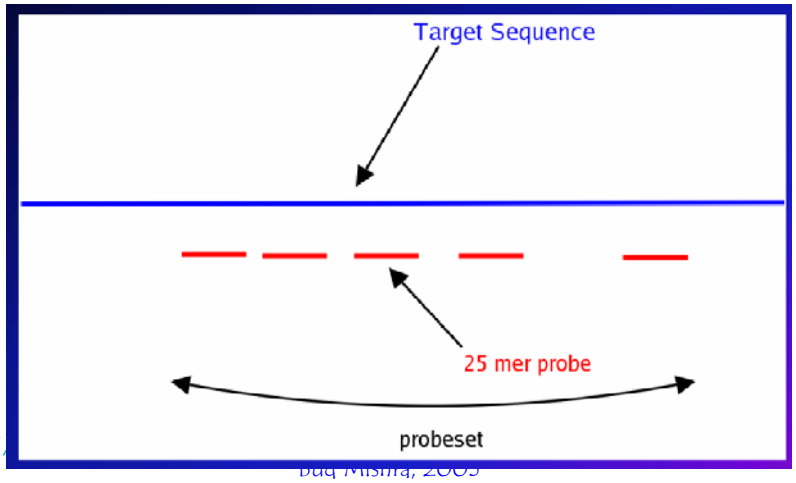
Computational Biology

Lecture #11: OMICS: Transcriptomics & Proteomics

Bud Mishra
Professor of Computer Science, Mathematics, & Cell Biology
Nov 28 2005

Probes & ProbeSets in Affymetrix Chips



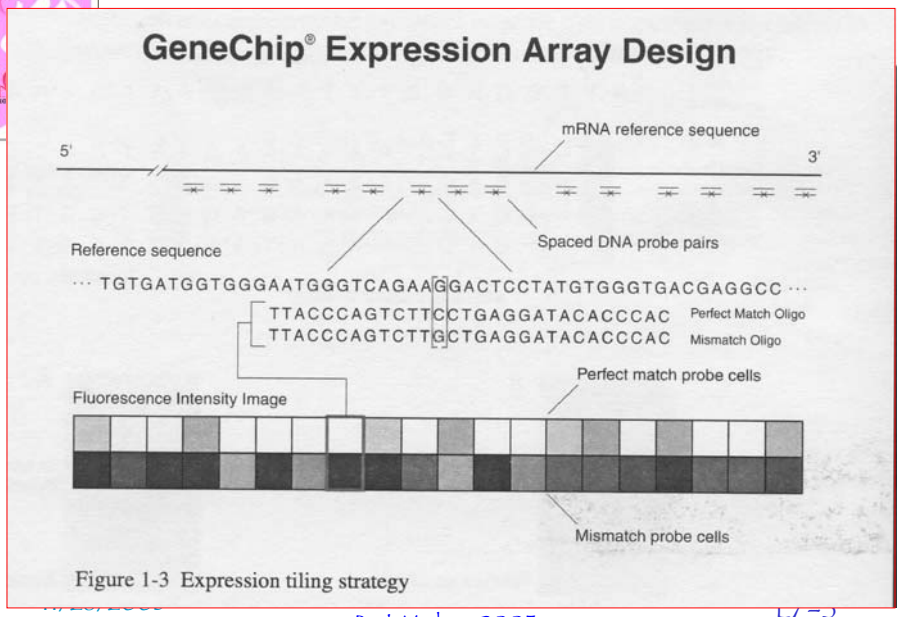


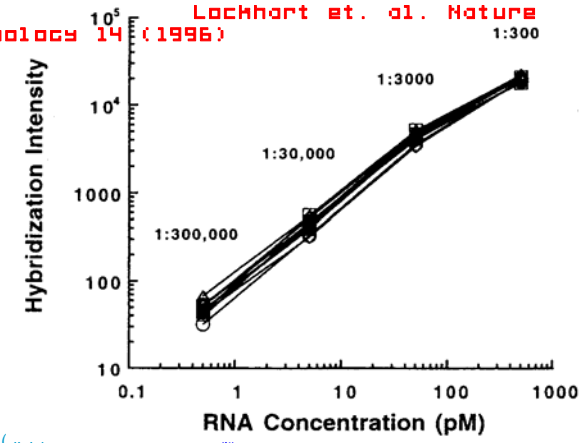
Figure 1-3 Expression tiling strategy

The big picture

- Summarize 20 PM,MM pairs (probe level data) into one number for each probe set (gene)
  - We call this number an expression measure
  - Affymetrix GeneChip Software has defaults.
- Does it work? Can it be improved?

Where is the evidence that it works?

Lockhart et. al. Nature Biotechnology 14 (1996)



Comments

- The chips used in Lockhart et. al. contained around 1000 probes per gene
- Current chips contain 11-20 probes per gene
- These are quite different situations

◇ We haven't seen a plot like the previous one for current chips

Data Processing

- The original GeneChip® software used AvDiff

$$AvDiff = |A|^{-1} \sum_{\{j \in A\}} (PM_j - MM_j)$$

  - where A is a suitable set of pairs chosen by the software. Here 30%-40-% could be <0, which was a major irritant.
  - Log PMj / MMj was also used in the above.

Data Processing

- Li and Wong (dChip) fit the following model to sets of chips

$$PM_{ij} - MM_{ij} = \theta_i \varphi_j + \varepsilon_{ij}$$

  - where $\varepsilon ij \sim N(0, \sigma^2)$. They consider θ_i to be expression in chip i. Their model is also fitted to PM only, or to both PM and MM. Note that by taking logs, assuming the LHS is ¸ 0, this is close to an additive model.
- Efron et al consider log PMj -0.5 log MMj. It is much less frequently <0.
- Another summary is the second largest PM, PM(2).

Data Processing

- The latest version of GeneChip® uses something else, namely

$$Log\{Signal\ Intensity\} = TukeyBiweight\{log(PM_j - MM_j^*)\}$$

  - with MMj* a version of MMj that is never bigger than PMj.
- Here TukeyBiweight can be regarded as a kind of robust/resistant mean.

Tukey Biweight
A robust mean

- Tukey Biweight mean of the dataset
  - Calculate the median (MED) of the data and the mean absolute deviation (MAD)
  - MED +- 5.0 * MAD comprise the limits outside which we consider the data to be outlier. (5.0 is a parameter)
  - X - MED is used to compute a weight that decays to zero outside the limits of outlier using the bi-square function.
  - Compute the weighted mean to eliminate the outliers.

Data Processing …

- *RMA* (Robust Multi-Array Averaging)
- 3 Step
  - Background removal
  - Normalization
  - Summarization

Data Processing …

- For example, dChip Background Removal was PM – MM, MAS-5 was somewhat similar
- RMA bg.correct uses a signal + plus noise model and uses the posterior mean to detect the signal.
- Works only on the PM values. The MM values serve in parameter estimation for this and normalization steps

RMA bg.correct …

- Signal: exponentially distributed
- Observed PM probe value: X = Y + Noise
- Noise: independent, mean μ, std dev = σ
- μ, σ, α (for the exponential distribution) are the three parameters to be estimated.
- Different methods for this.
  - All PM's
  - All MM's
  - Alpha from PM's mu and sigma from MM's

RMA bg.correct

- The last might be problematic
  - The MM's have a strong signal components and lead to mis-estimation of μ and σ
  - Result sensitive to mis-estimation of σ.
  - α is usually very small 0.001 – 0.002
  - We are looking at an improper flat prior being approximated by a slowly decaying exponential
  - We can take α = 0.0 in the final formula and formulate the estimation problem as estimating from an improper prior by taking limits.

