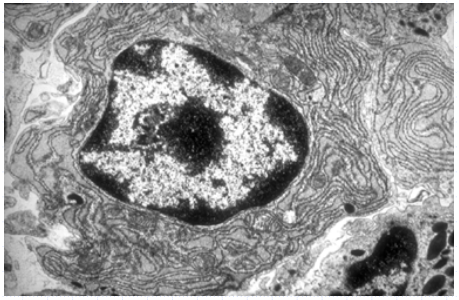
Exploiting the Full Potential of Microarray Data

X Liu*, V Vinciotti, K Fraser, S Swift and A Tucker
*Leiden Institute of Advanced Computer Science,
Leiden University; On leave from
School of Computing, Information Systems and Mathematics
Brunel University
www.ida-research.net; hui@ida-research.net

Xiaohui Liu

Cell Nucleus: Where the genes are.

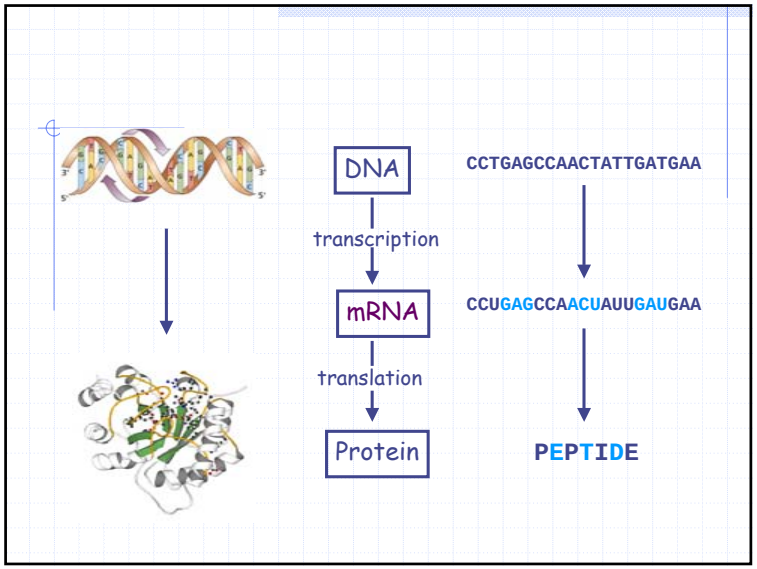


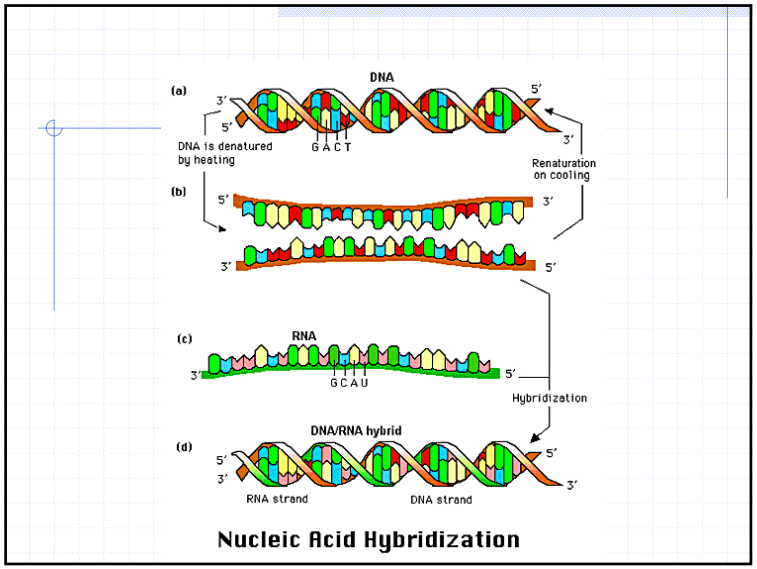
www.grad.ttuhsc.edu/courses/histo
Texas Tech University Health Sciences Center

