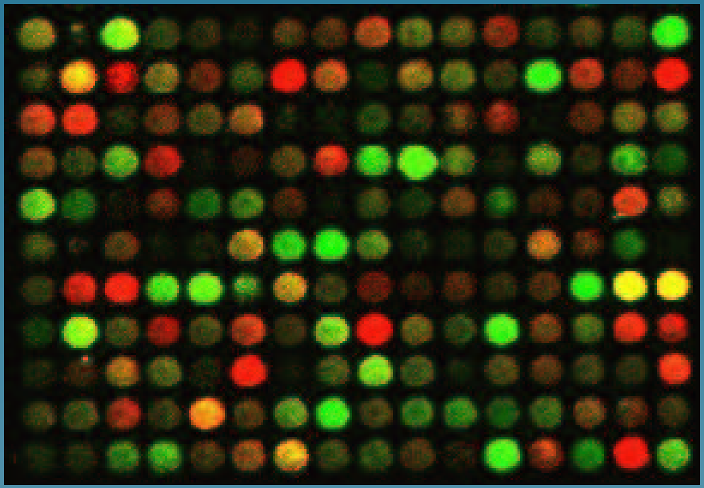
DNA microarrays



Jean-Philippe.Vert@mines.org

Ecole des Mines de Paris, France

'Mathematical aspects of molecular biology - Towards new mathematics', Nara, Japan, January 24-27, 2003.

Outline

1. The DNA microarray technology
2. Single gene analysis
3. Non-supervised clustering
4. Supervised classification
5. Systems biology

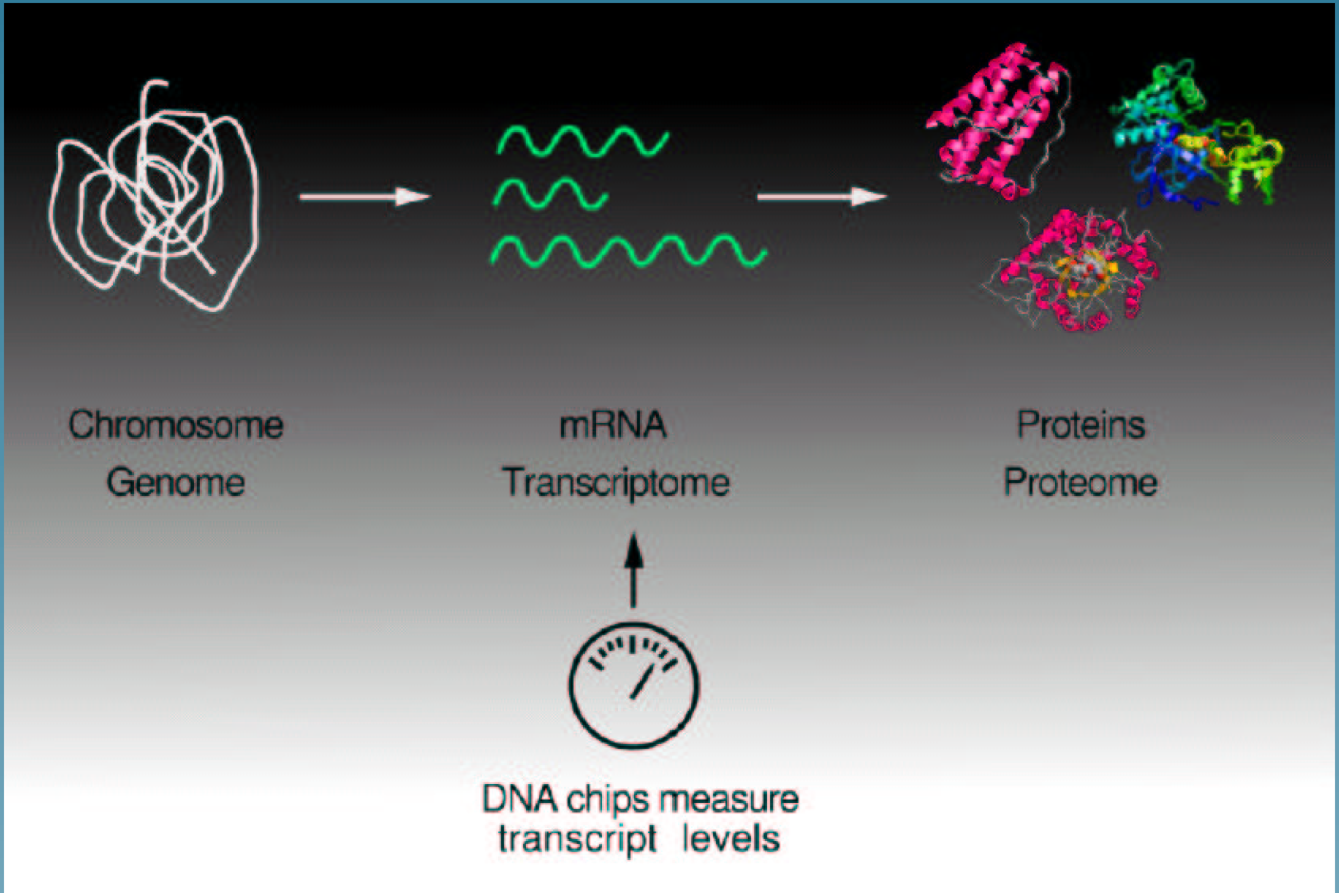
Part 1

The DNA microarray technology

Briefly...

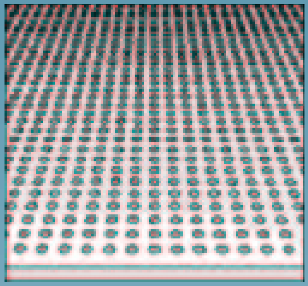
- Human DNA contains about 30,000 genes, encoding 100,000 proteins
- Understand life = understand how these proteins work together, are regulated ?
- DNA microarray is a tool to measure the quantity of mRNA (almost protein...) for all genes simultaneously, at a given instant.