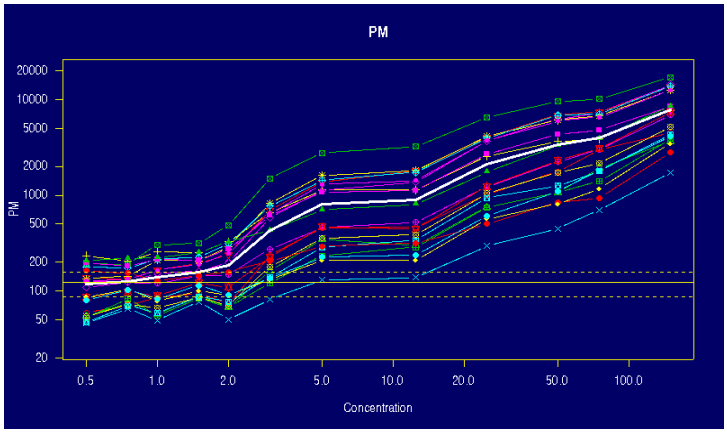
Normalization

- Goal:
  - Remove unwanted variability between chips/experiments
  - Combined with scaling to get the values between certain pre-fixed limits (MAS-5)
  - RMA: quantile normalization. Tries to achieve a linear relation between gene expression rank and response.

Summarization

- Combining the responses of the probes in the probeset to generate one value for the probeset.
- A form of mean.
  - Usually robustified
- RMA: median polish on the logged expression values
- MAS-5: Tukey Biweight (as explained earlier)
- dChip: Model based (see earlier)

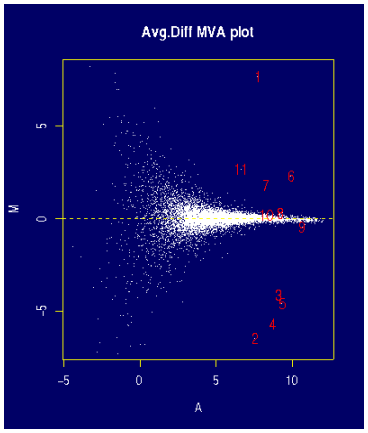
Probe level data exhibiting parallel behaviour on the log scale

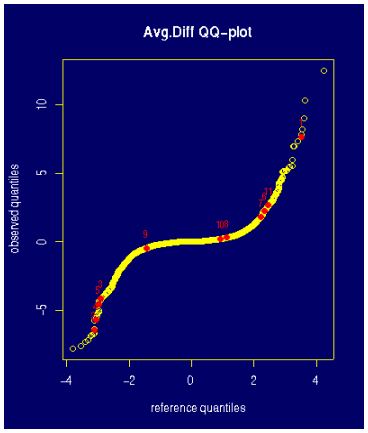


RMA in summary

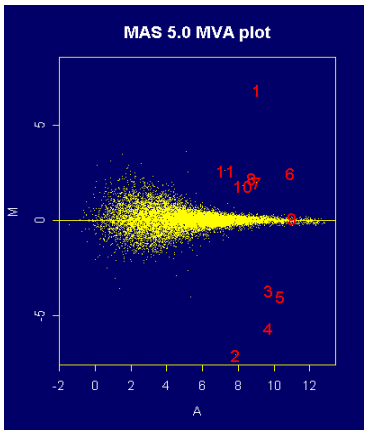
- We *background* correct PM on original scale
- We carry out *quantile* normalization
- We take *log2*
-
- Under the *additive* model
- log2 n(PMij –*BG) = m + ai + bj + εij
- We estimate chip effects *ai* and probe effects *bj* using a *robust/resistant* method.

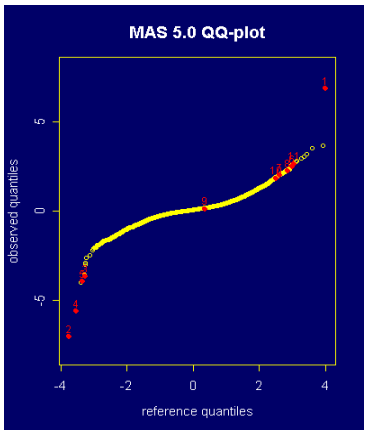
Performance (AvDiff)



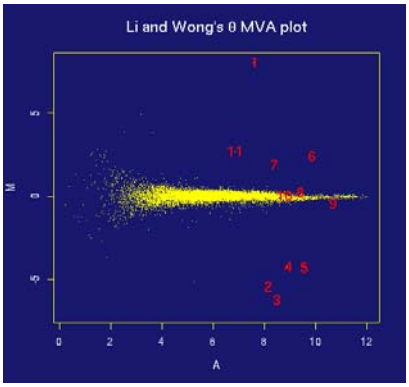


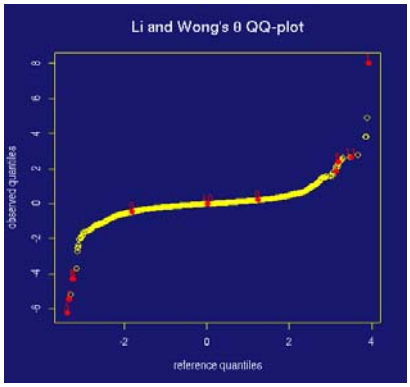
MAS-5



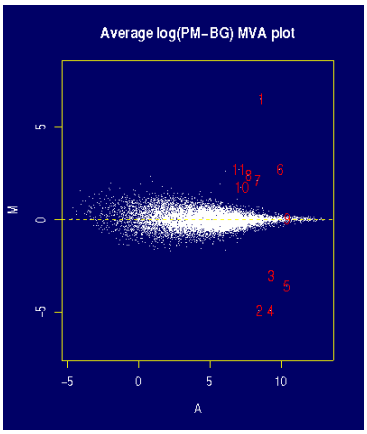


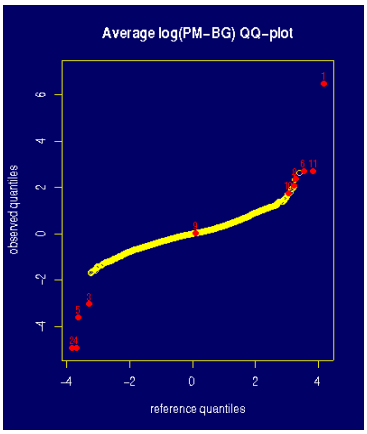
dChip (Li & Wong)



