags to form concatamers.

Summary of non-probe-based approaches

The mRNA sequences need not be known *a priori*.

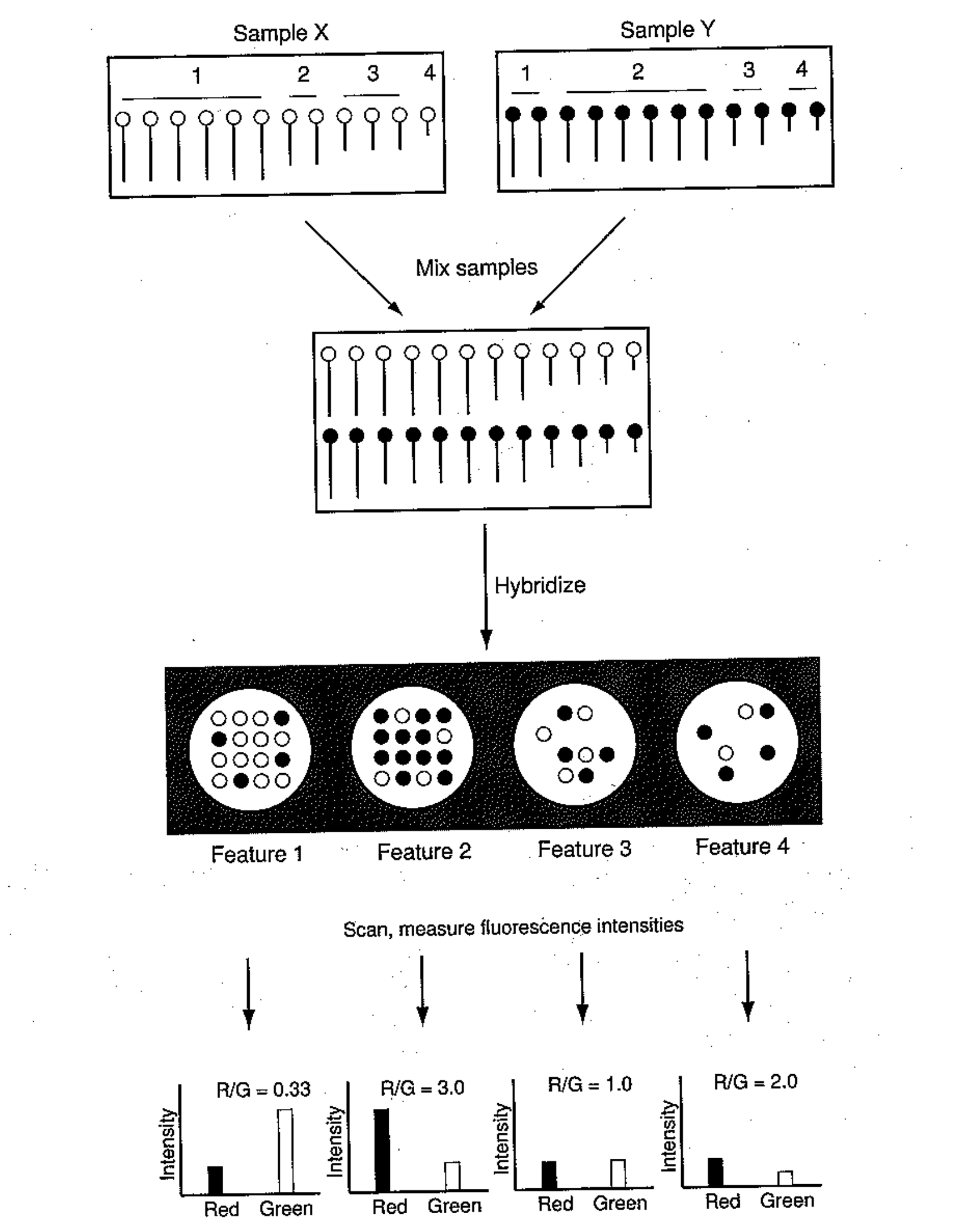
Neither will they be known *a posteriori* (without further analysis).

Invaluable for new species or even collections of species (samples of bacteria/algae etc.).

-Will be replaced by high-throughput sequencing methods within the next (few) years.

Measurement of differential expression by microarrays

Principle


Sample X
1
2
3
4
Sample Y
1
2
3
4
Mix samples
Hybridize
Feature 1
Feature 2
Feature 3
Feature 4
Scan, measure fluorescence intensities
R/G = 0.33
R/G = 3.0
R/G = 1.0
R/G = 2.0
Intensity
Red
Green

Background on microarrays

Probe: A template sequence, to which a matching mRNA (actually cDNA) binds. Usually (cDNA-) sequence from a specific gene.

cDNA: DNA complementary to RNA, produced by *reverse transcription*. When made of mRNA, it contains only the coding regions of a gene.

Target: The mRNA sample that is matched against the probes, to measure the amount of each mRNA type = activity of the gene.

Feature: (For microarrays:) A detector of a certain kind of mRNA. It has a specific location on the microarray

Microarray: A regular grid of features

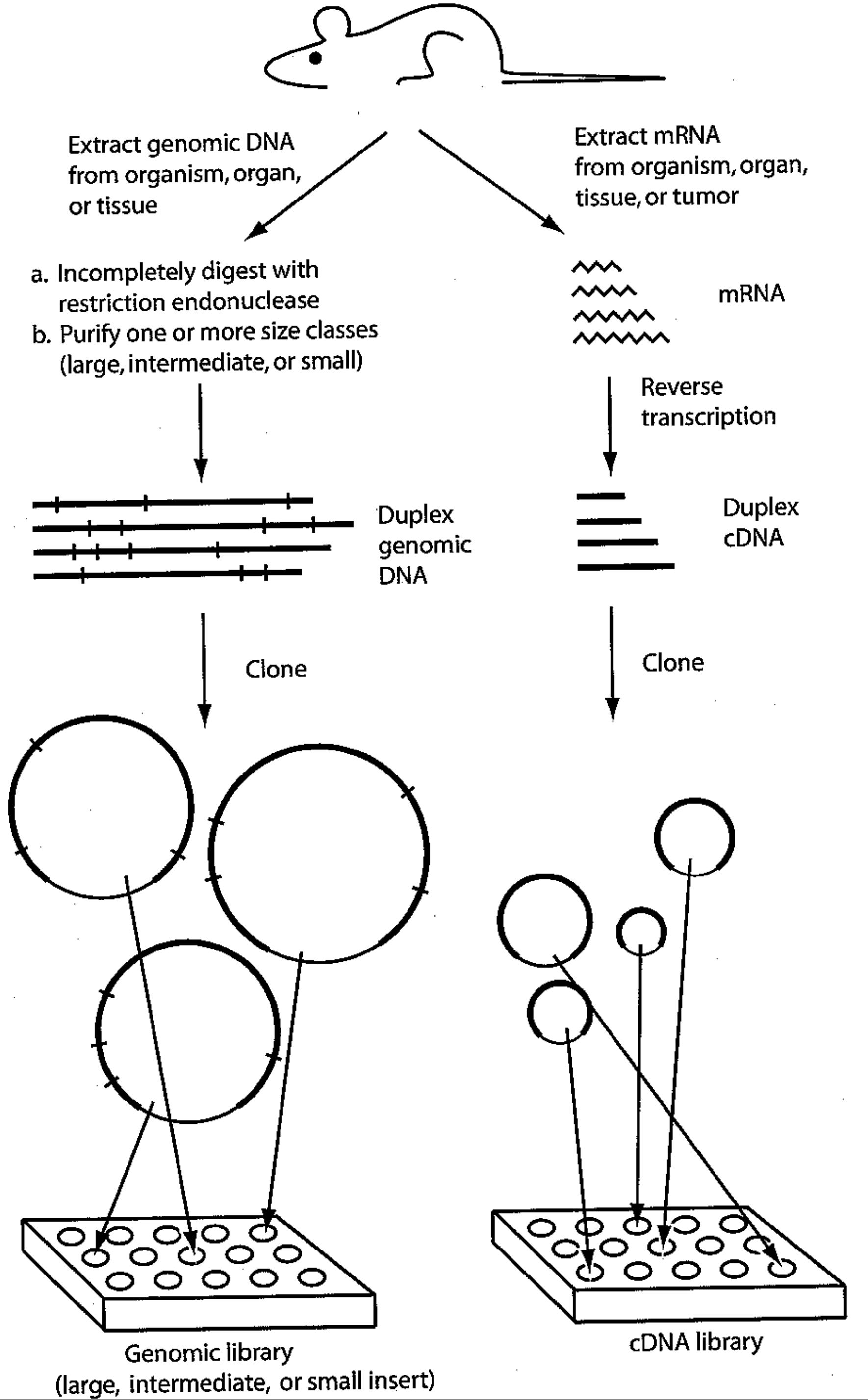
Background on microarrays, cntd.