DNA chips measure mRNA quantities



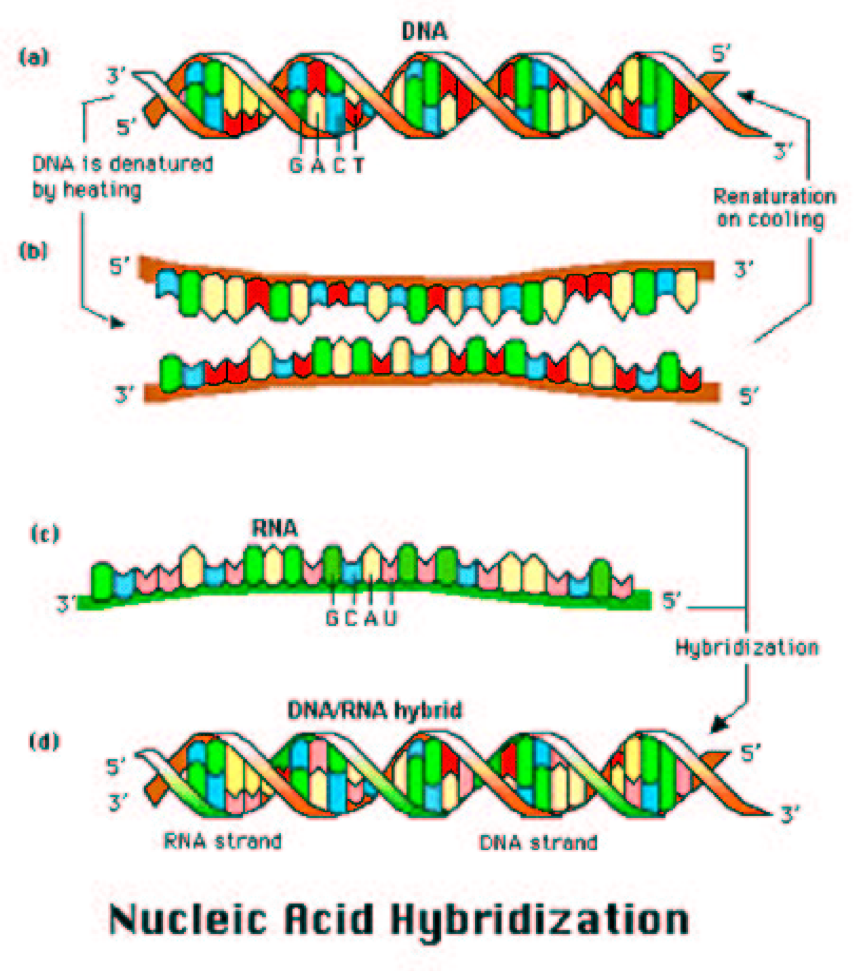
What are DNA arrays?

- A large number of DNA molecules spotted on a solid substrate (glass, nylon, or silicon)
- From 100 to 300,000 spots



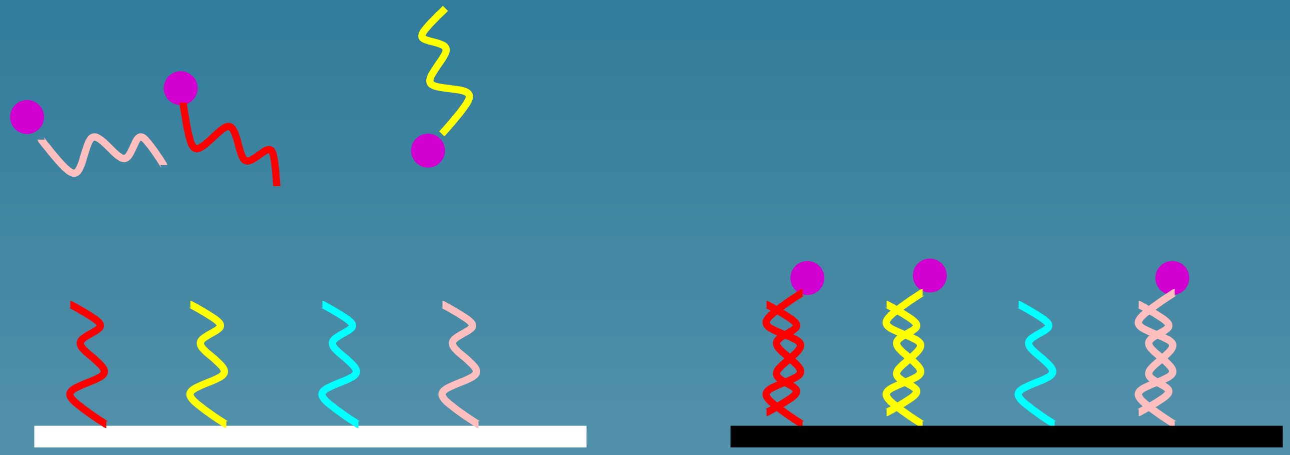
Affymetrix GeneChip® probe array. Image courtesy of Affymetrix.

How it works? Hybridization...