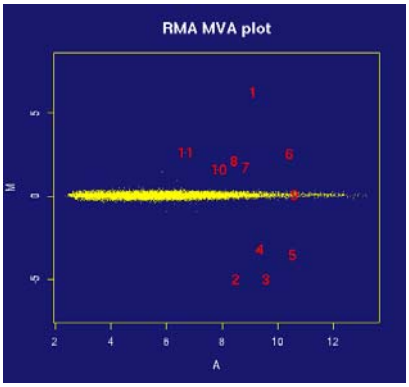


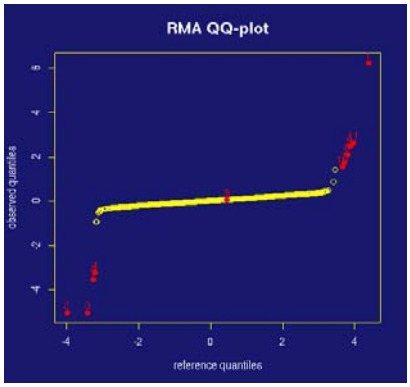
RMA (no median polish)





RMA (median Polish)



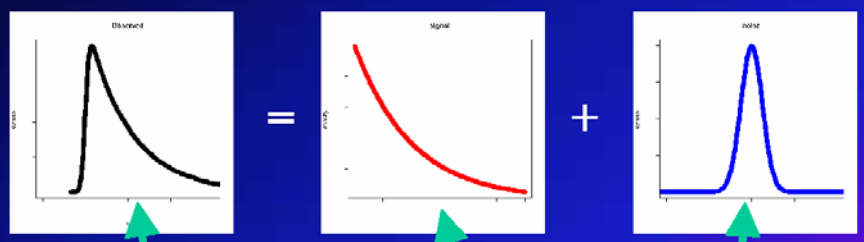


RMA background correction

- It assumes a model O = S + N,
- where S is an exponentially distributed (parameter α) signal, and N is a Gaussian noise with mean μ and standard deviation σ.
- Various truncation possibilities have also been suggested.
- The estimator used in all these papers is
- E [S | O=o].

RMA Background Approach

- Convolution Model



Observed
O

Signal
S
$Exp(\alpha)$

Noise
N
$N(\mu, \sigma^2)$

- $P(S=s) = \alpha \exp(-\alpha s)$, for $s \geq 0$
- $P(O = o \mid S = s) = \phi_\sigma(o - s - \mu)$
- The posterior is computed by
- $P(S = s \mid O = o)$

  $= P(S=s)\, P(O=o|S=s) / \int_0^1 ds\, P(S=s)\, P(O=o|S=s)$
- Numerator $= \alpha \exp(-\alpha s)\, \phi_\sigma(o - s - \mu)$

  $= \alpha / (\sqrt{2\pi}\, \sigma)$

  $\times \exp(-((s - (o - \mu))^2 + 2\sigma^2 \alpha s)/2\sigma^2)$
- Denominator $= \int_0^1 ds$ Numerator

- Thus, $P(S = s \mid O = o)$

  $= \phi_\sigma(s - a)/\Phi_\sigma(a)$

  $= o - \mu - \sigma^2 \alpha$

  $+ \sigma\, \phi((s - \mu - \sigma^2 \alpha)/\sigma)/\Phi((s - \mu - \sigma^2 \alpha)/\sigma)$

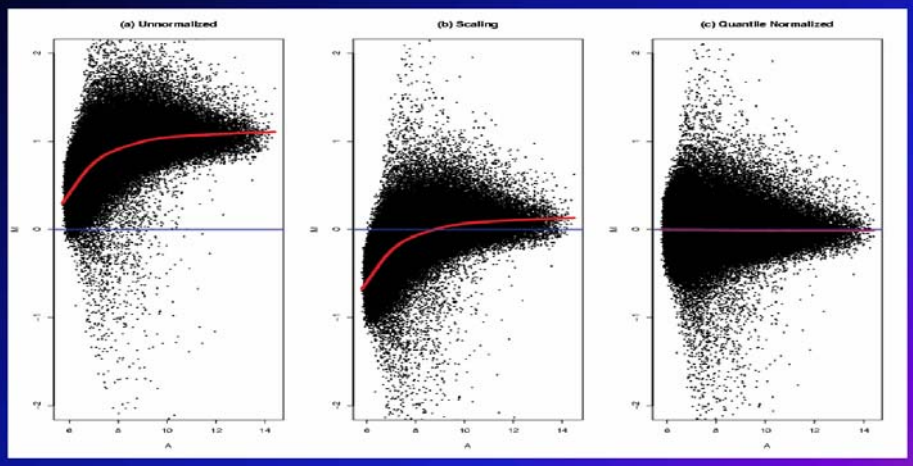
$= s -^* BG(\mu,\sigma)$

Correction is given by

$$E(S \mid O = o) = a + b \frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{o-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{o-a}{b}\right) - 1}$$

$$a = o - \mu - \sigma^2 \alpha, b = \sigma$$

Non-linear normalization needed

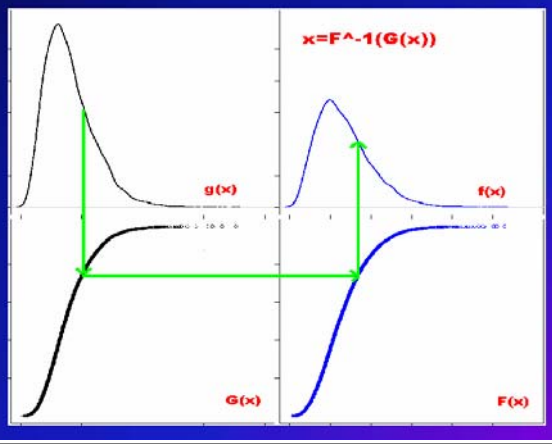


Unnormalized | Scaled | A Non-linear Normalization

Quantile Normalization

- Normalize so that the quantiles of each chip are equal. Simple and fast algorithm. Goal is to give same distribution to each chip.

Original Distribution



Target Distribution

Sort columns of original matrix

$$\begin{bmatrix} 1 & 5 & 3 & 5 \\ 2 & 1 & 6 & 7 \\ 3 & 2 & 2 & 6 \\ 4 & 6 & 1 & 8 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 1 & 5 \\ 2 & 2 & 2 & 6 \\ 3 & 5 & 3 & 7 \\ 4 & 6 & 6 & 8 \end{bmatrix}$$

Take averages across rows

$$\begin{bmatrix} 1 & 1 & 1 & 5 \\ 2 & 2 & 2 & 6 \\ 3 & 5 & 3 & 7 \\ 4 & 6 & 6 & 8 \end{bmatrix} \rightarrow \begin{bmatrix} 2 \\ 3 \\ 4.5 \\ 6 \end{bmatrix}$$