Synthesized oligonucleotide: Probes created directly, i.e., not by cloning. Length 25-60 nt.

Hybridization: Two single-stranded DNAs will bind to each other if they are close enough in space and their sequences are complementary.

Spotted (cDNA) microarrays

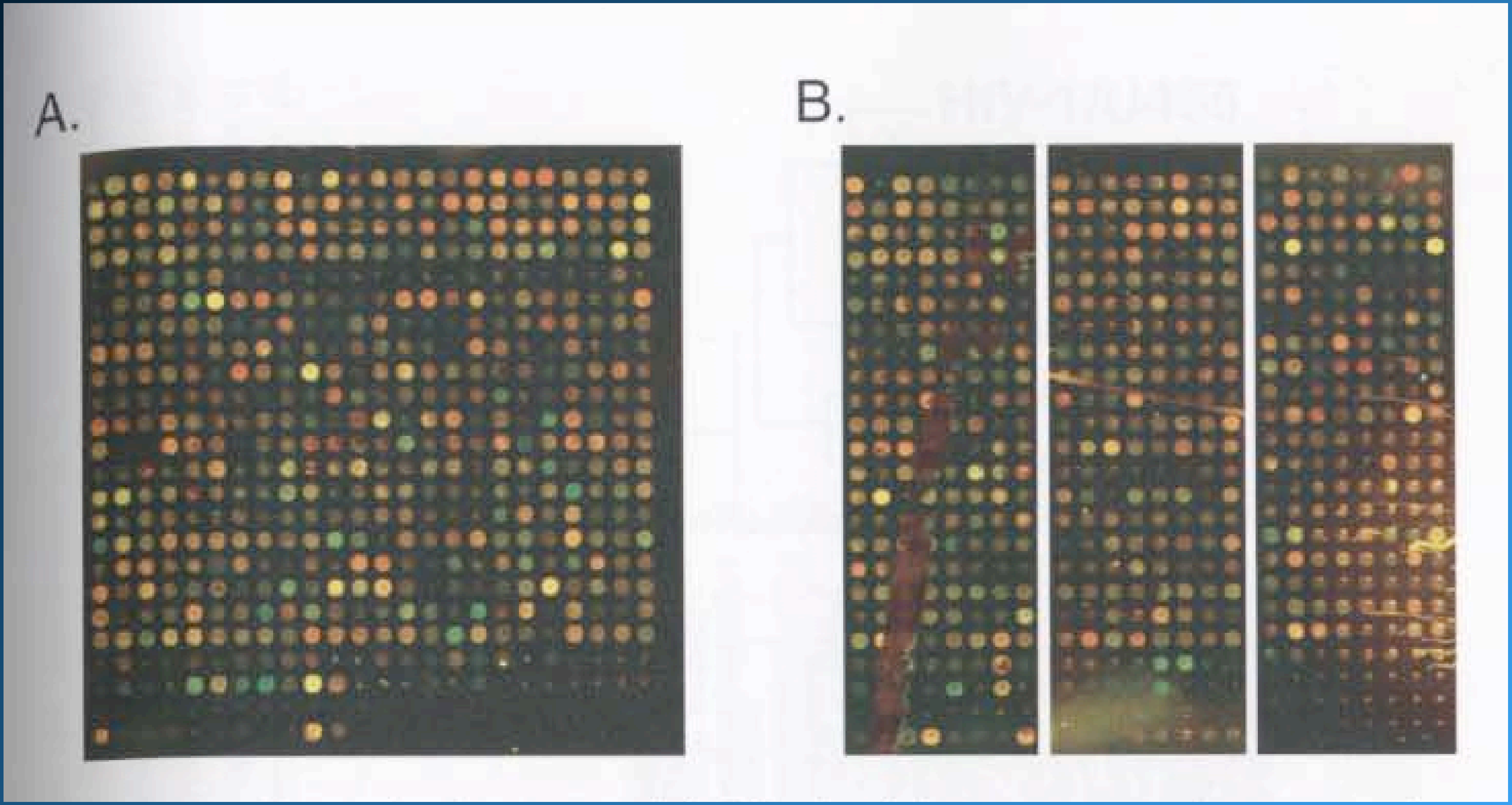
- Probes are cDNA stored beforehand in clone libraries. mRNA corresponding to genes can be recognized by the poly-A tails. Length > 200 nt.
- cDNA are denatured to single strands, and cDNA from one gene is spotted as a feature in a specific location on the array
- Spotting is done by printing robots: Printing heads are dipped into liquid containing cDNA, pressed onto the slide, and the cDNA then fixed to the slide.
- Accuracy: About one mRNA/cell when isolated from 10^6 cells (20pg per 20ug of mRNA)


Extract genomic DNA from organism, organ, or tissue
Extract mRNA from organism, organ, tissue, or tumor
a. Incompletely digest with restriction endonuclease
b. Purify one or more size classes (large, intermediate, or small)
mRNA
Reverse transcription
Duplex genomic DNA
Duplex cDNA
Clone
Clone
Genomic library (large, intermediate, or small insert)
cDNA library

Spotted microarrays ctd.

- Two targets are labeled differently by fluorescent dyes, Cy3 (green) and Cy5 (red)
- Both targets are hybridized on the same slide. cDNA from each binds to the same set of probes. The amount bound is (hopefully) proportional to the relative amount of mRNA in the two targets.
- Scanning: The slide is stimulated by “red” light to excite the Cy5 labels, and the amount of intensity at each location on the array is read. Same for green.
- This produces two large images

Examples of slides/arrays: Fruit fly mutant (Cy5, red) vs. wild type (Cy3, green)



First steps of data analysis

- Find the spots
- Quantify the intensities relative to background (?)
- Compute relative intensities
- Remove artefacts

Expression microarrays

Up to now: cDNA/spotted microarrays.

Alternatives:

1. Spotted, but instead of clone libraries use synthesized oligonucleotides
2. Synthesize the oligonucleotides directly on chips with litographic techniques (Affymetrix). These measure accurately one sample at the time (not two labeled samples as in spotted arrays)

Pros and cons

Of microarrays (vs. “open” sequencing):

+ large scale (10^4-10^5 features/genes)

- need to pre-define probes

Of spotted arrays (vs. oligonucleotide chips):

+ customizable

- noisy

The newest generation of oligonucleotide chips are customizable.

Gallup

- Did you learn something new?
- What is missing?

- Did you get an answer to your question?

HELSINKI UNIVERSITY OF TECHNOLOGY

LABORATORY OF COMPUTER AND INFORMATION SCIENCE

Data Analysis (Chapter 11: Measuring Expression of Genome Information)

Samuel Kaski

Assignment:

Think of at least one question for which you want to get an answer during this lecture.

Plan

Let’s again follow the book pretty closely.

Task of the lectures: Quick overview, views of the lecturer, opportunities to ask/discuss.