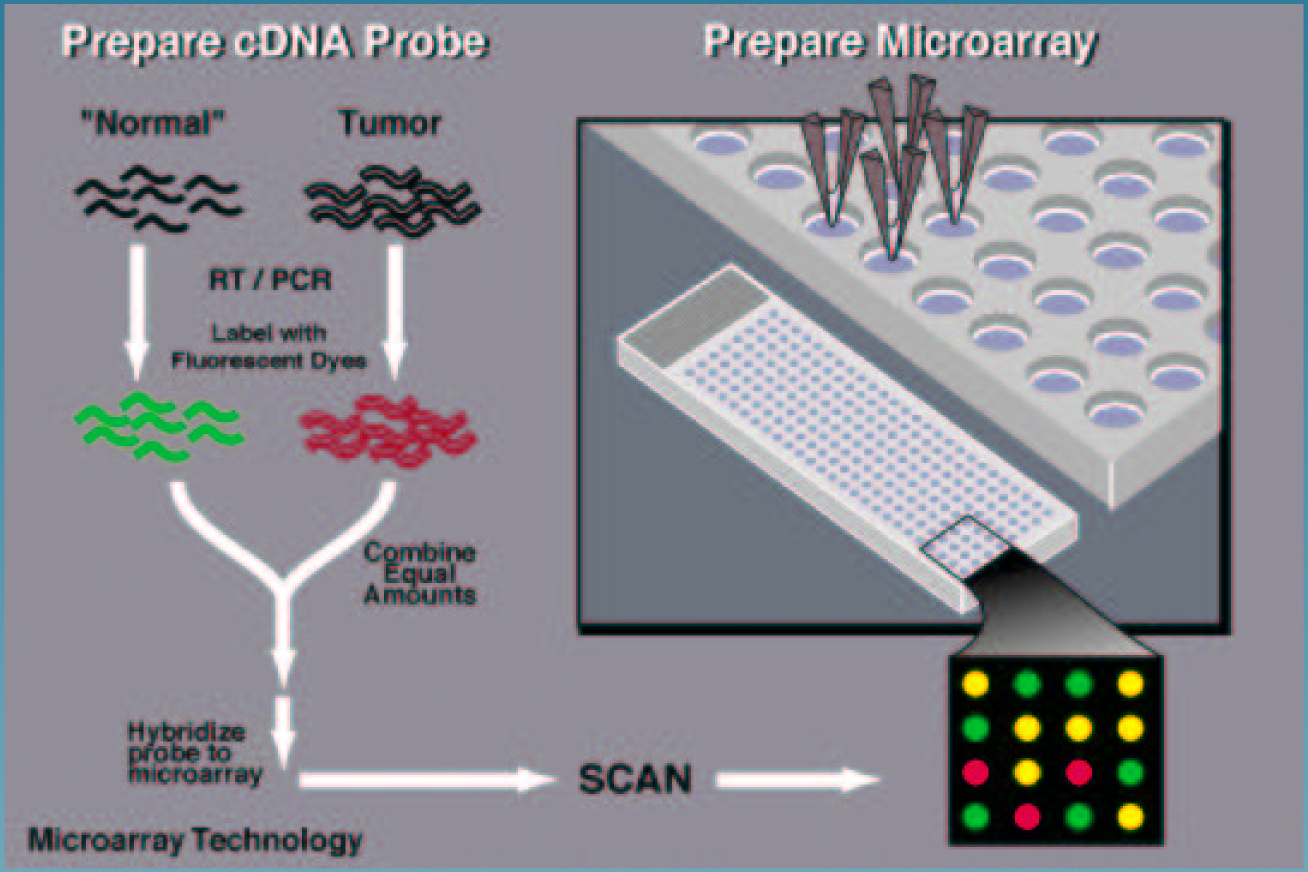


Nucleic Acid Hybridization

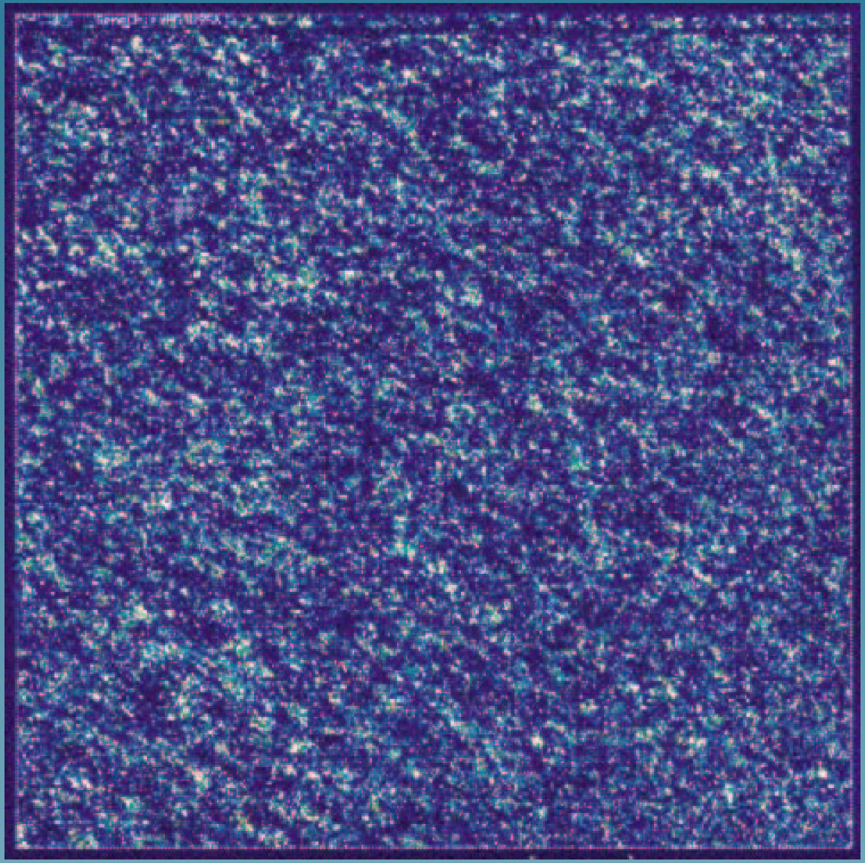
Hybridation on a chip

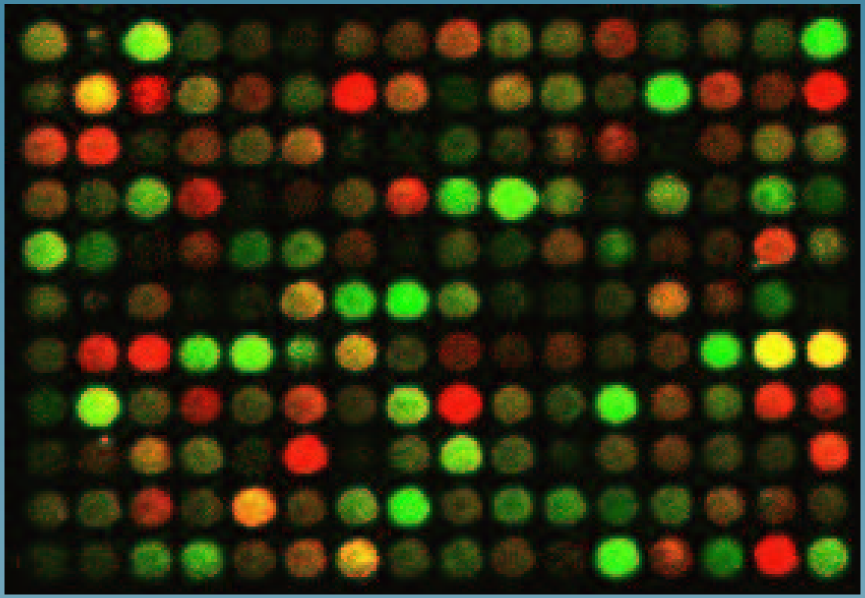


Classical experiment



What you get





The transcriptome

The transcriptome reflects

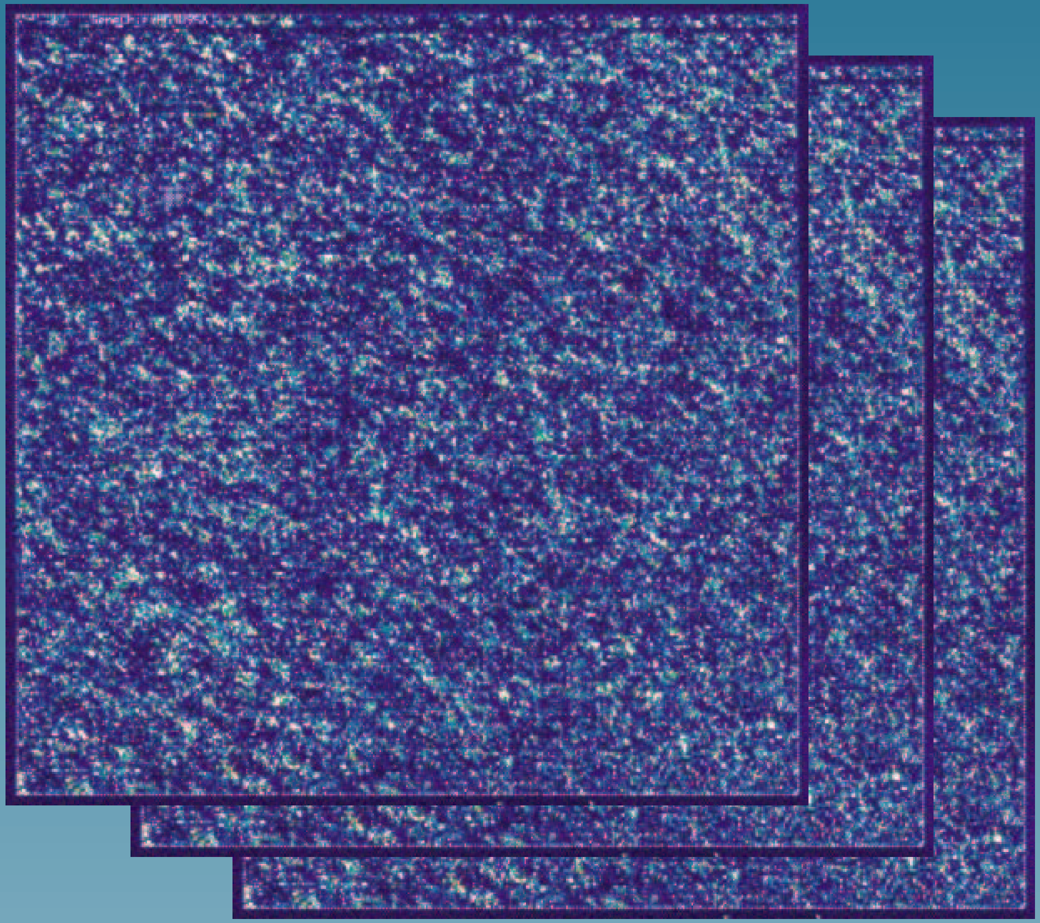
- tissue source, organe, cell type
- tissue activity and state