Set average as value for All elements in the row

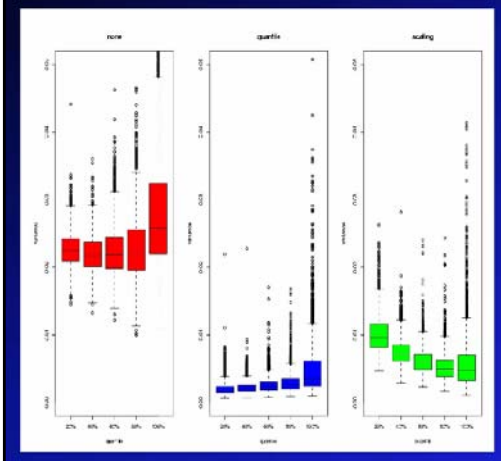
$$\begin{bmatrix} 2 \\ 3 \\ 4.5 \\ 6 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4.5 & 4.5 & 4.5 & 4.5 \\ 6 & 6 & 6 & 6 \end{bmatrix}$$

Unsort columns of matrix to original order

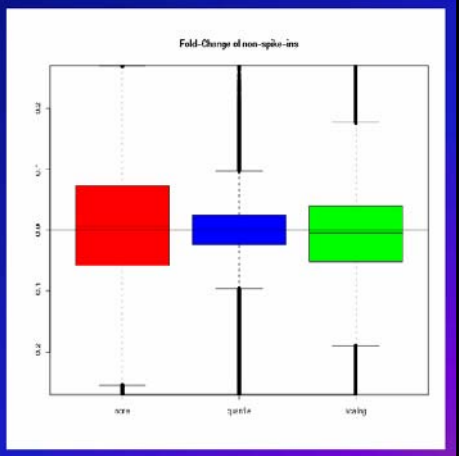
$$\begin{bmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4.5 & 4.5 & 4.5 & 4.5 \\ 6 & 6 & 6 & 6 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 4.5 & 4.5 & 2 \\ 3 & 2 & 6 & 4.5 \\ 4.5 & 3 & 3 & 3 \\ 6 & 6 & 2 & 6 \end{bmatrix}$$

It Reduces Variability

Expression Values



Fold change



Also no serious bias effects. For more see Bolstad et al (2003)

General Probe Level Model

$$y_{ij} = \mathrm{f}(\mathbf{X}) + \varepsilon_{ij}$$

- Where f(X) is function of factor (and possibly covariate) variables (our interest will be in linear functions)
- y_{ij} is a pre-processed probe intensity (usually log scale)
- Assume that $\mathrm{E}\left[\varepsilon_{ij}\right] = 0$

$$\mathrm{Var}\left[\varepsilon_{ij}\right] = \sigma^2$$

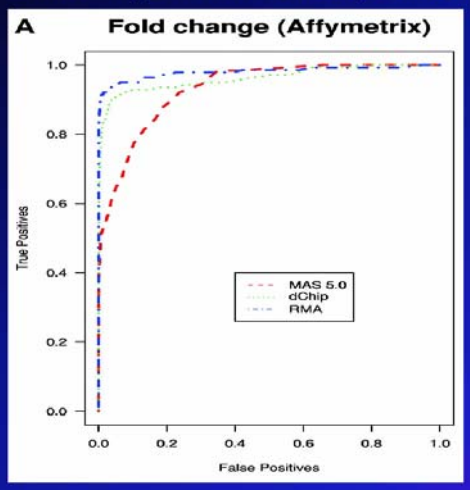
The Three Steps of RMA

1. Convolution Background
2. Quantile Normalization
3. Linear model on the log2 scale fit robustly.

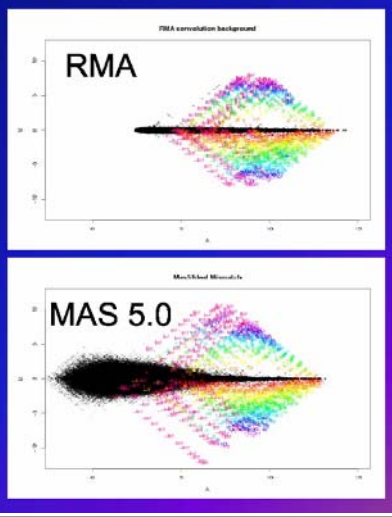
- Software for implementing RMA is in the Bioconductor *affy* package

RMA mostly does well in practice

Detecting Differential Expression



Not noisy in low intensities



The RMA model

$$y_{ij} = m + \alpha_i + \beta_j + \varepsilon_{ij}$$

where $y_{ij} = \log_2 \mathrm{N}\left(\mathrm{B}\left(PM_{ij}\right)\right)$

α_i is a probe-effect i= 1,…,I

β_j is chip-effect ($m+\beta_j$ is log2 gene expression on array *j)* j=1,…,J

Median Polish Algorithm

$$\begin{array}{ccc|c} y_{11} & \cdots & y_{1J} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ y_{I1} & \cdots & y_{IJ} & 0 \\ \hline 0 & \cdots & 0 & 0 \end{array}$$

Sweep Rows

Sweep Columns

Iterate

$$\begin{array}{ccc|c} \hat{\varepsilon}_{11} & \cdots & \hat{\varepsilon}_{1J} & \hat{\alpha}_1 \\ \vdots & \ddots & \vdots & \vdots \\ \hat{\varepsilon}_{I1} & \cdots & \hat{\varepsilon}_{IJ} & \hat{\alpha}_I \\ \hline \hat{\beta}_1 & \cdots & \hat{\beta}_J & \hat{m} \end{array}$$

Imposes Constraints

$$\text{median}\,\alpha_i = \text{median}\,\beta_j = 0$$

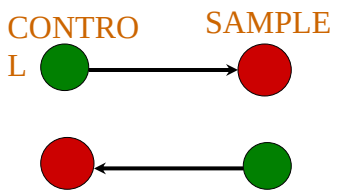
$$\text{median}_i\,\varepsilon_{ij} = \text{median}_j\,\varepsilon_{ij} = 0$$

Median Polish

- Advantages
  - Fast
  - Very robust
- Disadvantages
  - No algorithmic flexibility to fit alternative models
  - No standard error estimates

Within-slide Normalizations

Dye Bias



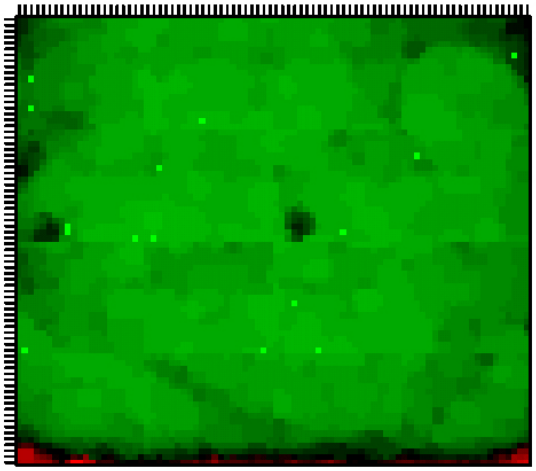
- Dye Bias
  - Two-channel microarrays
  - Intensity in one channel is higher than other
  - Dye swapping experiments
- Additionally, can be normalized
  - Take sum intensities for each signal
  - Normalize sums: Assumes most genes regulated at same level