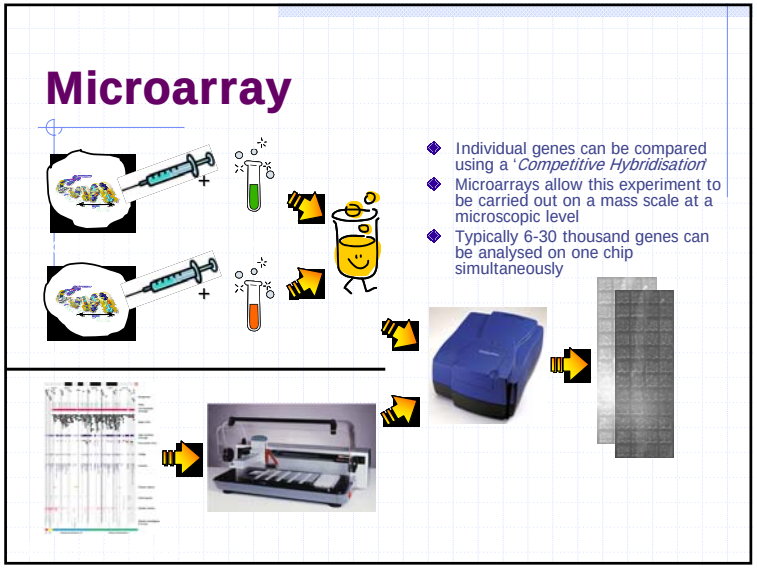
Genes are DNA sequences

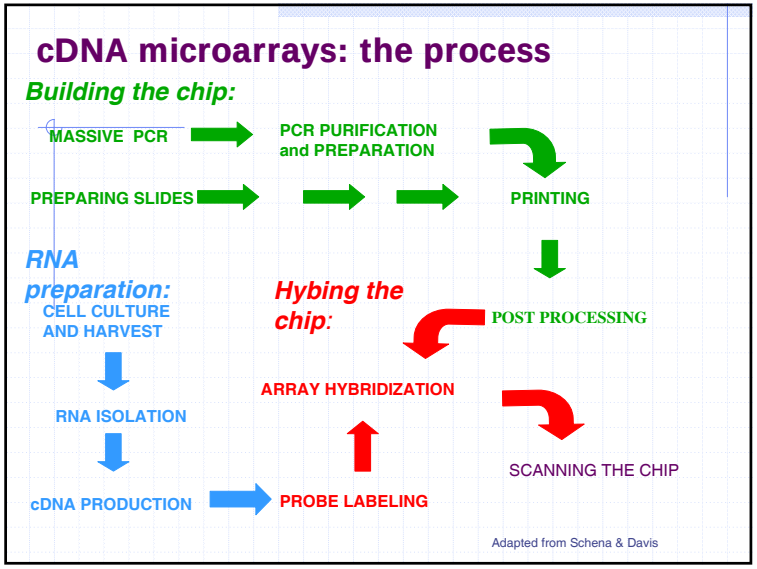
```
DEFINITION  Human breast cancer susceptibility (BRCA2) mRNA, complete cds.
  ORGANISM  Homo sapiens
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Eutheria; Primates; Catarrhini; Hominidae; Homo.
FEATURES             Location/Qualifiers
     source          /map="13q12-q13"
                     /chromosome="13"
     CDS             229..10485
                     /gene="BRCA2"
                     /codon_start=1
                     /product="BRCA2"
/gene="BRCA2"
ORIGIN
    1 ggtggcgcga gcttctgaaa ctaggcggca gaggcggagc cgctgtggca ctgctgcgcc
   61 tctgctgcgc ctcgggtgtc ttttgcggcg gtgggtcgcc gccgggagaa gcgtgagggg
  121 ctcgggtgtc ggtggcgcga gaggcggagc cgctgtggca atccaaactc gccgggagaa
[180 lines deleted]
10921 ttacaatcaa caaaatggtc atccaaactc aaacttgaga aaatatcttg ctttcaaatt
10981 gacacta
```

www.ncbi.nlm.nih.gov/Entrez

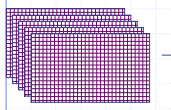

DNA
transcription
mRNA
translation
Protein
CCTGAGCCAACTATTGATGAA
CCUGAGCCAACUAUUGAUGAA
PEPTIDE


(a)
DNA
3′
5′
5′
3′
DNA is denatured by heating
G A C T
Renaturation on cooling
(b)
5′
3′
3′
5′
(c)
RNA
3′
5′
G C A U
Hybridization
(d)
DNA/RNA hybrid
5′
3′
5′
3′
RNA strand
DNA strand
Nucleic Acid Hybridization


Microarray
Individual genes can be compared using a 'Competitive Hybridisation'
Microarrays allow this experiment to be carried out on a mass scale at a microscopic level
Typically 6-30 thousand genes can be analysed on one chip simultaneously


cDNA microarrays: the process
Building the chip:
MASSIVE PCR
PCR PURIFICATION and PREPARATION
PREPARING SLIDES
PRINTING
RNA preparation:
CELL CULTURE AND HARVEST
RNA ISOLATION
cDNA PRODUCTION
PROBE LABELING
Hybing the chip:
POST PROCESSING
ARRAY HYBRIDIZATION
SCANNING THE CHIP
Adapted from Schena & Davis

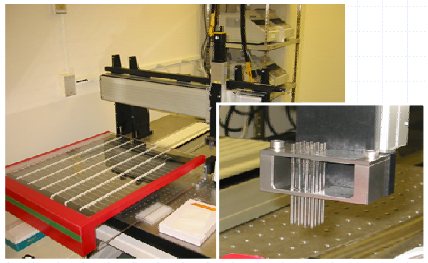
cDNA microarrays: Building the chip



Arrayed Library
(96 or 384-well plates)

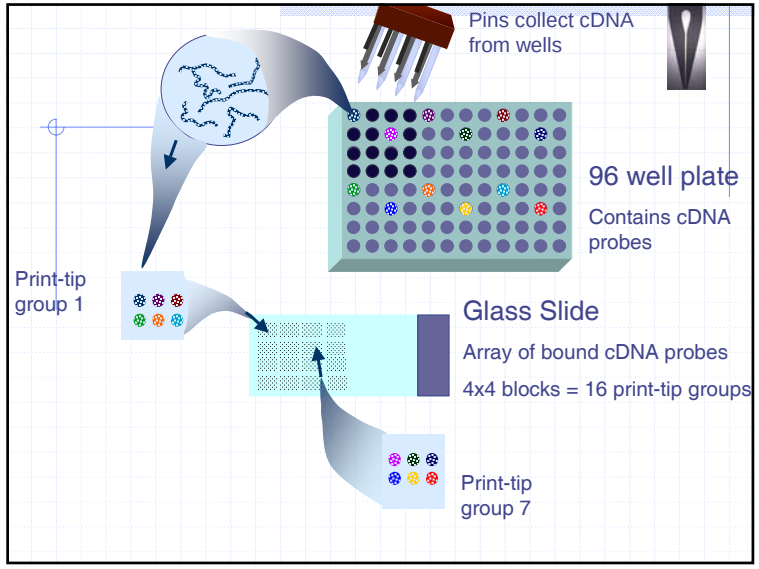
PCR amplification

Consolidate into well plates

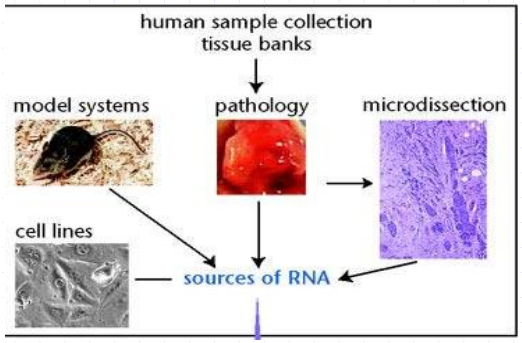


Spot as microarray on glass slides

(Ngai Lab, UC Berkeley)