Content (each very briefly)

- normalization
- statistical testing for differential expression
- experimental design

(- clustering)

- components of data
- examples of analyses

Main tasks:

1) Estimate the *gene expression matrix* based on the raw measurement values

Microarrays; time points or conditions

$$\begin{array}{cc|c|c|c|c|}
\hline
 & & & & & \\
\hline
\text{Genes} & \mathbf{x}_i & x_{i1} & x_{i2} & \dots & x_{in} \\
\hline
 & & & & & \\
\hline
\end{array}$$

2) Interpret the matrix. (Piece of cake...)

Note: It would in principle be better to do both 1 and 2 in a single step. It would make the estimation more accurate since all uncertainties in the data could in principle be properly taken into account.

Modularizing the process into two separate steps makes the computation and thinking easier but may produce sub-optimal results.

Normalization

Purpose: Remove biases resulting from experimental setting/parameters.

The book focuses on removal of dye bias; an even more important task is to make measurements done with different microarrays compatible.

Dye bias

Binding efficiency of cDNAs labeled by Cy3 and Cy5 may be different.

Empirical solution: *Dye swap*. Label sample A with Cy3 and sample B with Cy5 and measure relative expression with a microarray. Swap the labels and measure again. Take the average.

Pros: Very few assumptions needed.

Cons: Need two microarrays.

Dye bias cntd.: Global normalization

A modeling solution: Assume a relationship between values measured by the red dye (“R”) and the green dye (“G”) for the same sample. Linear dependency is sensible:

$R=kG$

Estimate the *parameter k* from data. Since we do not have same samples measured by R and G on the same chip, we can assume that the relationship holds for the set of all genes. (Sensible if we assume that most are noise and the active ones are symmetric.)

Finally: Correct *R* by normalizing with 1/*k*.

Dye bias cntd: Intensity-dependent normalization

The linear dependency $R=kG$ is not enough, since k turns out to depend on the intensity.

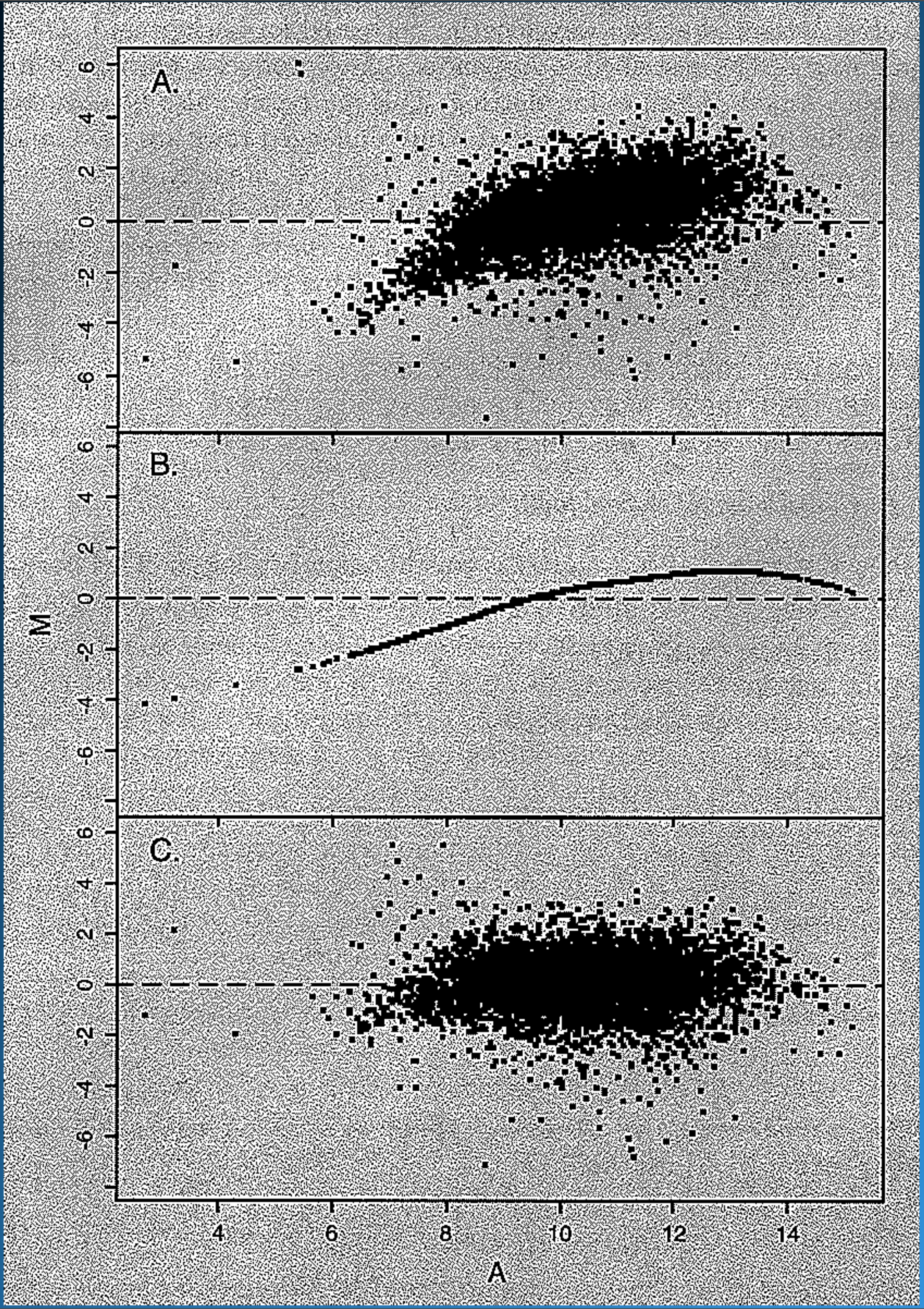
Define the average intensity by

$A = (\log R + \log G)/2 = \log (RG) / 2$

and the differential expression by

$\mathrm{M} = \log \mathrm{R} - \log \mathrm{G} = \log (R/G)$

Global correction makes the global average of M equal to zero. Local correction: make average of M zero for each A.


A.
B.
C.
M
A
-6
-4
-2
0
2
4
6
4
6
8
10
12
14

Summary about normalizations

We focused on the dye bias to make things concrete.

Other kinds of normalizations are needed to make measurements done with different microarrays (and especially different types of microarrays) compatible.

Lots of methods have been and are being proposed, and need to be developed.

Note that it is hard to automate this part. The data production process needs to be understood to some extent.

Testing for differential expression

Testing for differential expression

Problem: Assess whether expression of a gene in a *treatment* differs from the *control* (“standard condition”).

Trivial solution: Fold change (e.g. “Three-fold change”)

Obvious problem: Since there is noise in the data, we cannot know whether the difference is due to random fluctuations.

Need several *replicate measurements* and *statistical testing*.

Reminder: Statistical testing

Key idea: Assume the data comes from a baseline distribution (the *null hypothesis* holds). Evaluate how likely it is to have observed the data we have (or more extreme data). If it is very unlikely, then it is very unlikely that the null hypothesis holds either, and the null hypothesis will be rejected.

Example: Assume that the mean expression of a gene is the same in both control and treatment.

Choose a *risk level* or *significance level,* that is, the risk you are willing to tolerate that the null hypothesis will be rejected although it is in fact true.

Assuming the same standard deviations, the standard t-test is fairly robust to small sample sizes. Test whether the t-statistic has an extreme value, that is, integral from *t* to infinity is smaller than the chosen risk level.

$$t = \frac{\bar{X}_j^t - \bar{X}_j^c}{s\sqrt{\frac{1}{n_t} + \frac{1}{n_c}}}$$

Here *t* denotes treatment, *c* control, the *X*'s the sample means for gene *j*, the *s* the sample standard deviation, and the *n* the numbers of samples in control and treatment.

Multiple testing

Major problem for gene expression data: small *n*, large *p* (not “p-value” but number of genes...)