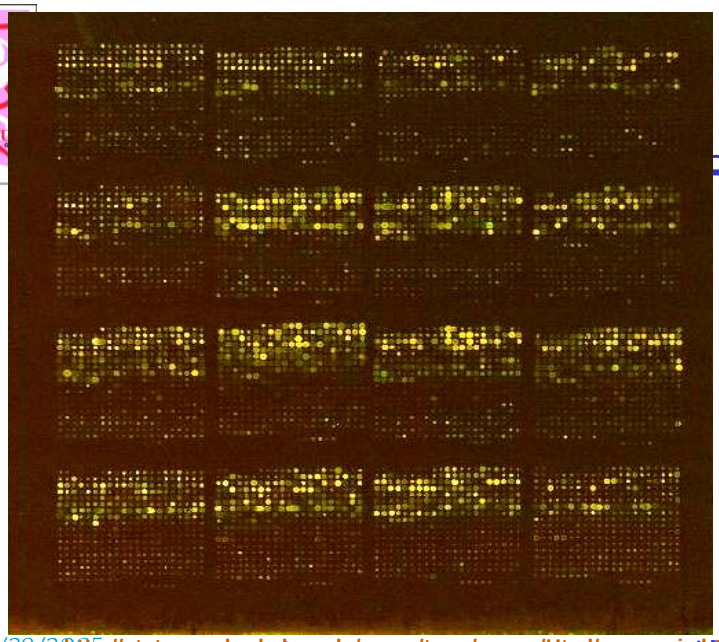
Spatial Normalization

- Signal varies according to spot location
  - Particularly, corners
    - Less hybridization solution
    - Susceptible to desiccation
  - Chip design
    - DO NOT cluster genes with similar expression profiles

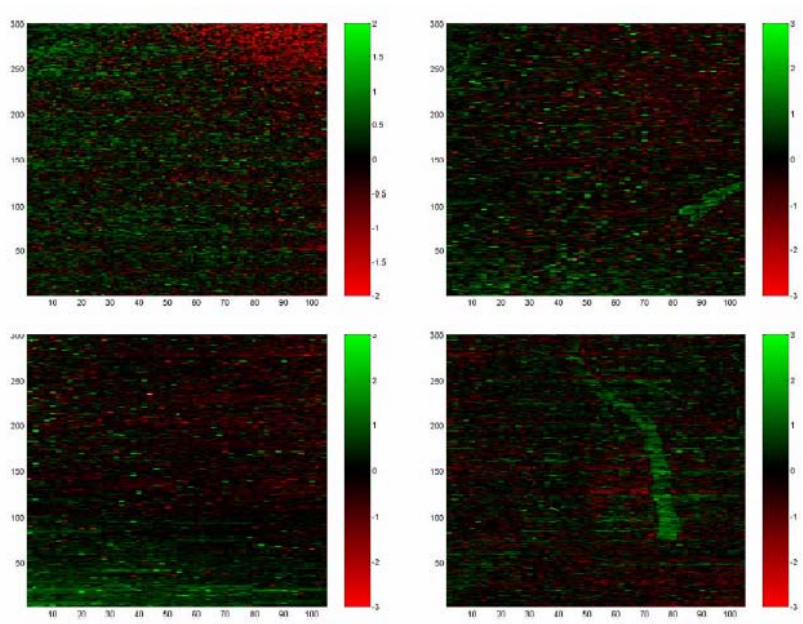
Spatial Bias

Source: http://www.csc.fi/oppaat/siru/sirupartII.pdf





Source: http://stat-www.berkeley.edu/users/terry/zarray/Html/normspie.html



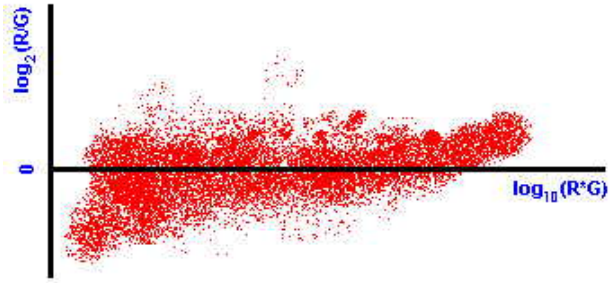
Source: http://www.cse.ucsc.edu/classes/bme210w/Winter04/lectures/Bio210w04-Lect06b-ComputationalNormalization.pdf

Intensity Dependent Biases

- Low intensities have much greater variation

RI Plots

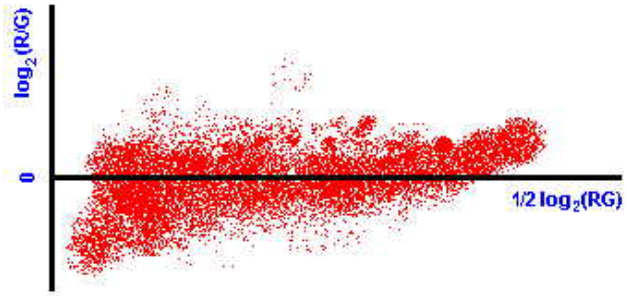
- Ratio-Intensity
- R: $\log_2(R/G)$
- I: $\log_{10}(R*G)$



http://www.ucl.ac.uk/oncology/MicroCore/HTML_resource/

MA Plots

- M: $\log_2(R/G)$
- A: $\log_2 SQRT(R*G) = ½ \log_2(R*G)$

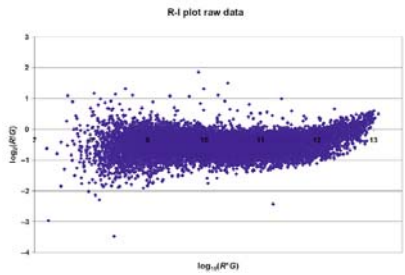


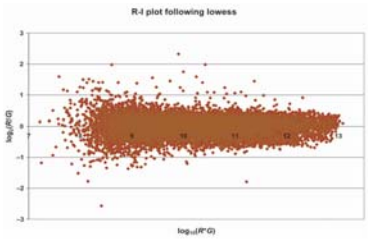
http://www.ucl.ac.uk/oncology/MicroCore/HTML_resource/MA_plots_popup.htm

Lowess (Loess) Normalization

- Locally Weighted Linear Regression
- Linearises Data

Loess Normalization





Figures from Quackenbush, 2002

Cross-slide Normalizations

- Comparisons between chips needed
- Slides normalized so comparisons can be made

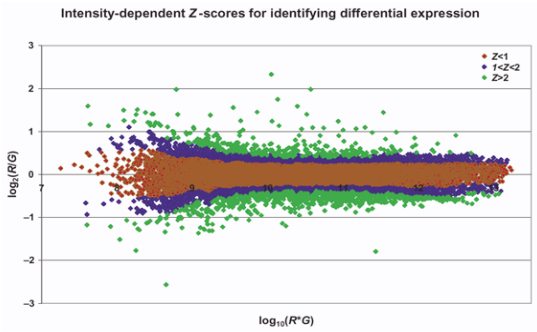
Per-Chip Normalization

- Mean/Median centering – mean/median intensity of every chip brought to same level
- Total intensity normalization – scaling factor determined by summing intensities
- Spiked-control, housekeeping normalization