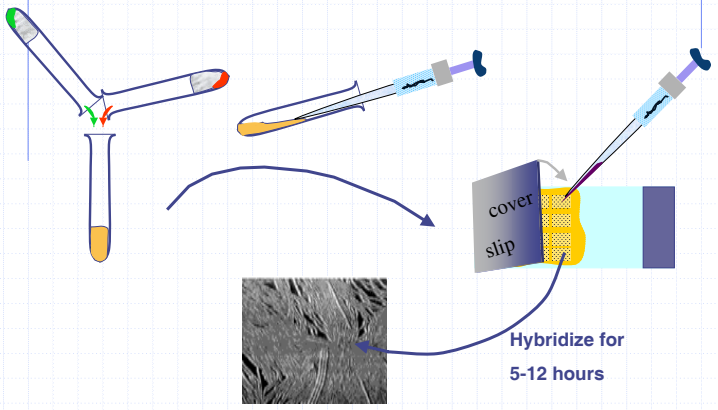
cDNA microarrays: RNA preparation



mRNA is reverse-transcribed into cDNA and labelled.

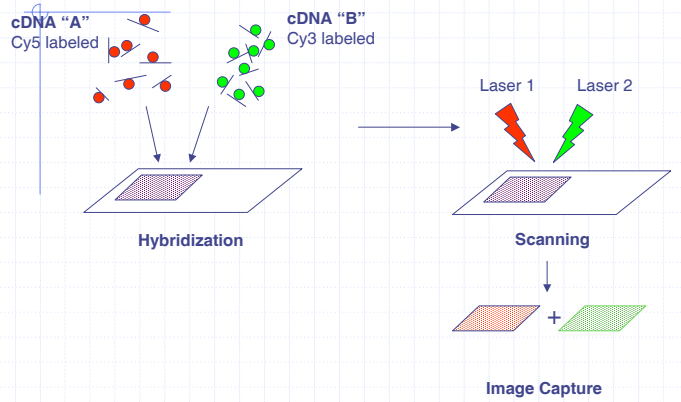
cDNA microarrays: Hybing the chip

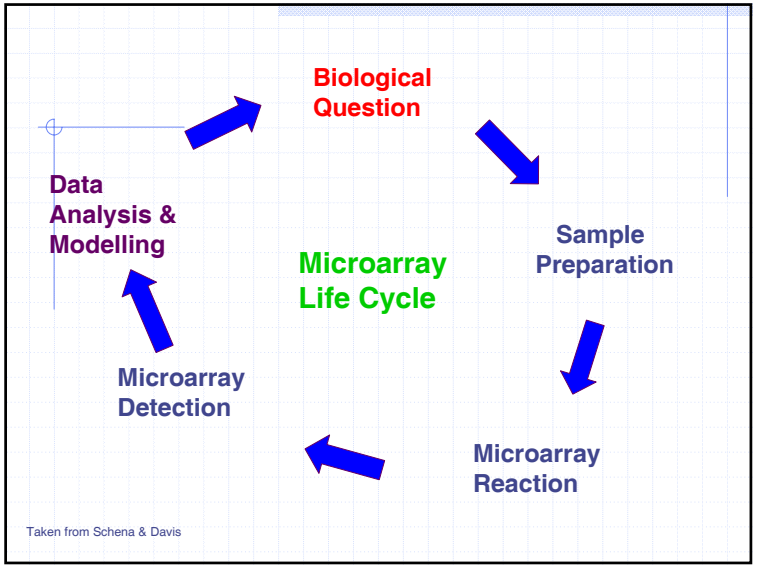
Hybridizing of labelled cDNA target samples to cDNA probes on the slide

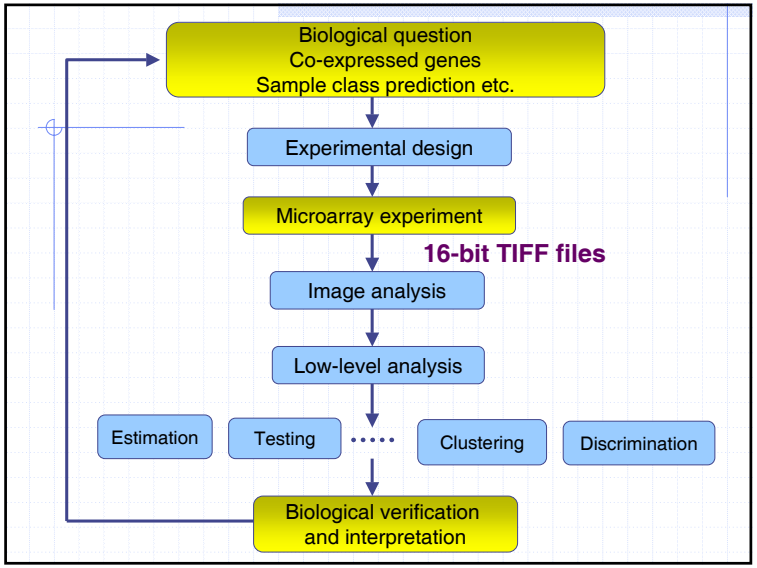


Hybridize for
5-12 hours

cDNA microarrays: brief summary







Microarray gene expression Data

Gene expression data on p genes for n samples

mRNA samples

Genes	sample1	sample2	sample3	sample4	sample5	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