When testing for several genes, the likelihood for finding differential expression in some of the tests increases, compared to when testing for only one. The number of false positives increases.

This *multiple hypothesis testing* needs to be taken into account.

Bonferroni correction is a very conservative way: To get a significance level α_B for the whole experiment, use $\alpha = \alpha_B / N$ for each single test.

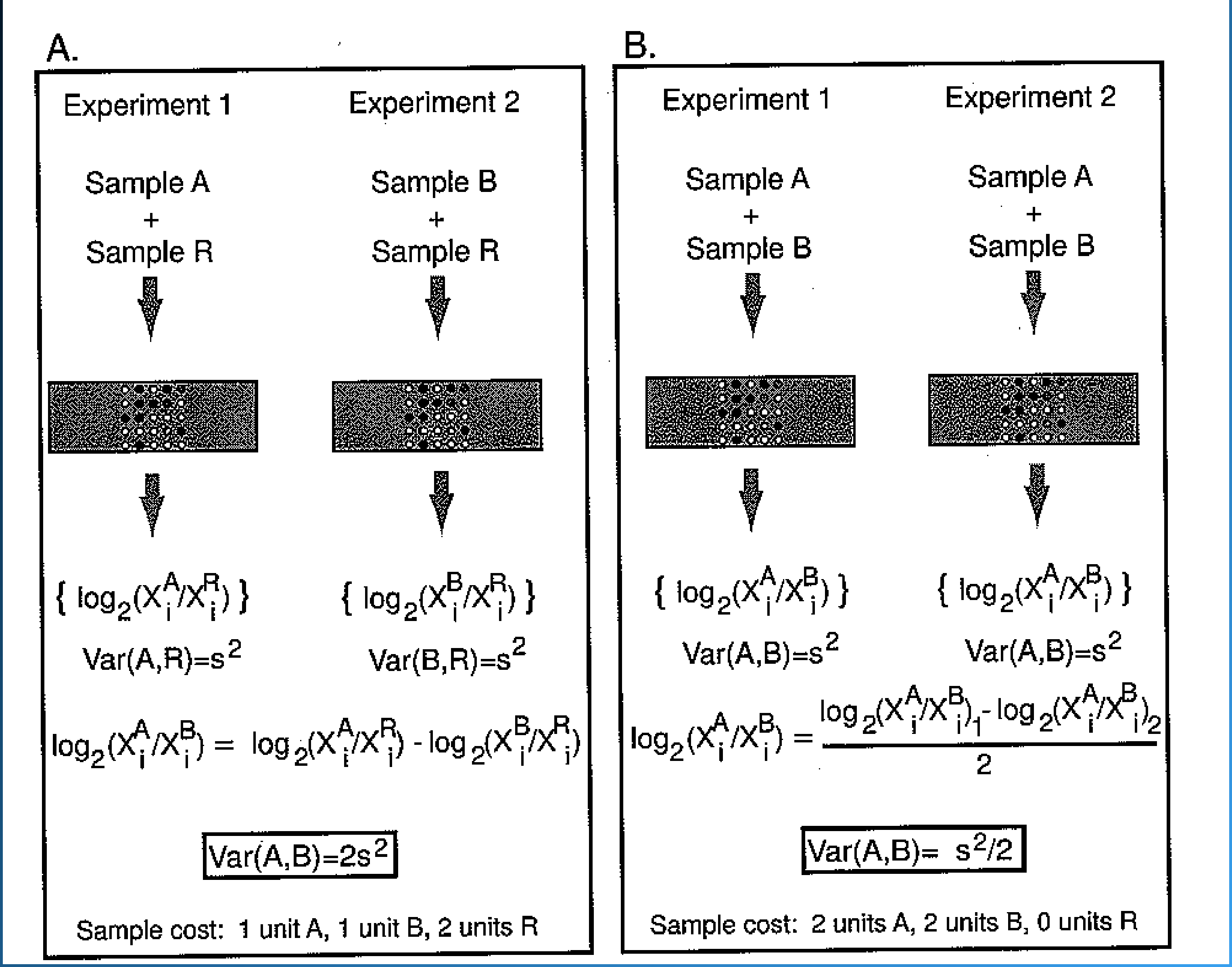
Experimental Design

Type of replication / sources of variation

So we need replicate measurements to evaluate whether differences are due to random fluctuations. But which kind of random fluctuation?

- Biological variation: samples from several individuals are needed to make conclusions about populations
- Technical (measurement) variation: Replicates of the sample preparation.
- Slide (microarray) and processing-specific variation: Replicates of microarrays

Common reference vs. replicated meas.



A.

Experiment 1

Sample A + Sample R

$\{ \log_2(X_i^A/X_i^R) \}$

$\mathrm{Var}(A,R)=s^2$

Experiment 2

Sample B + Sample R

$\{ \log_2(X_i^B/X_i^R) \}$

$\mathrm{Var}(B,R)=s^2$

$$\log_2(X_i^A/X_i^B) = \log_2(X_i^A/X_i^R) - \log_2(X_i^B/X_i^R)$$

$$\boxed{\mathrm{Var}(A,B)=2s^2}$$

Sample cost: 1 unit A, 1 unit B, 2 units R

B.