Differential Expression

- Crude filter
  - Genes over/underexpressed by a factor of two
  - Log_2 values of 1 and -1
  - Plus: Calculation very easy
  - Minus: Does not consider reliability of data

Localized Z-Scores



Figures from Quackenbush, 2002

Two sample t-test

- Calculates probability values sampled from same distribution
  - Considers mean, variance

Analysis Tools

- R, Bioconductor
- S+, ArrayAnalyzer
- Affymetrix Tools
- GeneSpring

MIAME Data

- MIAME: "Minimum Information About a Microarray Experiment"
  - Specifies content, not format
  - Specifies type of data to be published
- MIAME Checklist
  - Experiment design
  - Samples used; extract preparation and labeling
  - Hybridization procedures and parameters
  - Measurement data and specifications
  - Array design

MAGE-ML

- MicroArray Gene Expression Markup Language
- XML based
  - Object model
  - Document exchange format
  - Software toolkits

Repositories

- ArrayExpress (EBI)
- Gene Expression Omnibus (NCBI)
- Stanford Microarray Database (SMD)

- Microarray databases
  - AMAD: Another Microarray Database
  - LONGHORN:
  - MIDAS:
  - BASE:

Proteomics

Beyond Genomics

- Human Genome
  - 30,000 to 60,000 genes
- Human Proteome
  - 300,000 to 1,200,000 protein variants
- Human Metabalome
  - Metabolic products of the organism (lipids,carbohydrates, amino acids, peptides, prostaglandins, etc)

Proteome

- Proteome: The entire protein complement in a given cell, tissue or organism.
  - Protein Activities
  - 3D Structure
  - Modifications and Localization
  - Protein-Protein Interaction:Proteins in Complexes
  - Protein Profile: Global patterns of protein content and activity (particularly in response to a disease state.)
  - Understanding system-level cellular behavior

Technology & Databases

- Identify proteins and protein complexes in biological samples comprehensively and quantitatively with both high sensitivity and fidelity.
  - Develop new diagnostic markers
  - Identification of new drug-target
- HUPO (Human Proteome Organization)
  - Coordinating proteomics projects worldwide.

Integration

- Complementary to other functional genomic approaches:
  - Micro-array based expression profiles
  - Systematic phenotypic profiles at the cell and organism level
  - Systematic genetics
  - Small-molecule-based arrays

Applications of Proteomics

- Protein Mining
  - Catalog all the proteins present in a tissue, cell, organelle, etc.
- Differential Expression Profiling
  - Identification of proteins in a sample as a function of a particular state: differentiation, stage of development, disease state, response to drug or stimulus
- Network Mapping
  - Identification of proteins in functional networks: biosynthetic pathways, signal transduction pathways, multiprotein complexes
- Mapping Protein Modifications
  - Characterization of posttranslational modifications: phosphorylation, glycosylation, oxidation, etc.

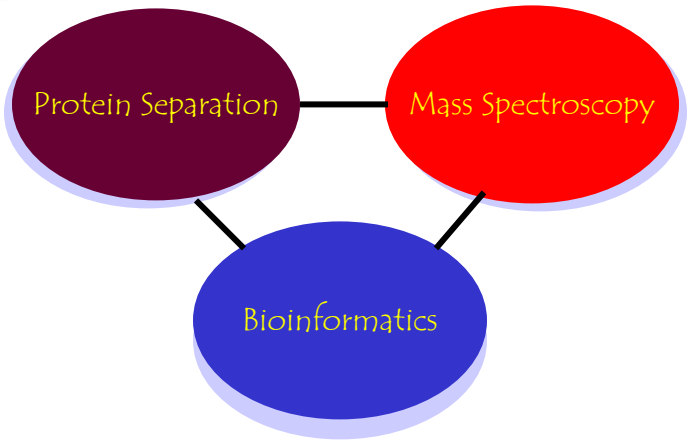
Challenges of Proteomics

- Limited and Variable Sample Material
- Sample Degradation
- Vast Dynamic Range
  - (more than 10^6-fold for protein abundance)
- Post-translational Modifications
- Unlimited tissue, developmental and temporal specificty
- Disease and drug perturbation.

Proteomics Technologies

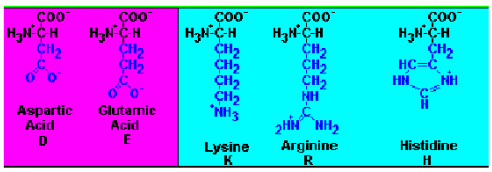
- Development of genome and protein sequence databases
  - Bioinformatics and Data mining software
- Development of mass spectrometry instrumentation suitable to analyze biomolecules
  - Protein mass, Peptide mass, Peptide sequence
- Development of analytical protein separation technology
  - IEF, 2D-SDS-PAGE, HPLC, Capillary Electrophoesis, Affinity Chromatography

Components of Proteomics



2 D Electrophoresis

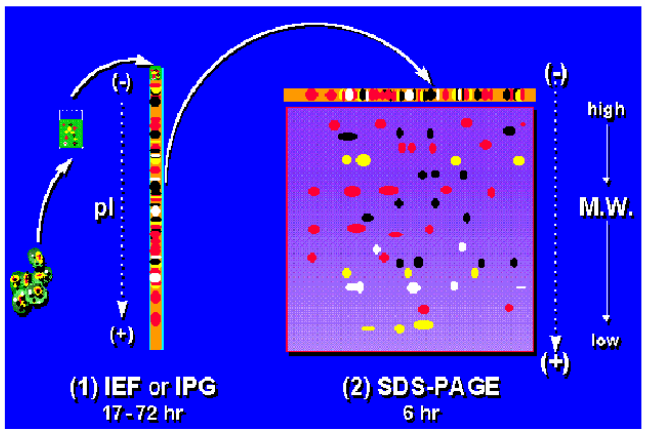
- Property of proteins
  - Some amino acids are acidic / basic (donate / accept H+)
  - Collection of amino acids in protein determines its pI value
    - pI = pH at which molecular charge = zero
- 2D electrophoresis
  - Separate proteins according to both pI & molecular weight



2 D Electrophoresis

- Method
- 1. Extract & prepare protein sample in solution
- 2. Separate proteins (in each dimension)
  - I. Based on pH
    - Using isoelectric focusing (IEF)
    - Using immobilized pH gradient (IPG) strips
  - II. Based on molecular weight (size)
    - Using gel electrophoresis