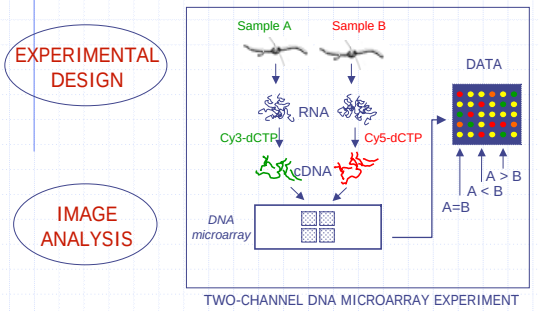
Gene expression level of gene i in mRNA sample j

= Log(Red intensity / Green intensity), or

Log(Signal)

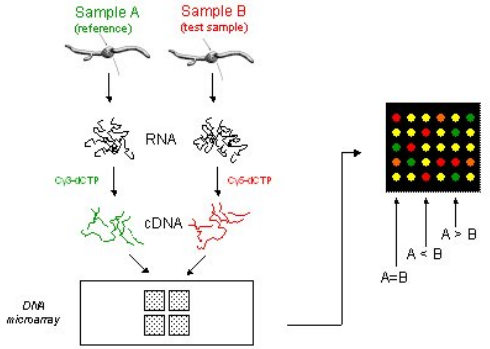
Veronica Vinciotti

From experimental design to gene networks



TWO-CHANNEL DNA MICROARRAY EXPERIMENT

Experimental Design

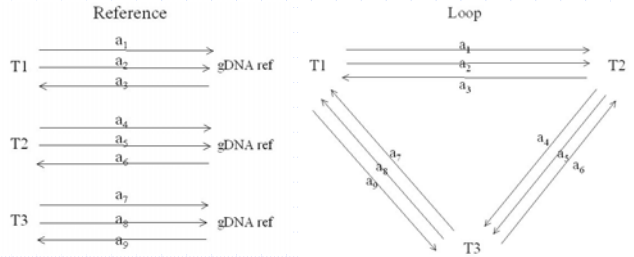


Experimental Design Issues

- Technical variation
  - Replicated genes on the slide
- Biological variation
  - Samples from different individuals
- How to allocate samples to arrays
  - Which two conditions should be compared on one array?

Choice of Design

- **Question**: Given number of conditions (e.g. time points) we wish to compare and a number of arrays we can afford to make, what is the most efficient design?



Which Design is Best?

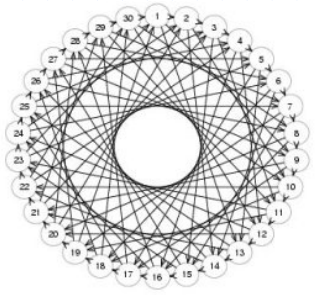
- Loop-type of designs have been shown to be more efficient than reference designs
  - Both theoretically and experimentally
- Loop designs allocate the resources more efficiently to compare the conditions of interest
  - In a reference design, 50% of the resources are used on a reference sample, of little interest to biologists

However…

- The data that come out of loop-type of designs are less intuitive
  - One can use a simple linear model to obtain estimates of the contrasts
- How to extend loop designs to large studies
  - Comparing all possible pairs of conditions becomes unrealistic for large studies
- How to measure the efficiency of a design
  - The design should provide precise estimates of the parameters of interest
  - The design should be robust to the situation when arrays get missing/damaged or the experiment has to be extended in future

Design for Large Studies

(Vinciotti et al, 2004, *Bioinformatics*)



A–Optimality Score for Contrasts: Tr[Inv(X'X)] = 159.4576 .
Nr of arrays = 90 . Nr of conditions = 30

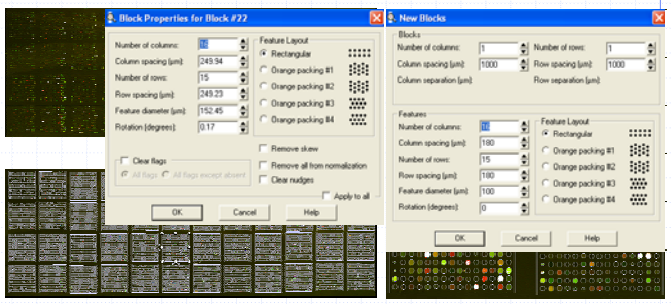
Karl Fraser

Current Image Processing Techniques

- Current techniques rely upon operator assistance and prior knowledge
- At present, no one method has been successful in blindly processing a slide with excess noise
- Rather than focus on one technique, we instead propose an adaptable framework which can be developed to combine multiple techniques