Experiment 1

Sample A + Sample B

$\{ \log_2(X_i^A/X_i^B) \}$

$\mathrm{Var}(A,B)=s^2$

Experiment 2

Sample A + Sample B

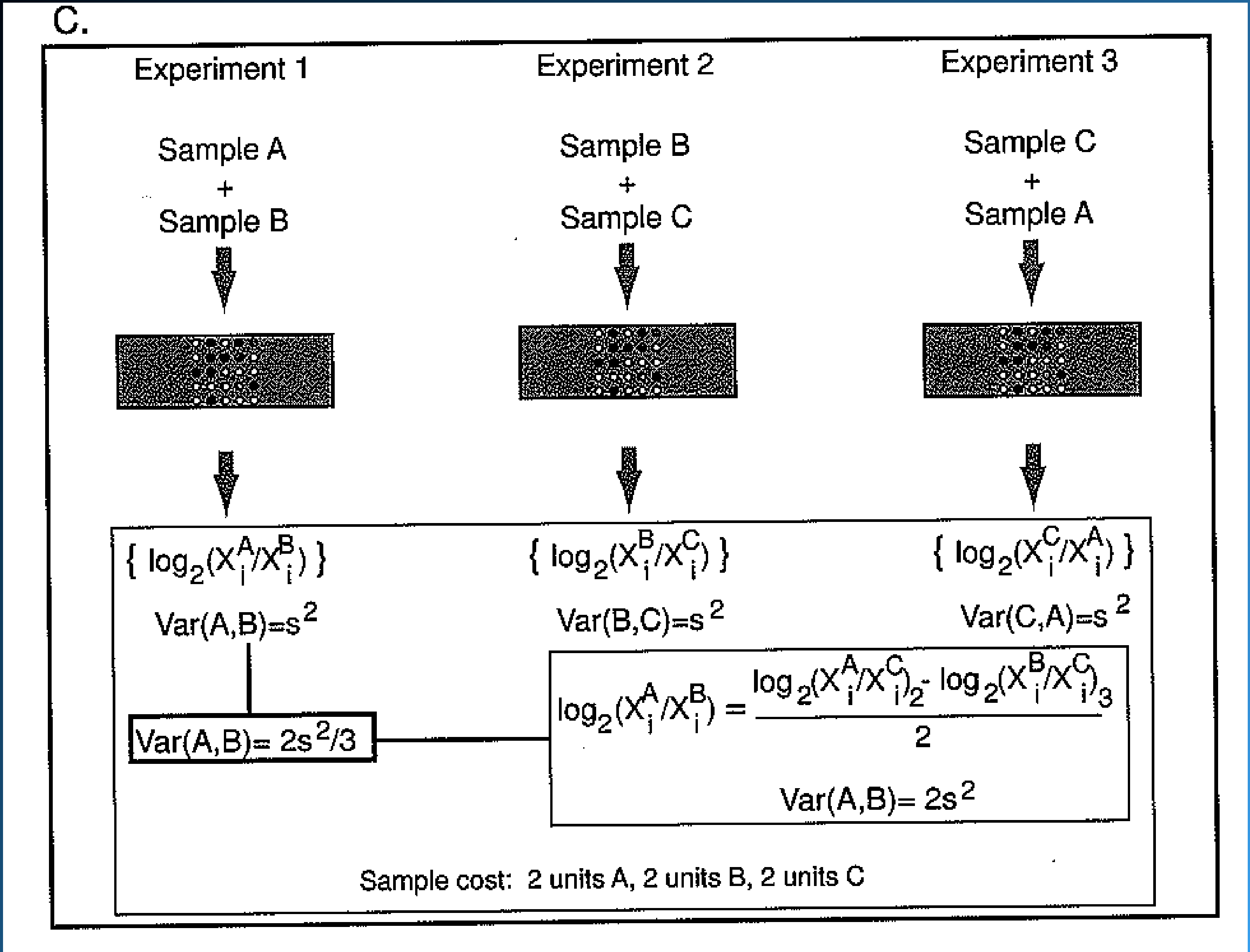
$\{ \log_2(X_i^A/X_i^B) \}$

$\mathrm{Var}(A,B)=s^2$

$$\log_2(X_i^A/X_i^B) = \frac{\log_2(X_i^A/X_i^B)_1 - \log_2(X_i^A/X_i^B)_2}{2}$$

$$\boxed{\mathrm{Var}(A,B)= s^2/2}$$

Sample cost: 2 units A, 2 units B, 0 units R



C.

Experiment 1

Sample A + Sample B

$\{ \log_2(X_i^A/X_i^B) \}$

$\mathrm{Var}(A,B)=s^2$

$\mathrm{Var}(A,B)= 2s^2/3$

Experiment 2

Sample B + Sample C

$\{ \log_2(X_i^B/X_i^C) \}$

$\mathrm{Var}(B,C)=s^2$

Experiment 3

Sample C + Sample A

$\{ \log_2(X_i^C/X_i^A) \}$

$\mathrm{Var}(C,A)=s^2$

$$\log_2(X_i^A/X_i^B) = \frac{\log_2(X_i^A/X_i^C)_2 - \log_2(X_i^B/X_i^C)_3}{2}$$

$\mathrm{Var}(A,B)= 2s^2$

Sample cost: 2 units A, 2 units B, 2 units C

“Data interpretation”

Tasks

Start from a gene expression matrix. Common tasks include:

- Annotation
- Search for co-regulated gene groups
- Classify tissues/conditions

This is a non-exhaustive list. Note that in this broadly-scoped course we can only get a glimpse of the very basics.

Background

Supervised methods: Predict c given **x**.

- Examples: Regression, classification

Unsupervised methods: Characterize **x** / find regularities in **x**.

- Examples: Clustering, component analysis

Here: Response variables: c

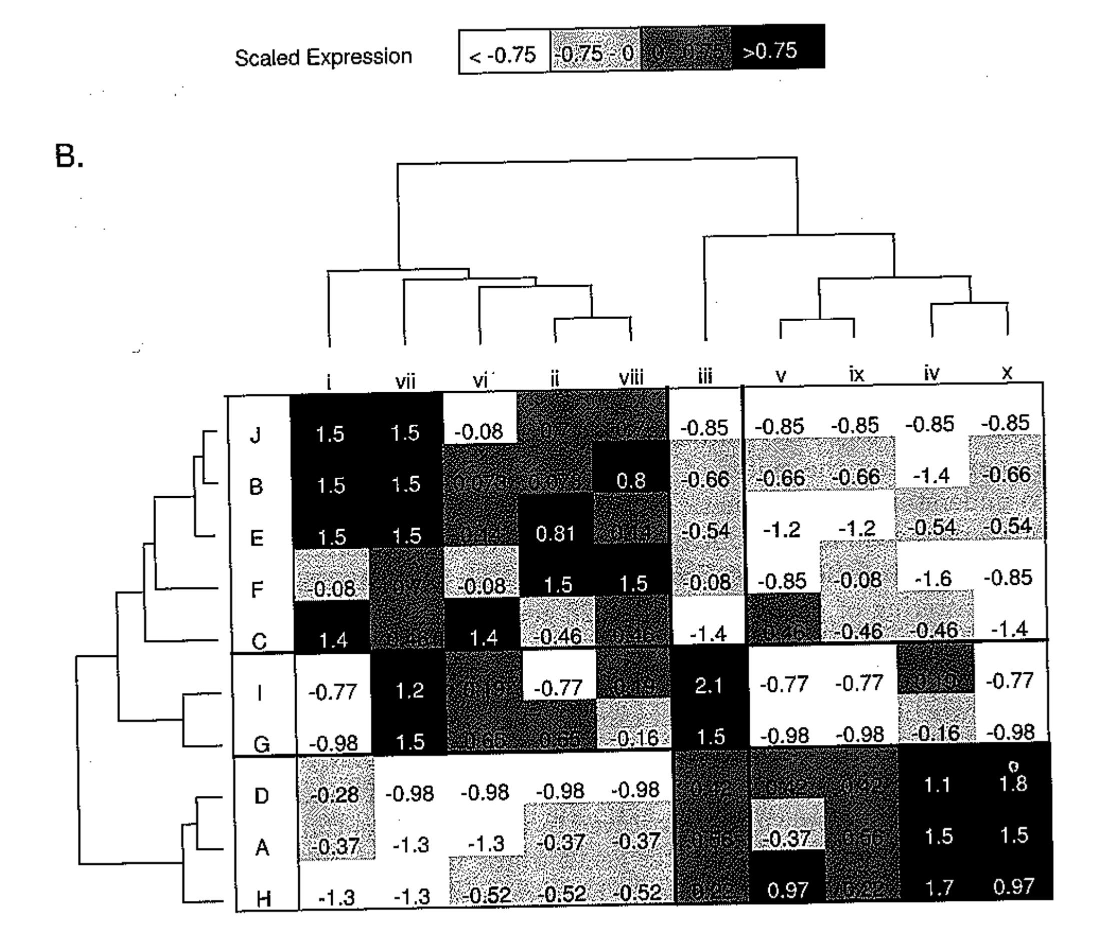
Predictor variables or attributes: **x**

Profile: **x**

Clustering

In a separate slide set.

Clustering both ways for visualization



	i	vii	vi	ii	viii	iii	v	ix	iv	x
J	1.5	1.5	-0.08	0.72	0.72	-0.85	-0.85	-0.85	-0.85	-0.85
B	1.5	1.5	0.075	0.075	0.8	-0.66	-0.66	-0.66	-1.4	-0.66
E	1.5	1.5	0.14	0.81	0.14	-0.54	-1.2	-1.2	-0.54	-0.54
F	-0.08	0.72	-0.08	1.5	1.5	-0.08	-0.85	-0.08	-1.6	-0.85
C	1.4	0.46	1.4	-0.46	0.46	-1.4	0.46	-0.46	-0.46	-1.4
I	-0.77	1.2	0.19	-0.77	0.19	2.1	-0.77	-0.77	0.19	-0.77
G	-0.98	1.5	0.65	0.65	-0.16	1.5	-0.98	-0.98	-0.16	-0.98
D	-0.28	-0.98	-0.98	-0.98	-0.98	0.42	0.42	0.42	1.1	1.8
A	-0.37	-1.3	-1.3	-0.37	-0.37	0.56	-0.37	0.56	1.5	1.5
H	-1.3	-1.3	-0.52	-0.52	-0.52	0.22	0.97	0.22	1.7	0.97

Ultimately: bi-clustering

Components of data

Clustering reduces the number of profiles by grouping them. Component analysis reduces the number of variables (attributes) by combining them into *components*.