2 D Electrophoresis



Proteomic Technology

Sequence the proteome

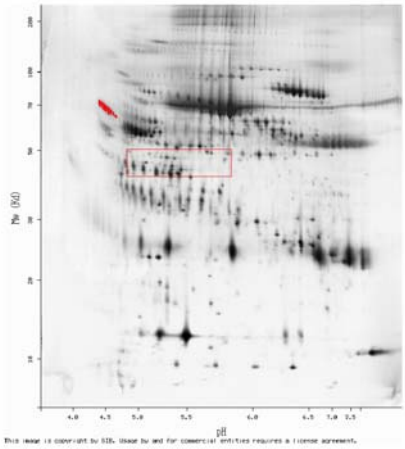
Tissue

Isolate proteins

Run a 2 D SDS-PAGE

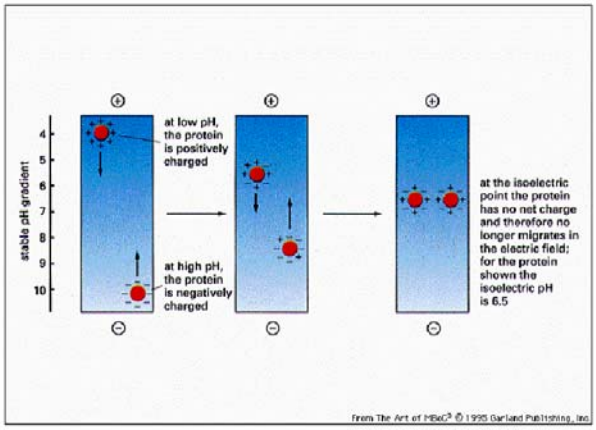
Isolate single protein dots

Sequence the protein

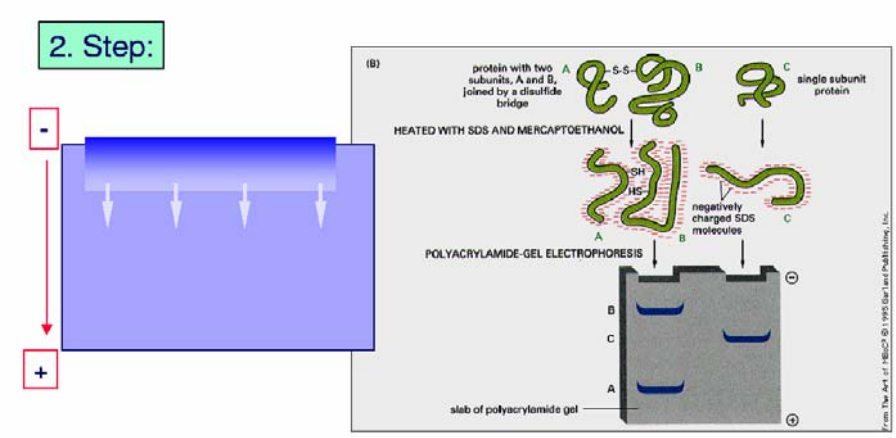


2-D SDS PAGE

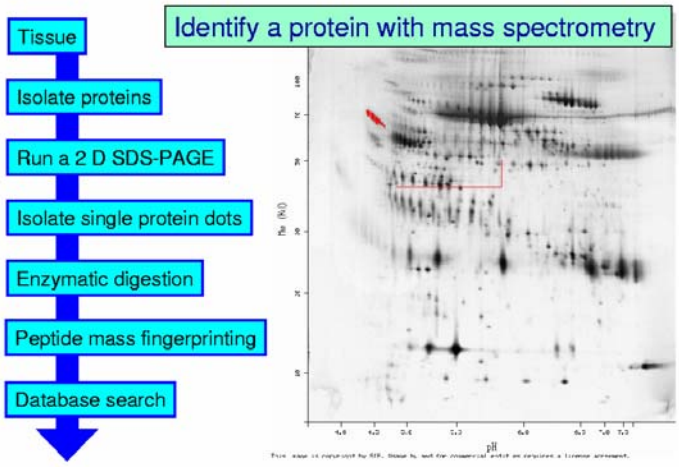
1. Step:



2D SDS PAGE



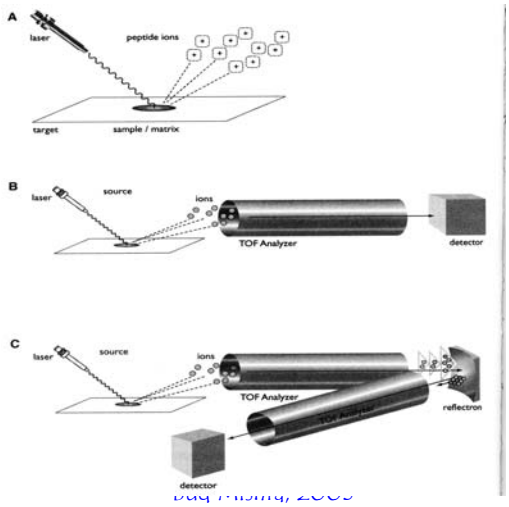
Integrating with Mass-Spec



Mass Spectrometry

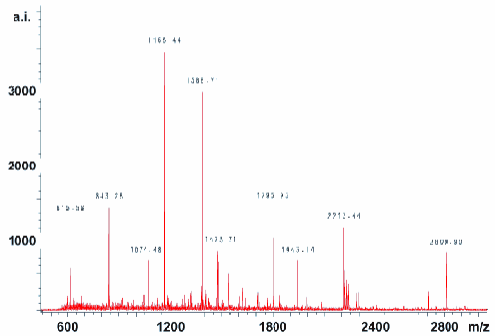
- Method
  - 1. Excise individual dots from 2D electrophoresis
  - 2. Digest protein into fragments with enzyme (e.g., trypsin)
  - 3. Ionize protein fragments (without breaking)
    - Matrix Assisted Laser Desorption Ionization (MALDI)
    - Electrospray Ionization (ESI)
  - 4. Accelerate through mass spectrometer
  - 5. Produces peptide mass fingerprint

Principles of MALDI-TOF Mass Spectroscopy



Fingerprint

- **Peptide mass fingerprint**



Identifying peptide mass fingerprint

- Compare with Fingerprint for actual protein in database
  - Predicted fingerprint for predicted / hypothetical protein (Precompute for efficiency)
  - May fail to distinguish Post-translation modifications to protein
- Protein databases / web servers (e.g., SWISS-2D PAGE)
  - For each protein, record its
  - (1) Protein pI, molecular weight, peptide mass fingerprint…
  - (2) Experimentally determined location in 2D gel

Peptide Chips

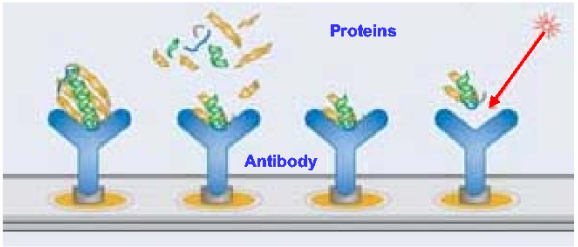
- Protein-protein interaction,
- Unravel signal transduction pathways,
- Perform multi-parameter diagnosis,
- Study individual immunological repertoires
  - e.g. autoimmune reactions.