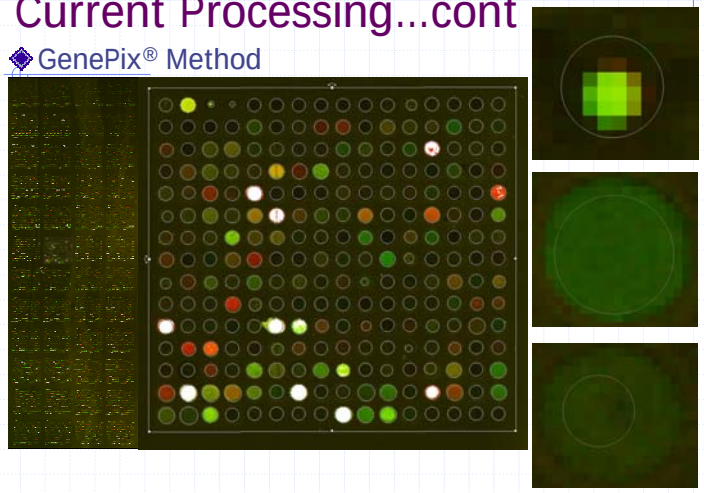
Current Processing

- GenePix® Method

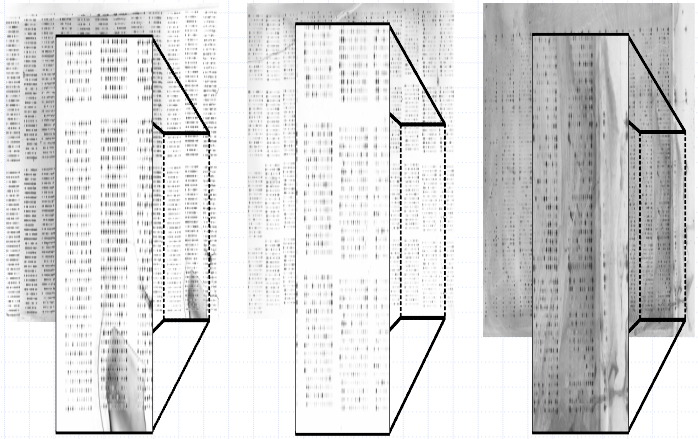


Current Processing...cont

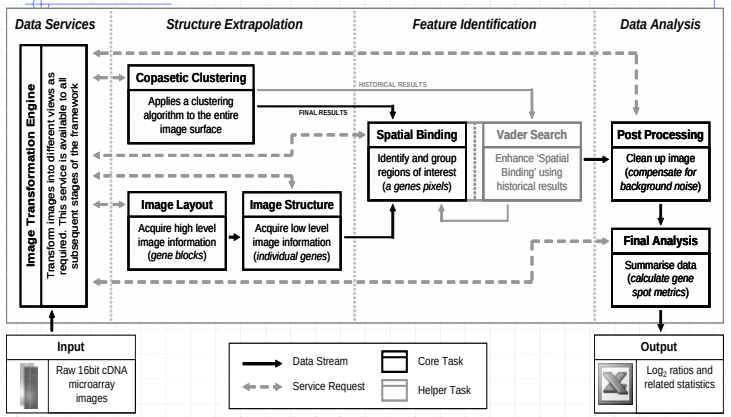
- GenePix® Method



Problems

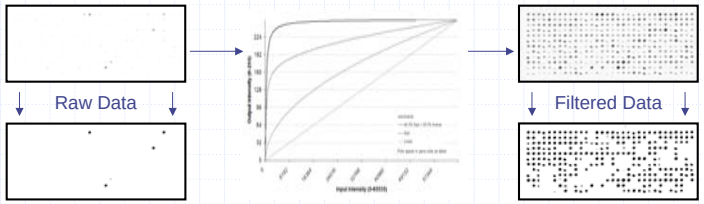


Copasetic Analysis



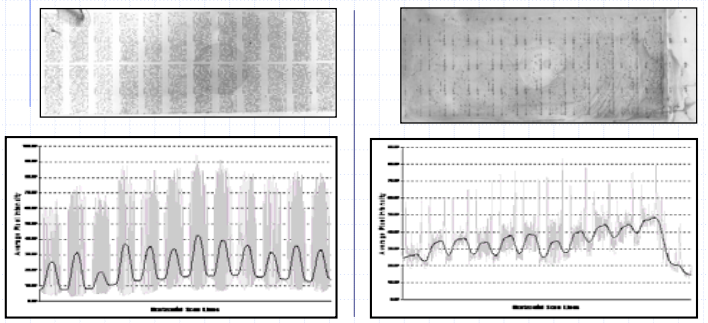
ITE: Data Filtering

- Analysing the raw data may not be beneficial
- Filtering can clean, emphasis genes
- For example, input-output response curves

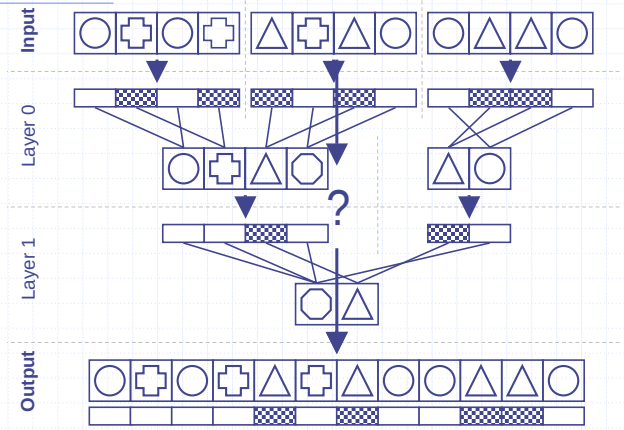


ITE: Data Transformations

- Sometimes a different perspective can help…

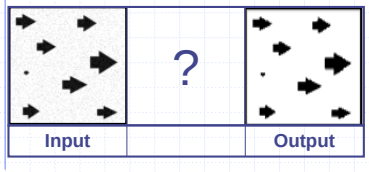


CC 2: Conceptual Overview



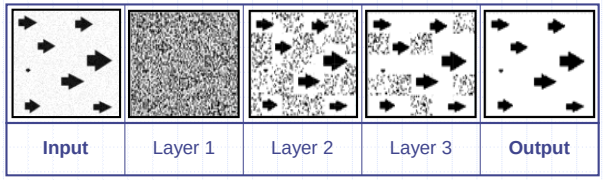
CC: Standard Clustering

Standard Clustering



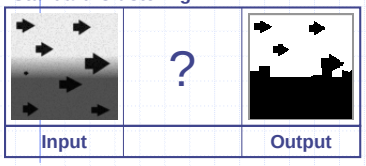
- Clustering is performed locally
- Pyramidic process used to combine results
- Standard clustering approach is unfeasible due to large datasets