  - stage of development, grotwth, death
  - cell cycle
  - disease / healthy
  - response to therapy

Applications

- gene discovery for drug target
- disease diagnosis
- systems biology
- pharmacogenomics, genetic testing etc...

Single gene analysis

The problem



Experiments

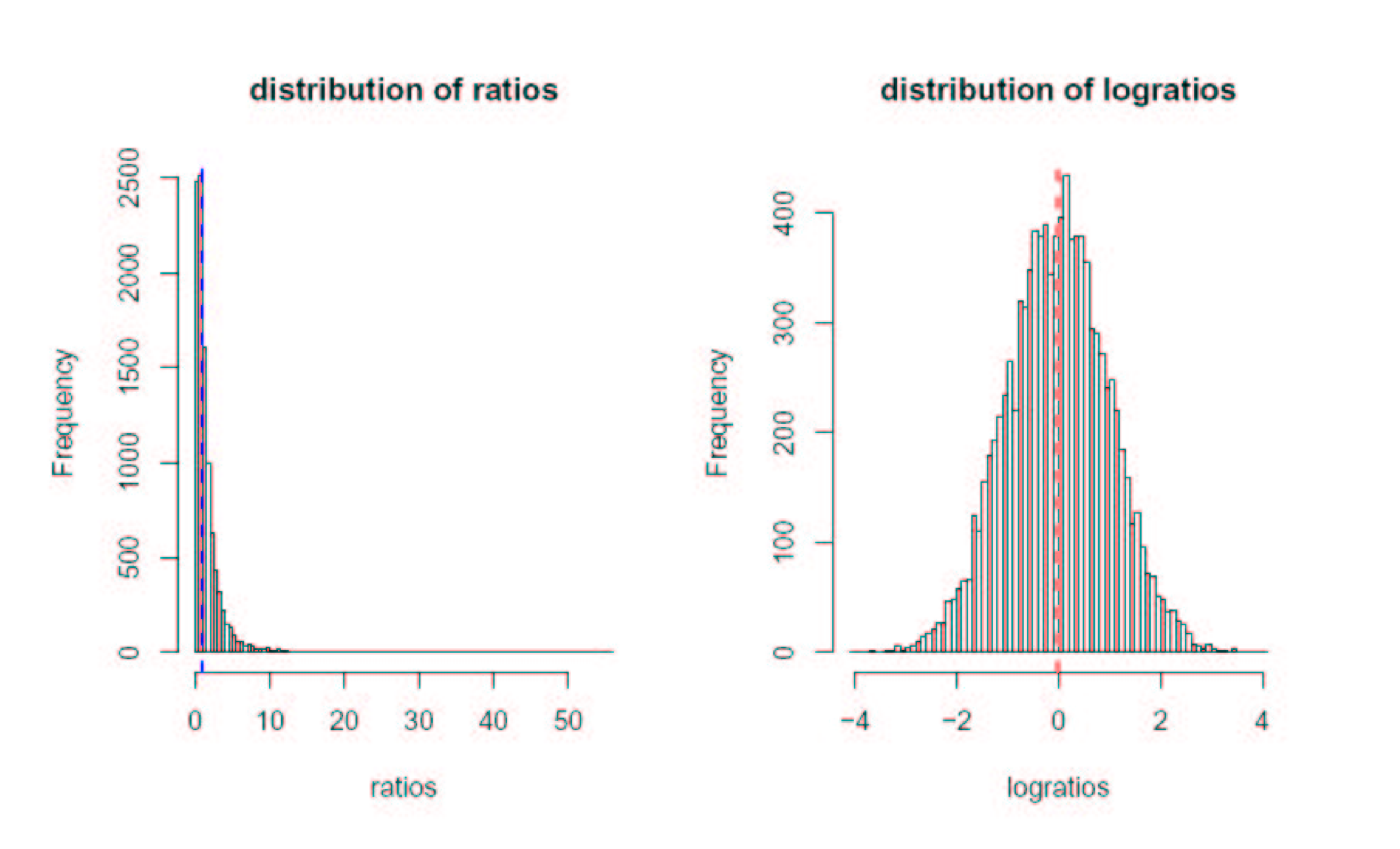
Genes

1.5	−2	0.2	3.4	−2.1	...
−4	2.1	0.5	1.1	0.9	...
...	...	...	...		