Principal components analysis (PCA): Combine the variables into uncorrelated linear components, such that the first component captures the maximal amount of variance, the second one maximal amount of the rest while being uncorrelated with the first, etc.

Can be computed with an eigenvalue decomposition.

Confirmation of results

Clustering and components analyses are essentially exploratory techniques, sophisticated ways of *looking at the data*. We usually do not even assume that they give a “correct” description of the data.

Hence, the results need to be verified by further experiments.

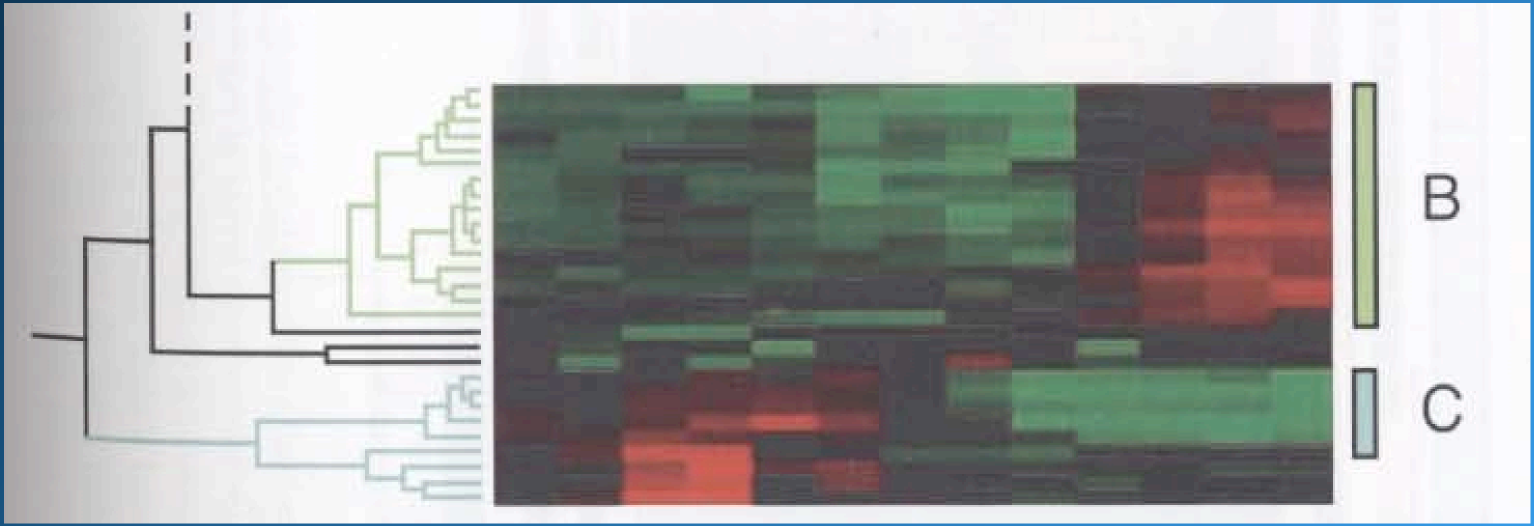
At the minimum, the measurements should be replicated. RT-PCR gives more accurate measurements (but is more laborious).

Ultimately: The analysis should produce new hypotheses which are tested in new biological experiments.

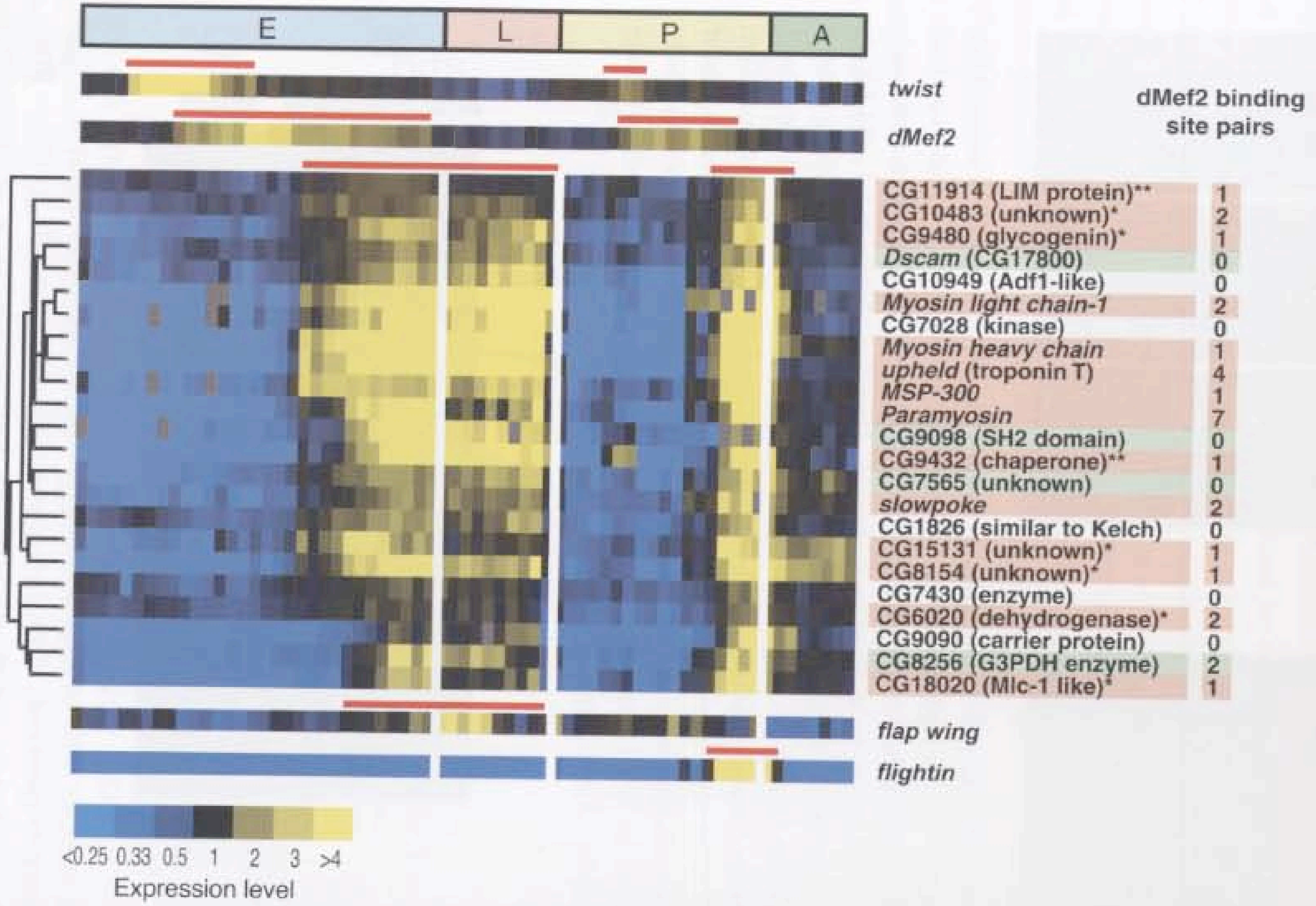
Examples of experimental applications

Gene expression in human fibroblasts

In animal cells growth is prompted by growth factors. Growth was synchronized by first depriving cells of growth factors and then giving them



Gene expression during Drosophila Development



Protein expression

Goals of proteomics studies

Measuring the protein content is among the most important but at the same time most difficult tasks.