Copasetic Clustering

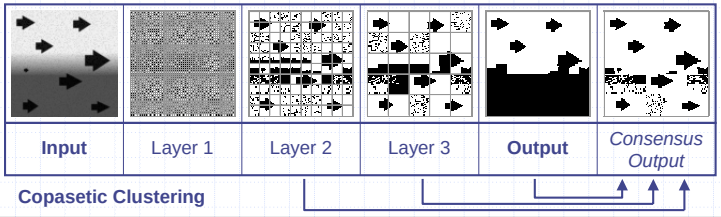


CC: Historical Information

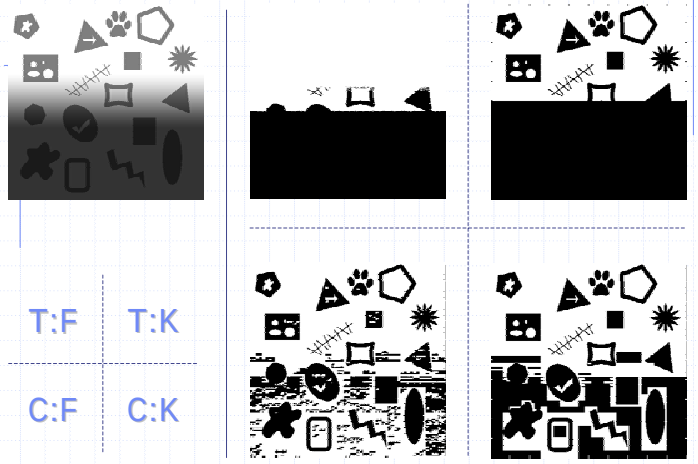
Standard Clustering



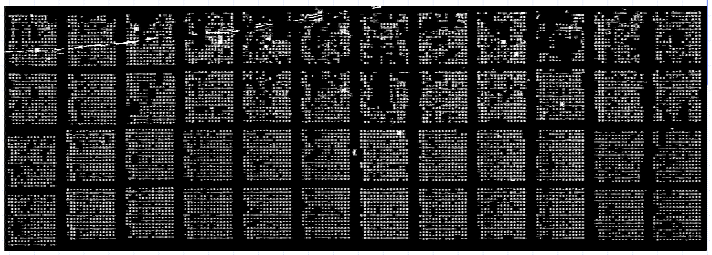
- Search is biased to local regions
- This is very powerful when combined with the historical information saved from different levels
- Still makes use of traditional techniques



Example process

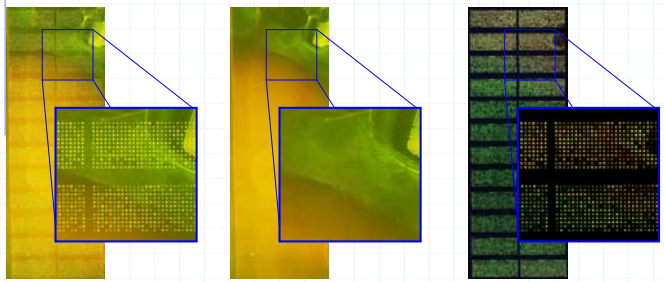


CC: Microarray Results



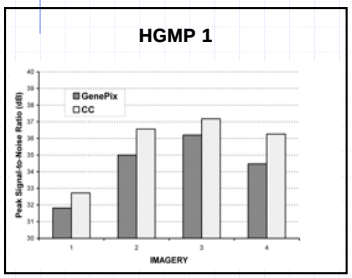
- A microarray slide that contains ~10 million observations (1.2M FG)
- Black squares show regions where extreme values have distorted local area

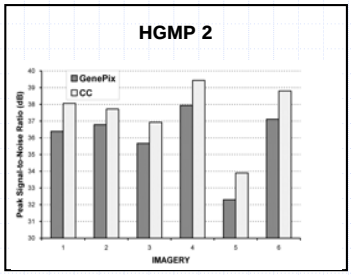
Post-Processing & Final Analysis



Overall Results

- Provided a 1 – 3dB (PSNR) improvement over GenePix® as used by an expert operator



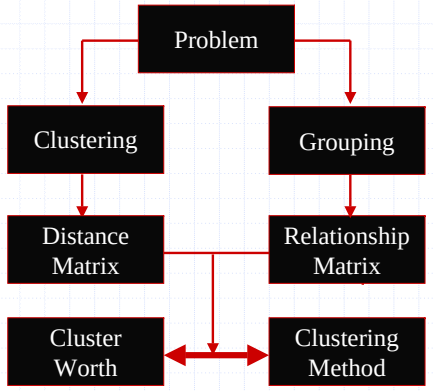


Stephen Swift

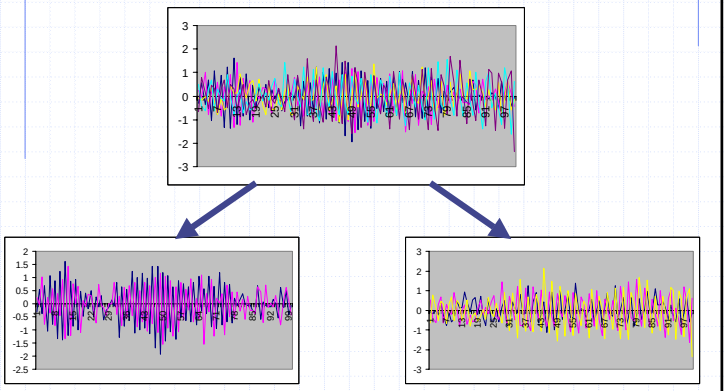
Clustering and Grouping (1)

- Clustering
  - Arranging Objects (as Points) into Sets According to “Distance” on a Hyper-Graph
- Grouping
  - Arranging Objects into Sets According to Some Inter-Object Relationship
- Each Set is Usually Mutually Exclusive
- Will Not Consider “Fuzzy” Clustering