Spot intensity

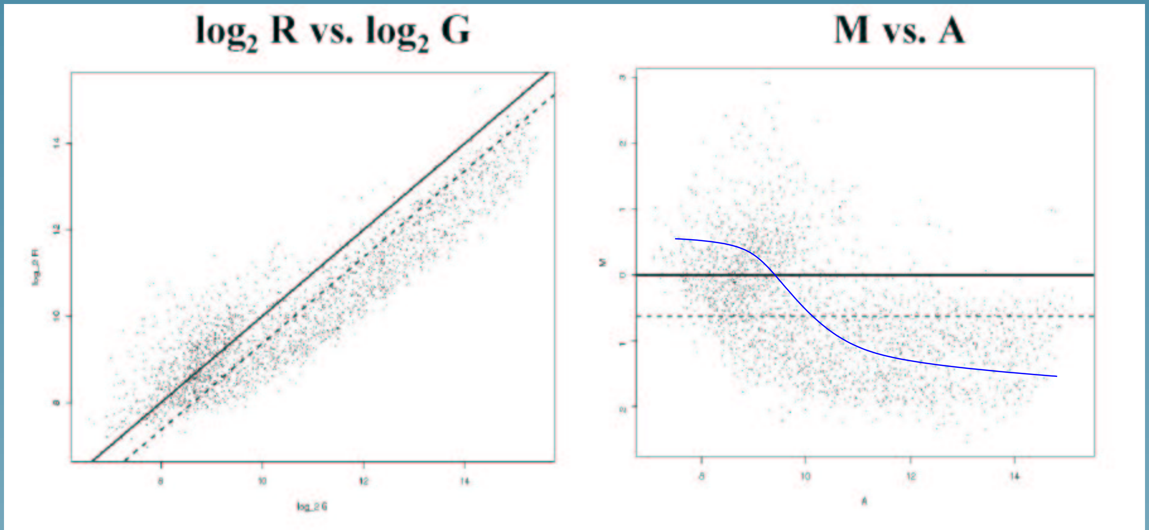
- Let R and G the intensity of the red and green spot, for a given gene
- The ratio R/G is indicative of the relative abundance of the mRNA quantity in the two samples
- R and G are estimated by image analysis algorithms

Ratio logarithm



Self-self hybridation

$$\begin{cases} M & = \log R - \log G \\ A & = \log R + \log G \end{cases}$$



Normalization

- Normalization is required to ensure that differences in intensities are due to differential expression, and not printing, hybridation or scaning effects

- Several statistical techniques to remove the 'noise'.

- Result: for each gene, a number to indicate over/under-expression.

Application

- input: microarrays for two different conditions
- Output: a list of differentially expressed genes
- Suggests more investigations on this genes, but limited.

Non-supervised clustering

Motivations

- Find some hidden structure in the data
- In cluster analysis, the goal is to find groups, or clusters, of similar objects
- Object = genes and/or experiments