Possible tasks:

- Which proteins, among a set of known proteins, are co-regulated
- Measure protein abundances
- Differential expression of proteins
- Measuring ligand-protein binding
- Measure protein-protein interactions

Technique #1: 2DE / MALDI-MS

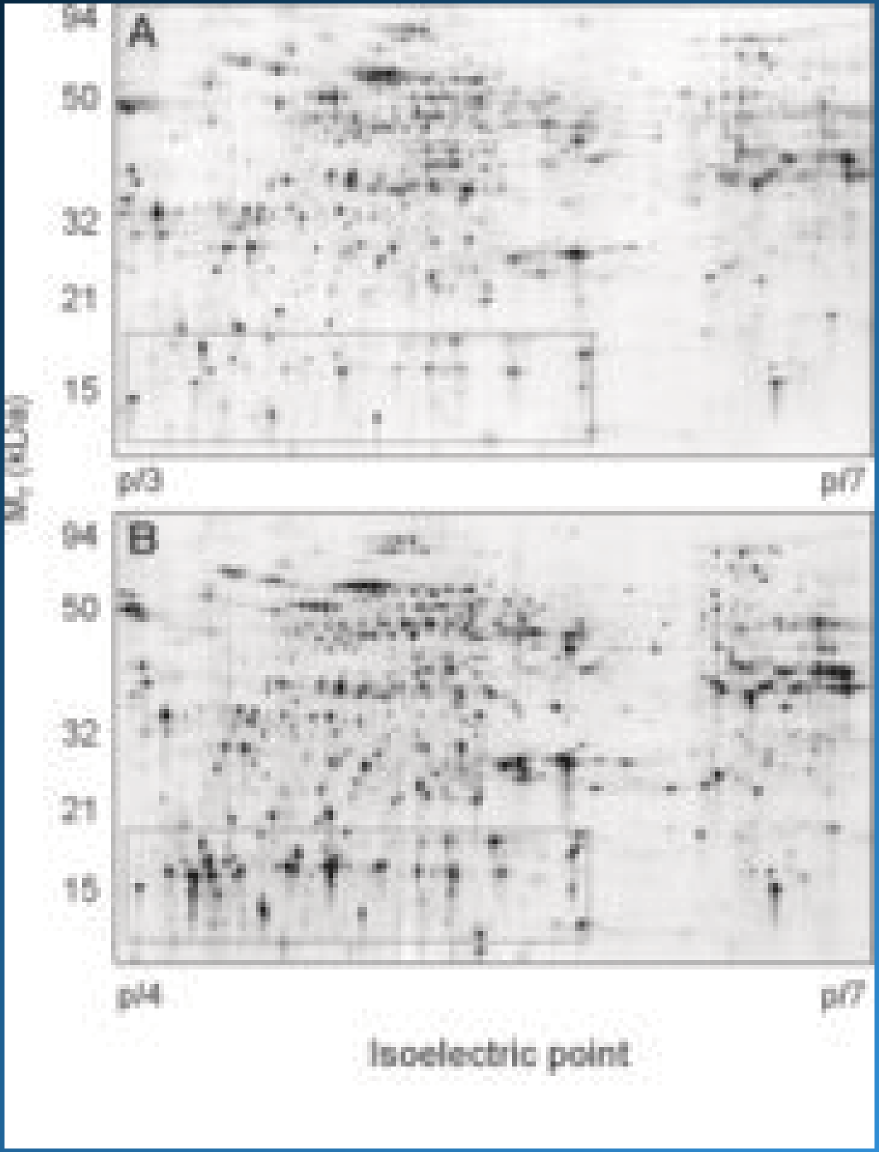
1. Separate polypeptides by 2D gel electrophoresis into spots typically containing only one polypeptide each

2. Identify each spot:

2a: Ionize the sample with laser

2b: Identify with a mass spectrometer

2D gel electrophoresis



Identify the polypeptide of one 2DE spot

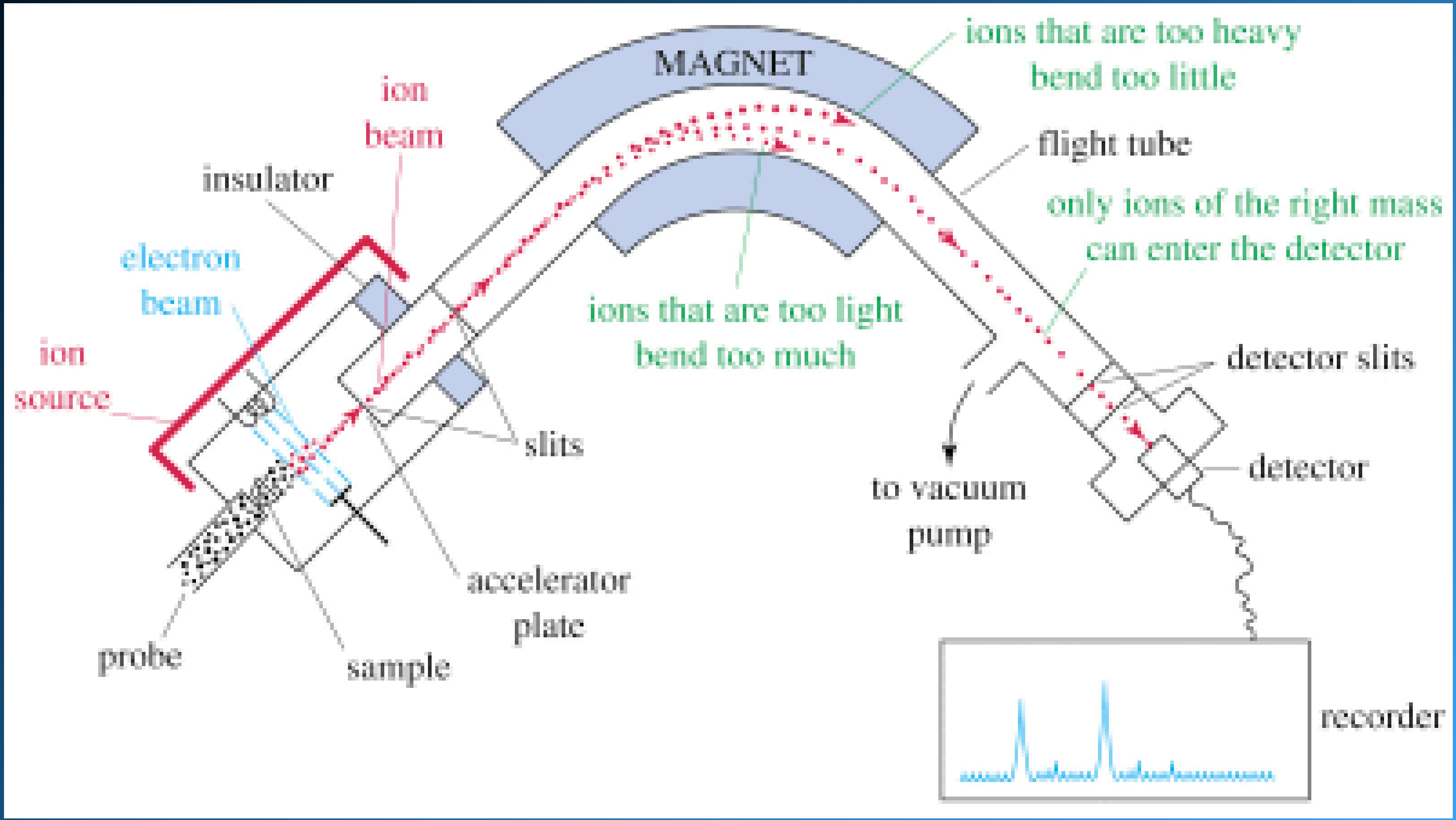
1. Cut the spot out from the gel. Cleave them into smaller pieces by a suitable enzyme.

2. MALDI: Embed in an organic matrix (substance), dry, and excite with a laser beam. The polypeptides get loose and become charged by picking one or more protons (H^+).