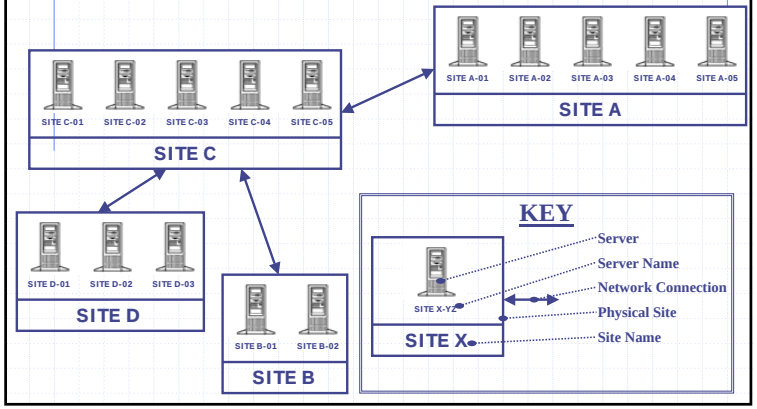
Clustering and Grouping (2)



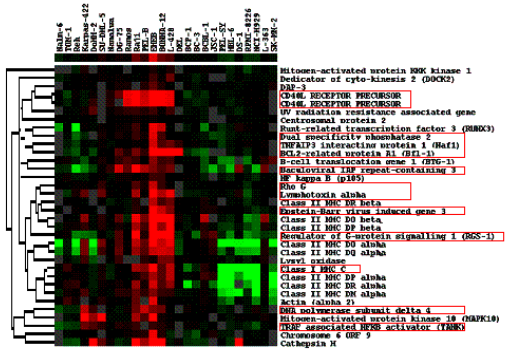
Application 1 – MTS Decomposition



Application 2 – Email Logfiles



Application 3 – Microarrays



Vectors and Distances (1)

- Many Methods are Designed to Work on Distance Metrics, e.g. K-Means
- They Assume that the "Triangle Inequality" Holds
- This is NOT the Case for Many Applications, e.g. MTS Decomposition Using Cross-Correlation
- More General "Grouping" Methods Must be Chosen

Vectors and Distances (2)

- Distance Matrices
  - Euclidean
  - Correlation
  - Minkowski
  - Manhattan
  - Mahalanobis
- Relationship Matrices
  - How Long is a Piece of String?
  - Often Application Dependant

Cluster Worth (1)

- The Choice of Correct Metric for Judging the Worth of a Clustering Arrangement is Vital for Success
- There are as Many Metrics as Methods!
- Each has Their Own Merits and Drawbacks

Cluster Worth (2)

- Sum of Squares by Cluster
- Homogeneity (H)
- Separation (S)
- H/S
- Maximum Likelihood
- Minimum Description Length
- Etc...

The Number of Clusters

- Many Applications Specify the Number of Clusters a Solution Requires, e.g. the Email Server Application
- Many Do Not, e.g. Microarray Data
- Determining the Number of Clusters is Very Difficult
- A Choice of Method that Locates the Number of Clusters and Their Contents is Often Desirable

Methods

- Statistical
  - K-Means
  - Hierarchical
  - PAM
- Optimisation / Search /AI
  - Evolutionary Computing
  - SOM
  - Hill Climbing and Simulated Annealing
- KDD and Others, e.g. CLARIS, EM

Comparing Clusters and Methods

	H40	K40	SOM40	HC40	SA40
H40	-	0.609	0.041	0.640	0.647
K40	-	-	0.053	0.536	0.540
SOM40	-	-	-	0.082	0.074
HC40	-	-	-	-	0.879
SA40	-	-	-	-	-

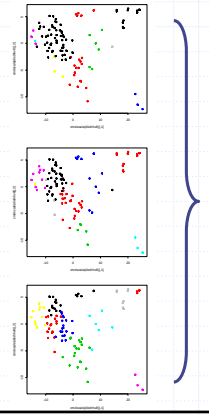
- Metrics Can Be Used To Compare Method Result Similarity

Consensus Clustering

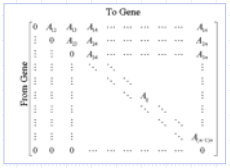
- Clustering Results Can Vary Depending on the Method Used
- Combine the Results of Multiple Methods into One Set of Consensus Results
- An Algorithm is Needed For Generating *Consensus Clusters* Given the Agreement Matrix
- We Use an Approximate Stochastic Algorithm Called Simulated Annealing

Consensus Clustering

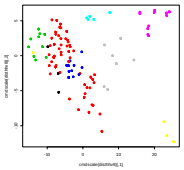
Input Cluster Results



Agreement Matrix



Consensus Clusters



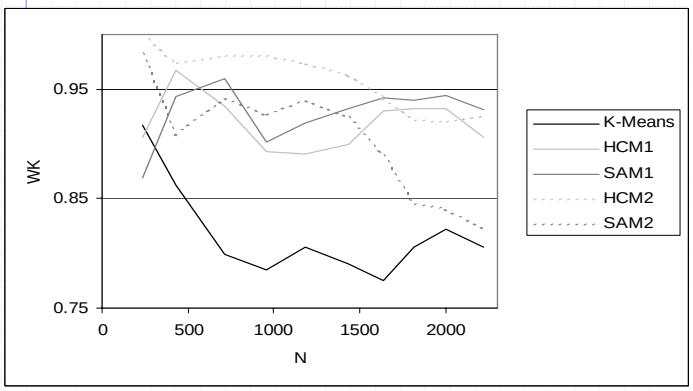
The Agreement Matrix

To Gene

From Gene

$$\begin{bmatrix} 0 & A_{12} & A_{13} & A_{14} & \cdots & \cdots & \cdots & \cdots & \cdots & A_{1n} \\ \vdots & 0 & A_{23} & A_{24} & \cdots & \cdots & \cdots & \cdots & \cdots & A_{2n} \\ \vdots & \vdots & 0 & A_{34} & \cdots & \cdots & \cdots & \cdots & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \ddots & & & & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \ddots & & & & \vdots \\ \vdots & \vdots & \vdots & & & \ddots & A_{ij} & & & \vdots \\ \vdots & \vdots & \vdots & & & & \ddots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & & & & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & & & & & & \ddots & A_{(n-1)n} \\ 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix}$$