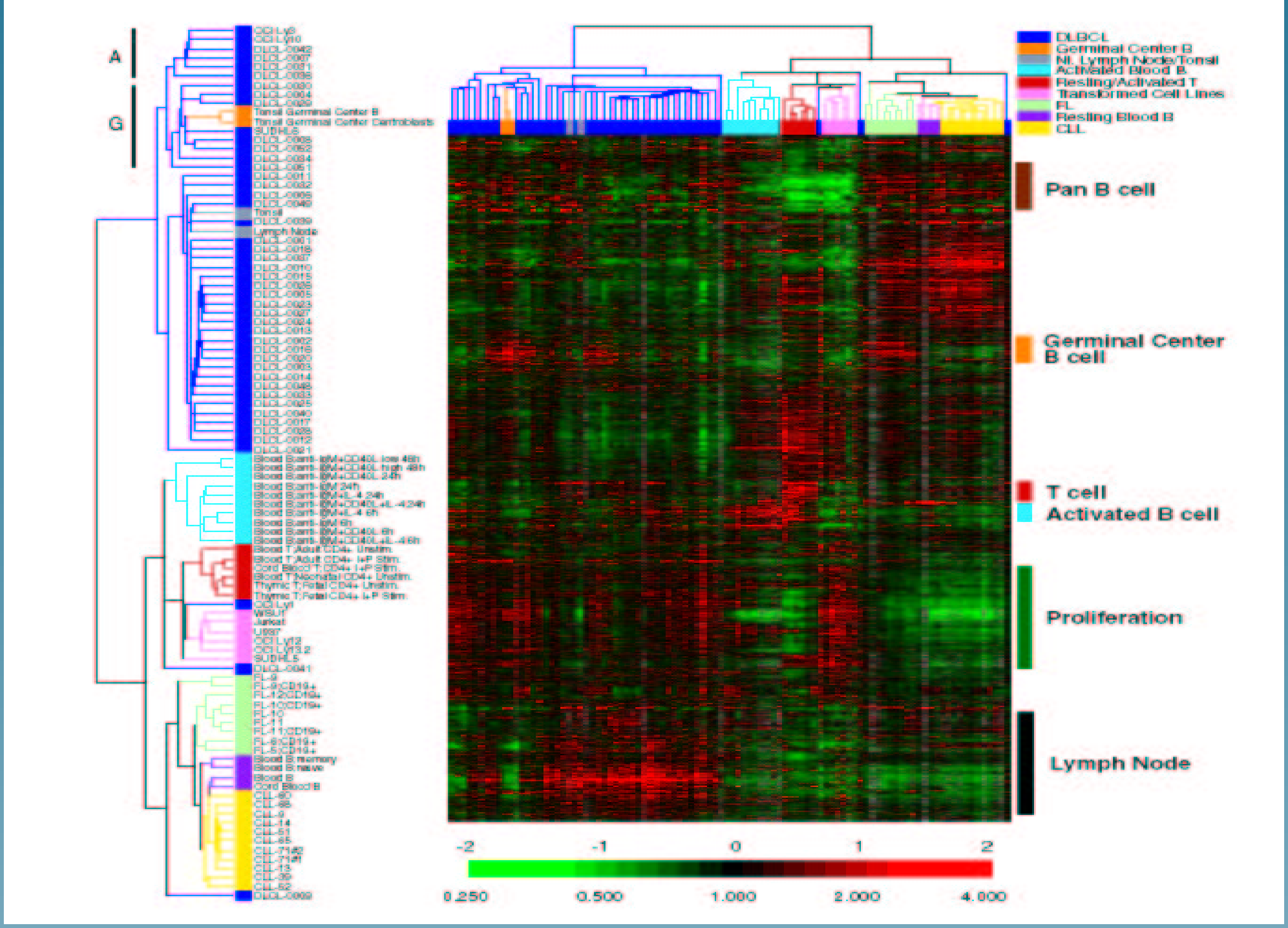
Gene clustering

- For vizualization
- To detect biologically related genes (interact, participate in a common biological process...)
- To detect spatial or temporal patterns (depends on the experiments)

Experiment clustering

- To detect clusters of experiments such as tumor classes, cell types, and the relations among them.
- To detect experimental artifact
- For vizualization

Example (Alizadeth et al., 2000)



Clustering overview

- Define a distance for objects to be clustered
- Choose a clustering algorithm:
  - ⋆ hierarchical methods (either divisive or agglomerative) provide a hierachy of clusters, from the smallest (singletons) to the largest (whole set)
  - ⋆ partitioning methods output K clusters, where K must be specified

Define a distance

- Each object (gene or experiment) is represented as a vector $x = (x_1, \ldots, x_n)$.
- Euclidian distance is natural
- Centering ($\sum x_i = 0$) and scaling to unit norm ($\sum x_i^2 = 1$) can be useful

Distance between clusters

Let A and B two clusters, and d a distance between objects. Then d can be extended by:

$$d(A,B) = \min_{(x,y)\in A\times B} d(x,y) \qquad \text{single linkage}$$

$$d(A,B) = \frac{1}{|A||B|} \sum_{(x,y)\in A\times B} d(x,y) \qquad \text{average linkage}$$

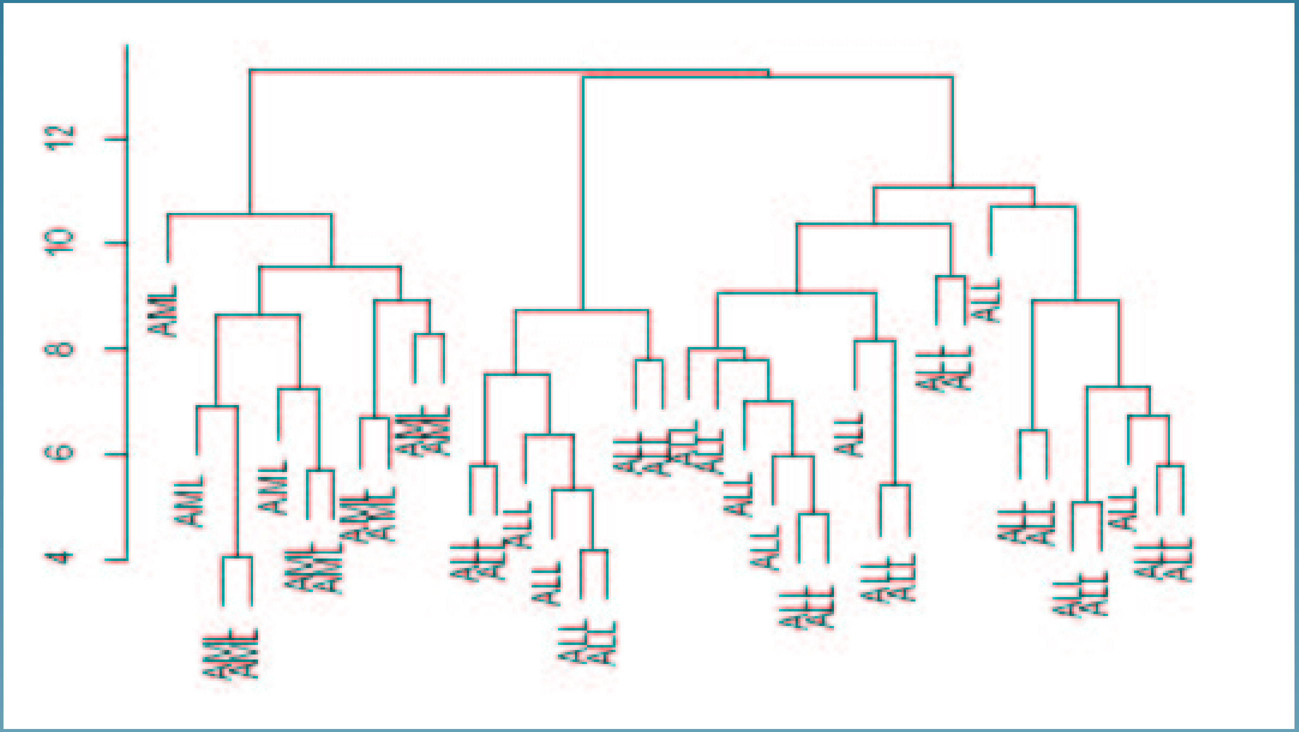
$$d(A,B) = \max_{(x,y)\in A\times B} d(x,y) \qquad \text{complete linkage}$$

Hierarchical clustering

It is a widely-used agglomerative hierarchical method:

- Start with all singletons as clusters
- At each step, merge the two clusters with the minimum distance between them, until only one cluster remains.
- Output the hierarchy as a tree/dendogram.

Example



From Golub et al., clustering of two cancer types.

k-means clustering

This is a simple and widely used partitioning method. The number of clusters k is fixed.