Peptide Chips

- Goal
  - High-throughput analysis of protein expression / interaction
  - Adapt approach similar to DNA microarrays
  - Improves on speed vs. 2D electrophoresis
- Approach
  - No equivalent of hybridization for proteins
  - Exploit other biochemical binding reactions
    - Antibody–antigen
    - Receptor–ligand
    - DNA–protein…

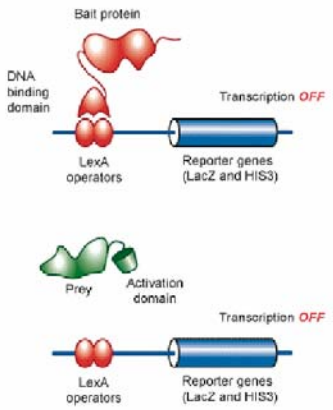
Ciphergen

Ciphergen Antibody Capture Protein Chip

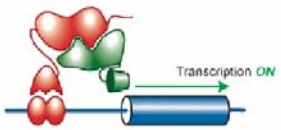


1) Protein with antigen bound to antibody probe	2) Remainder of protein digested enzyme, leaving peptide antigen	3) Wash away protein fragments	4) SELDI laser ionizes & desorbs epitope binding peptide, sends to mass spectrometer

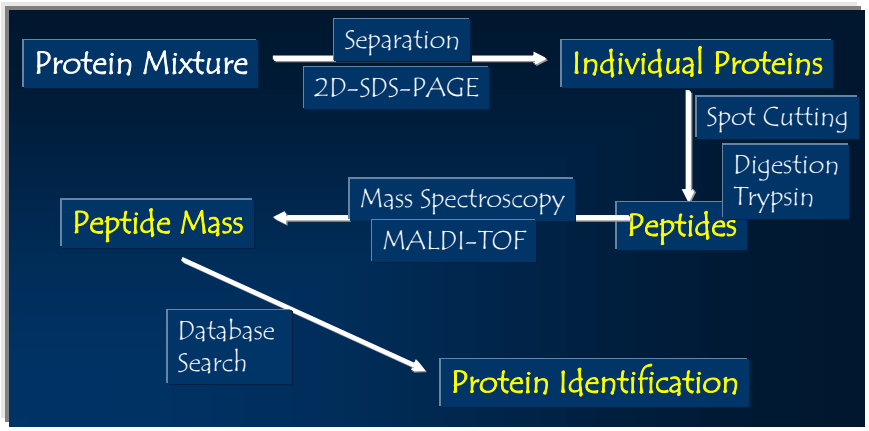
Protein-Protein Interaction



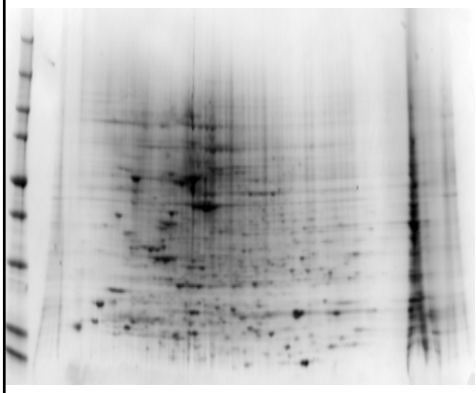
Yeast two hybrid system



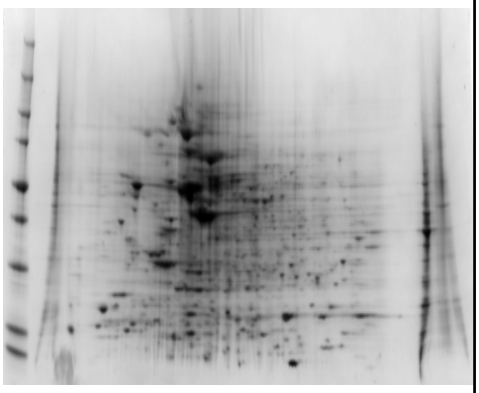
Basic Proteomic Analysis Scheme



2D-SDS-PAGE of 2 Types of Cells

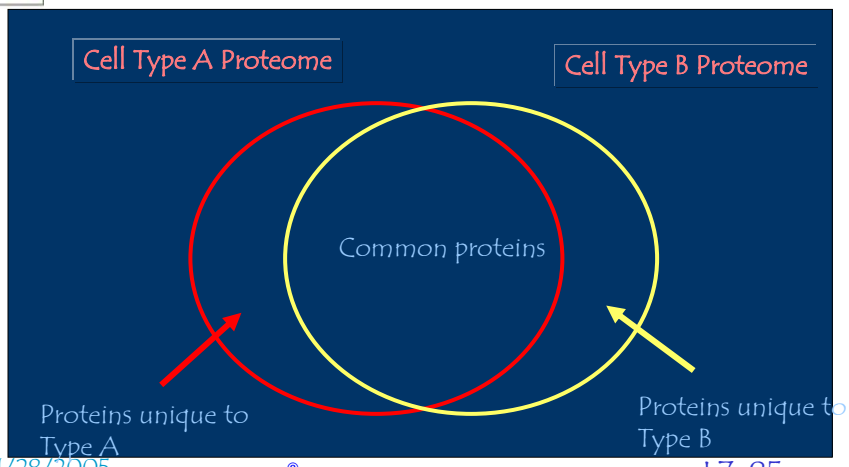


Cell Type A

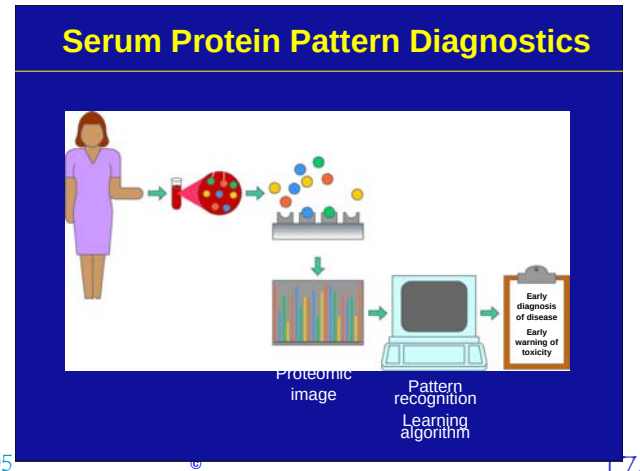


Cell Type B

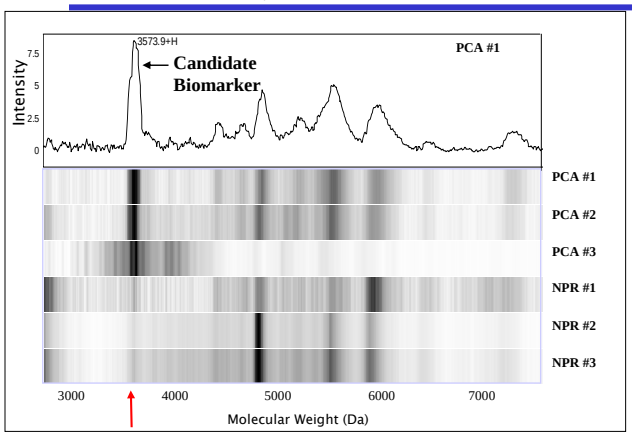
Differential Expression



Clinical Diagnostics Proteomics: Protein Profiling



Protein Profiles of (3 patients)



To be continued…

…