Scalability Issues



Summary (1)

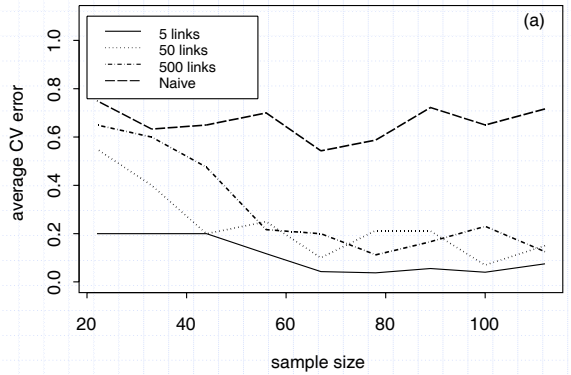
- Clustering and Grouping Problems are Hard!
  - Especially Microarray Data
- Difficult Choice of Metric, Cluster Worth and Method Against Problem
- There is No Free Lunch!

Allan Tucker

MicroArray Data

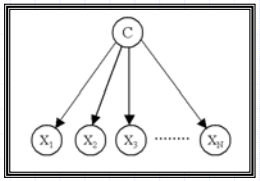
- High dimensional
- Small number of samples
- Model the data
  - Classification
  - Feature selection
  - Knowledge discovery
- Model complexity issues

Effect of Model Complexity



Identifying Predictive Genes

- Naïve Bayes Classifier
  - Well established
  - Minimises parameters
- Feature selection
  - Local stepwise methods
  - Global search (SA)
- Resampling methods
  - Cross validation

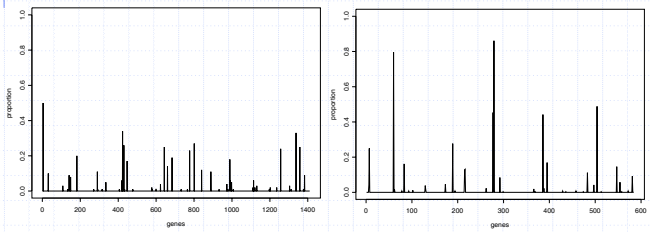


Identifying Predictive Genes

- Identify genes robustly
  - Data perturbed during CV
  - Repeats of stochastic SA search
- Assign confidence based upon the frequencies of genes being selected
- Limit maximum number of links - MDL

Confidence Scores

- Relatively small number of genes
- Identified with high confidence
- Consistency between runs



Identified Genes

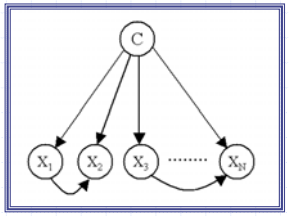
B-CELL		PROSTATE	
GeneBank	Proportion	GeneBank	Proportion
AK023995	0.862	AA055368*	0.5
U15173*	0.796	N64741	0.34
L21936	0.488	AA487560*	0.33
D83785	0.454	W47179	0.27
BC014433	0.442	AA486727	0.26
U59309	0.277	AA455925	0.25
-47202	0.25	H29252	0.25
Z14982*	0.169	AA010110	0.24
BC016182*	0.162	AA180237	0.23
U82130	0.146	AA443302	0.2
Z80783	0.131		
BC009914	0.127		
U77949	0.112		

Expert Knowledge

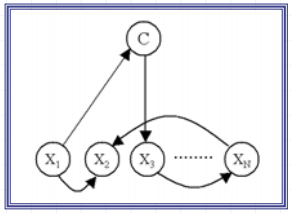
- Lots of other information available
  - Pathway Information
  - Gene Ontology
  - Sequence Information
  - Functional information
- Use this data as *prior knowledge*
- Update with data

Bayesian Classifiers

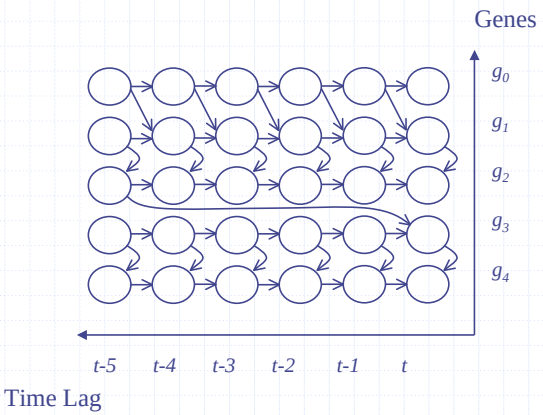
TAN - No longer assume independence between features



BNC – Include class node as a normal variable



Dynamic Bayesian Networks



Summary (2)

- When micro-array data only has small samples:
  - Simple models with small parameters best
  - Global search for parameters better
- Bayesian networks can incorporate different types of data
- Update expert knowledge with data

Conclusion

- Biological data are very noisy
- Modelling biological systems, at systems level?
- More integrated computational methods for organising and analysing data

Acknowledgements

- LIACS, LUMC, IBL
- MARIE, BIOMAP, BBSRC, EPSRC, Wellcome Trust
- Larry Hunter, Terry Speed
- Data kindly provided by
  - Paul Kellam from the Dept. of Immunology and Molecular Pathology, University College London.
  - Dr Li from the Dept. of Biological Sciences, Brunel University, Uxbridge.