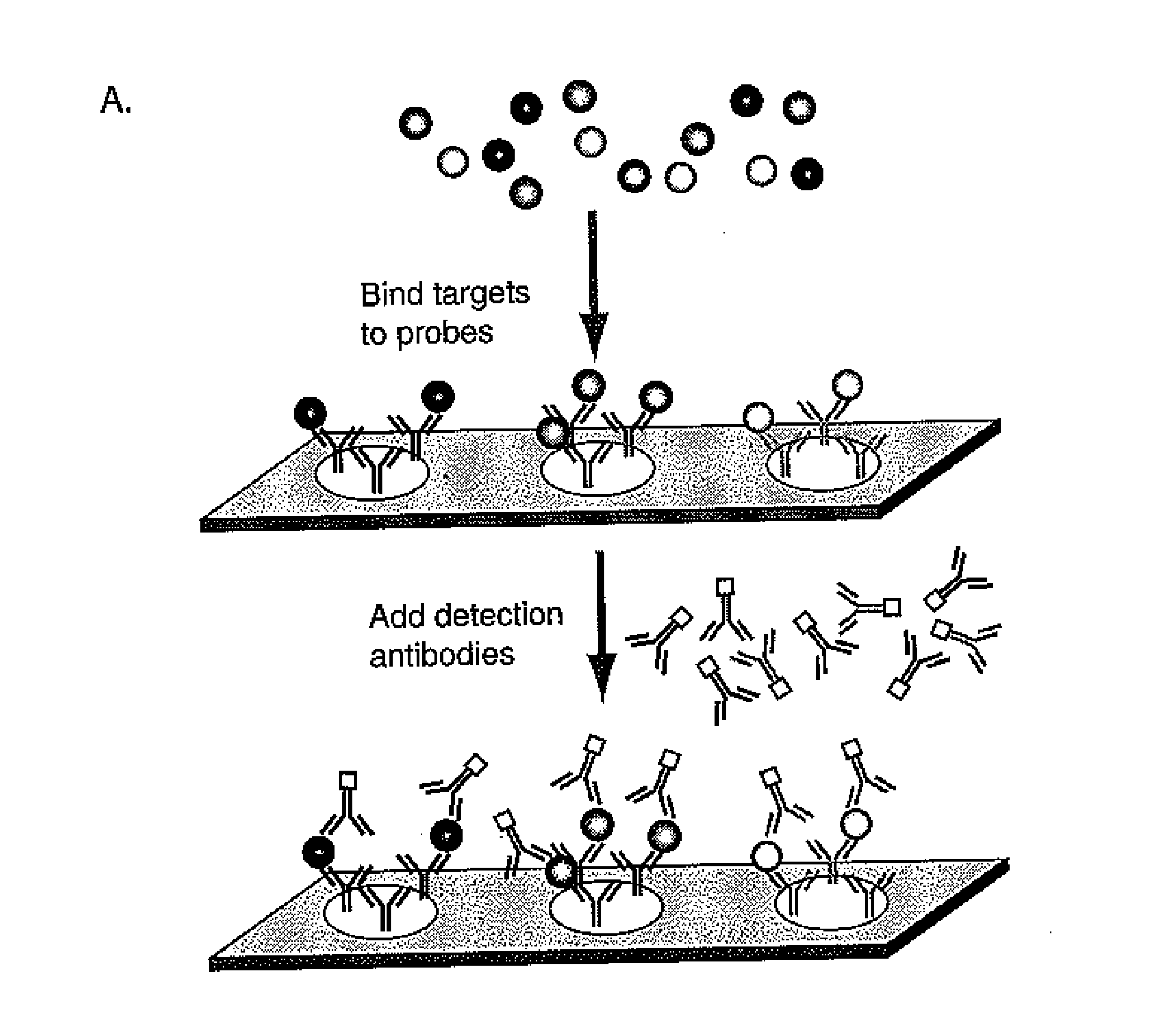
2. Mass Spectrometry: Accelerate the particles in an electric field. Their speed (or curvature) depends on their mass/charge (m/z) ratio. Measure the spectrum of m/z values produced.

3. Check from databases which polypeptide the spectrum best resembles

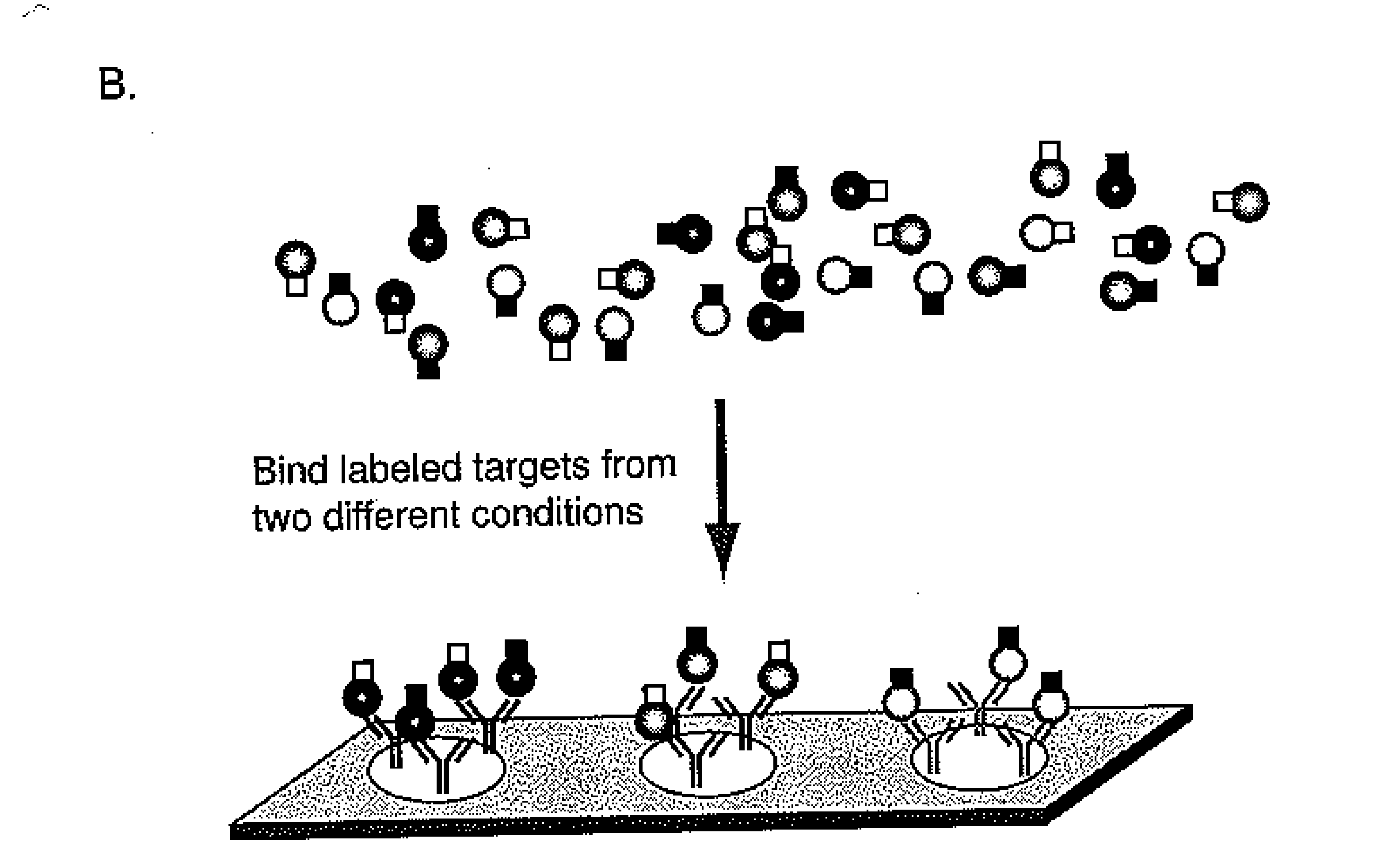
Mass Spectrometer



Protein microarrays



Protein microarrays



Gallup

- How much did you know already?
- What is missing?

- Did you get an answer to your question?