- Chose k points as initial centroids.
- At each step: assign each object to the cluster with the closest centroid, and adjust centroids (e.g., average the members of a cluster)
- Iterate until convergence

k-means clustering properties

- Many variants (choice of distance, averaging, etc...)
- When the cost function corresponds to an underlying probabilistic mixture model (e.g., Euclidean distance and mixture of Gaussians) then k-means is an online approximation to the EM algorithm, and converges toward a local maximum likelihood.
- Chosing k is problematic

Advantages of clustering

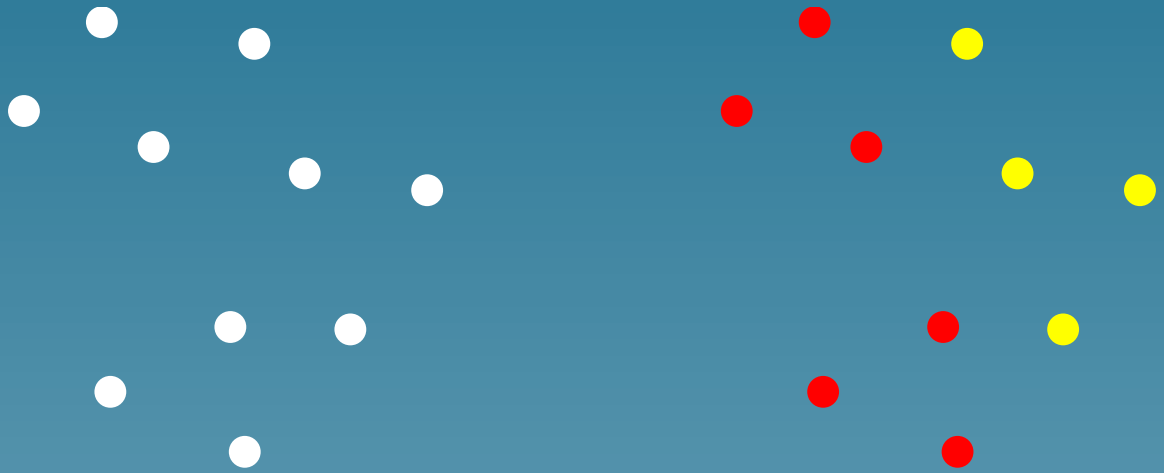
- Intuitive and quick algorithms
- Vizualization of the results
- Has proved to be useful as a first data mining tool for microarray data

Pitfalls of clustering

- the clustering problem is usually ill-posed.
- We will always find clusters, even if there is no structure in the data.
- If several cluster structures are super-imposed, what will we get?
- So easy to use that few precautions are taken.

Supervised classification

From clustering to classification



From clustering to classification

- Clustering: a set of points $(X_1, \ldots, X_N)$ is given, we are looking for intrinsic structures.

- Classification: in addition, a set of values $(Y_1, \ldots, Y_N)$ is given, we want to find the link between X and Y.

- Classification is easier than clustering. If the problem is simple, both can be the same, but not in general.

Application of classification

- predict cell type, cancer type, response to a treatment, type of bacterial pathogen from microarray data

- predict gene class (function, localization...) from its expression profile

- but impossible to discover new class.

Mathematical formulation

- Let (X, Y) be a $\mathbb{R}^d \times \{0, 1\}$ valued random pair, with joint probability P.

- A classifier is a function $g : \mathbb{R}^d \longrightarrow \{0, 1\}$

- We observe an i.i.d. sample $(X_i, Y_i)_{i=1,\dots,N}$

- We must infer a classifier $\hat{g}$ such that $P(\hat{g}(X) \neq Y)$ be as small as possible.

Remarks

- There exists an (unknown) best classifier, given by:

$$g(x) = \begin{cases} 1 & \text{if } P(Y = 1|X = x) > \frac{1}{2} \\ 0 & \text{if } P(Y = 1|X = x) \leq \frac{1}{2} \end{cases}$$

- P is unknown, only the i.i.d. sample is observed

Learning algorithm

- a set $\mathcal{G}$ of candidate classifier
- a mapping $\left(\mathbb{R}^d \times \{0,1\}\right)^N \longrightarrow \mathcal{G}$ which choses a classifier $\hat{g}$ from the observed data

Empirical risk

The risk of a classifier g is:

$$R(g) = P(g(X) \neq Y).$$

The empirical risk is

$$R_{emp}(g) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(g(X_i) \neq Y_i).$$

When N is large, $R_{emp}(g) \to R(g)$ but...

Overfitting

- Overfitting occurs when:

$$R_{emp}(\hat{g}) << R(\hat{g}).$$

  - ⋆ N is too small,
  - ⋆ $\mathcal{G}$ is too large.

- This is typically the case in most microarray-related problems!

Statistical learning theory

- Studies under which conditions $R(\hat{g})$ is small
- Main results: an algorithm which minimizes $R_{emp}(g)$ is good, if the capacity of $\mathcal{G}$ is small
- A trade-off must usually be found between
  - ⋆ $\mathcal{G}$ too small (too poor to mimic $P(Y|X)$)
  - ⋆ $\mathcal{G}$ too large (overfitting)

Classification algorithms

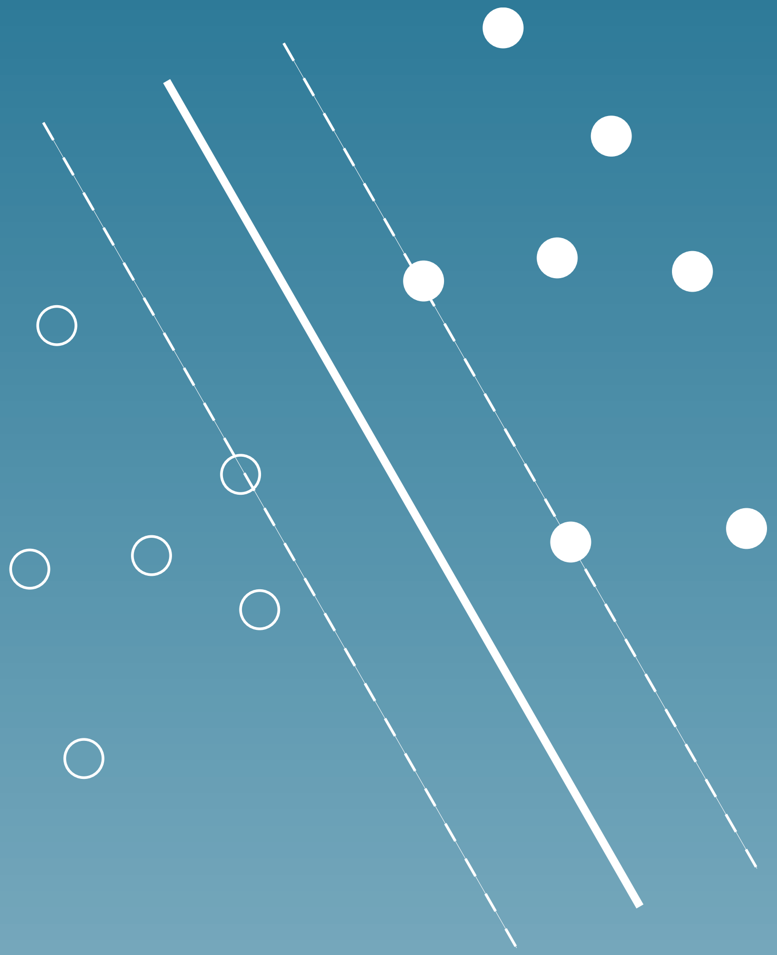
- A long list: Fisher linear discrimination, discriminant analysis, naive Bayes, Bayesian belief networks, logistic regression, neural networks, classification trees, nearest neighbour classifiers, support vector machines, bagging, boosting...
- Performance depends on the problem.

Example: FLDA

Introduced in 1936:

- Find directions to project the points, with large ratios of between-groups to within-groups sums of square
- predict the class of a new observation by the class whose mean vector is closest in terms of projection.

Example: linear SVM



Remarks about classification

- Theory progressed a lot recently
- Still unadapted to microarrays: how to learn from 100 points in a 100,000 dimensional space?
- This is going to be a major topic of research in the coming years

Systems biology

Motivation

- Individual genes interact, are regulated, and are part of a complex system (life)
- For the first time, with microarrays, we can have a global view of the system (at the transcriptome level)
- Can we then reconstruct / understand the system?

Possible Applications

- Basic biology: understand biological process
- Medicine: any action on an individual gene or protein can have consequences on the system

A general approach

- Make a formal model for biological system
- Design experiments with microarray and fit the model to observations.

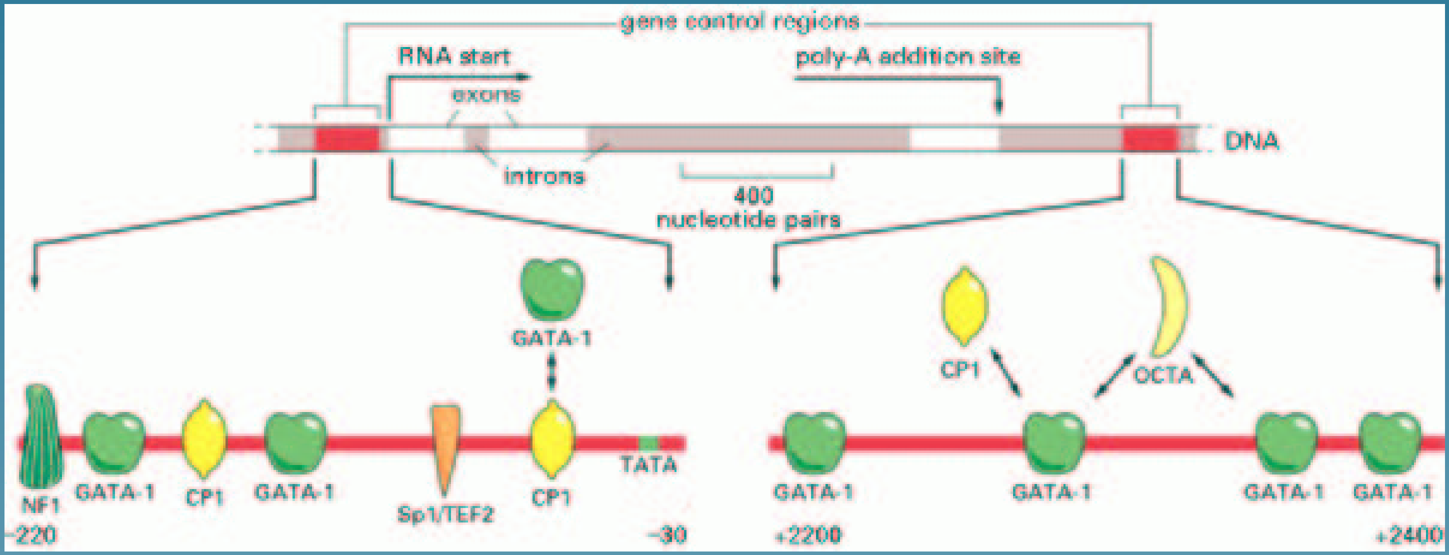
This is a topic where mathematics and biology must meet!

Biological considerations

A formal model for biological system could include biological evidences:

- gene regulation (observed through the transcriptome)
- chemical interaction (interactome)
- information transmission (signalling pathways)
- chemical process (metabolic pathways)
- evolutionary relationships

Example: B-Globin expression regulation



Computational models of regulatory networks

- boolean networks
- differential equations
- stochastic networks
- Bayesian networks

Boolean networks

- The expression of gene i at time t is represented by a $\{0,1\}$-valued variabel $X_i(t)$

- Evolution equation:

$$X_i(t+1) = F_i(X_1(t), \ldots, X_N(t)).$$

- F can be inferred, to some extent, by expression profiles

Remarks

- For a given model, one can study attractors / cycles / bifurcation, topological properties of the graph (connectivity...), global properties of large random networks etc...
- However: binary deterministic model not very realistic
- This can be generalized to a variety of continuous-time and continuous-value models (S-systems...)

Continuous models

- Generalize boolean networks: continuous-time and real-valued systems

- Example: S-sytems

$$\frac{dX_i}{dt} = \sum_k T_{ik} \prod_j X_j^{g_{ijk}} - \sum_k U_{ik} \prod_j X_j^{h_{ijk}} + I_i(t).$$

- Universal approximation properties (idem neural networks)

Model fitting

- For a fixed model structure, parameters learned by minimization

- Big problems: how to infer the model? Curse of dimensionality for the parameters?

- Currently, some small models for the best-studied regulatory switches in bacteria...

Probabilistic modelling

- A microarray experiment seen as a random vector
- Goal = estimate a probability distribution for the expression vector, based on a series of experiments
- Big problem: how to infer the law of a 100,000-dimensional vector from 100 observations?

Example: Bayesian models

- A convenient way to represent a probability distribution for N variables
- It is based on a graph whose vertices are the variable indexes
- Conditionnaly to its neighbours, the law of a variable X_i is independant of the other variables.
- Methods exist to estimate the graph and the parameters

Summary: challenges in systems biology

- Formal models for biological systems
- Learning from few points in high dimension

Conclusion

Conclusion

- Microarray technology is a new and revolutionnary technology
- Can be used to answer practical questions (e.g., diagnosis)
- Gives a snapshot of the whole transcriptome at a given instant: can be used to better understand biological systems
- Can be combined with several other new high-throughput technologies
- Does not fit current mathematics