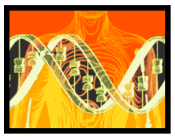
MASTER IBM 2010

INTRODUCTION À LA FOUILLE DE DONNÉES EN BIOINFORMATIQUE

COURS MASTER IBM 2010



JEAN-DANIEL ZUCKER

DR À L'IRD UR UMMISCO
(MODÈLISATION MATHÉMATIQUES ET INFORMATIQUES DES SYSTÈMES COMPLEXES)
UMR U755 INSERM/PARIS 6 ET LIM&BIO/PARIS 13

UPMC PARIS UNIVERSITAS — IRD Institut de recherche pour le développement — Inserm Institut national de la santé et de la recherche médicale — Lim&Bio

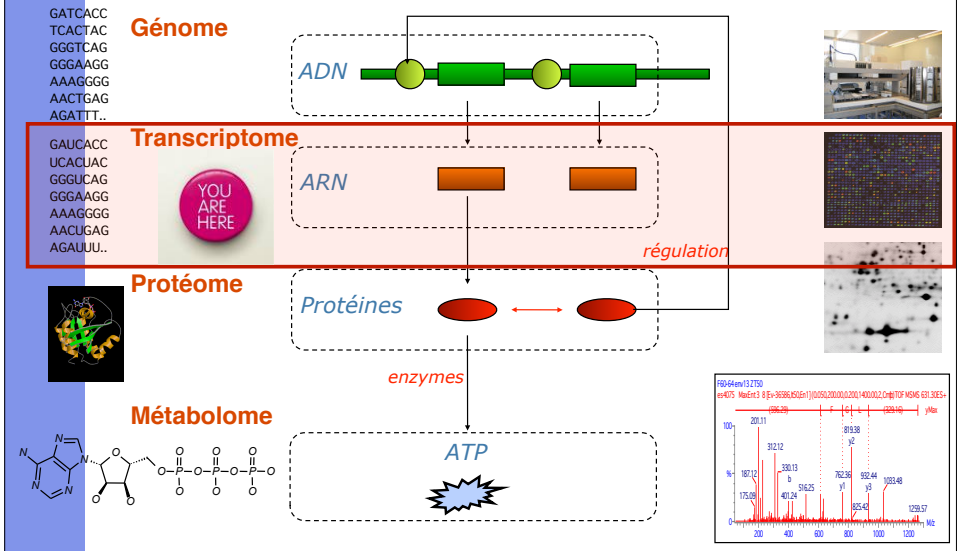
MODULE BIOINFO M2 IBM : DIDACTIQUE

- **Objectifs:**
  - COMPRENDRE LE DOMAINE DE LA BIO-INFORMATIQUE
  - LE RÔLE DE L'INFORMATIQUE ET L'IA DANS CES PROBLÉMATIQUES
  - COMPRENDRE LES ALGORITHMES DE BASE DE LA FOUILLE DE DONNÉES POUR LA BIOINFORMATIQUE
  - UTILISER R ET DES APPLETS
- **Examen: Controle.**

ADMINISTRATIF: MODULE NT-BIO MASTER IBM

- **Mardi 17 Nov. 2010 –** INTRO GÉNÉRALE
  - **La bioinformatique : les 'omics'**
  - **La génomique : BD et algorithmes**
- **Mercredi 18 Nov 2010 –** ALGORITHME POUR LA BIOINFO ET OMIQUE
  - **Les réseaux, les BD, les ontologies.**

QUELLES DONNÉES DE l'analyse des «omes» ?



Plan

A. Les données BioPuces: du signal aux données brutes

B. Stockage et standardisation des données

C. Le traitement statistique des données

D. Analyse et Fouille de données: Clustering

E. Analyse et Fouille de données: Annotations

F. Analyse et Fouille de données: Réseaux

Transcriptome: perspective historique

Dans l'ère de la "post-génomique": identifier la fonction des gènes

- **Mécanisme d'hybridation de l'ADN (1960s)**
- **« un fragment d'ADN simple brin ou d'ARN messager est capable de reconnaître son brin complémentaire parmi des milliers d'autres: c'est le phénomène d'hybridation »**
  - **Détection des hybrides**
  - **Fixation sur les supports:**
    - »**Southern blots (1970s),** **Northern blots, Dot blots**

Une (r)évolution ?

- **La nouveauté (1990): passage à l'échelle**

« Il est devenu courant de déposer 20 000 préparations différentes sur des membranes de Nylon de 20 centimètres de côté. La puce à ADN (DNA chip ou biochip en anglais) résulte de l'évolution de ce format vers une miniaturisation plus poussée, qui atteint une densité de 250 000 unités réactionnelles par centimètre carré. »

–**centaines, milliers de sondes au lieu de 10**

- Sondes **sont attachées à des supports physiques**
- La robotique est **largement utilisée**
- L'informatique joue un rôle clef dans toutes les étapes

Principales technologies

- **Sondes cDNA (> 200 nt) sur nylon ou verre**
- **Oligonucleotides (25-50 nt) sur du verre**
- **Oligonucleotides (25-60 nt) synthétisées in situ sur du silicon**
- **(et d'autres puces chromosomiques, …)**

Des puces possédant une unique opération: « quantifier les transcrits »

Biopuces Commercialisées (liste non exhaustives)

- **Clontech, Incyte, Research Genetics : jusqu'à 8000 clones**
- **Incyte / Synteni - Biopuces à 10000 sondes, non distribuées (il faut envoyer l'ARN)**
- **Affymetrix - Biopuce basée sur des oligon.** Ex: HG-U133A Affymetrix contient des sondes pour 22.000 gènes humain.
- …

Un facteur limitant: le prix d'achat

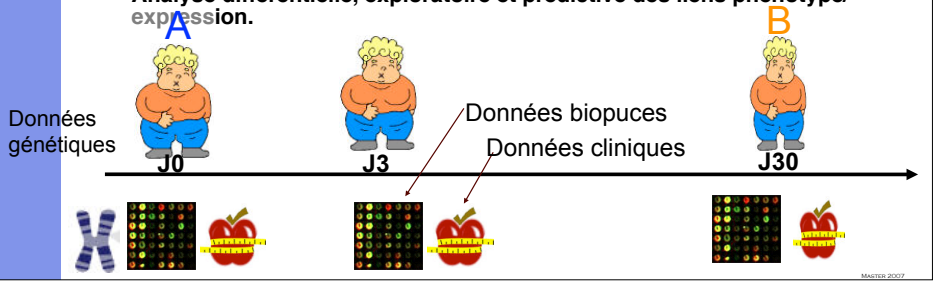
- **Prix des puces Affymetrix GeneChip (Prix 3 Mars 2008)**
- Chaque puce « GeneChip » représente entre 10,000 et 40,000 gènes différents ou EST.
- Les prix ci-dessous tiennent compte d'un rabais …

* Rhesus Macaque Genome Array -	$460
* Human Genome HG-U133 Plus 2.0 -	$460
* Human Genome HG-U133A 2.0 -	$310
* Murine Genome MOE 430 2.0 -	$460
* Murine Genome MOE 430A 2.0 -	$310
* Murine Genome MOE 430A -	$410
* Murine Genome MOE 430B -	$410
* Rat Genome RAE 230 2.0 -	$435
* Rat Genome RAE 230A -	$385
* Rat Genome RAE 230B -	$385
* C. elegans Genome Array	$310
* Canine Genome Array -	$310
* Chicken Genome Array -	$ 435
* Drosophila Genome 2.0 -	$310

http://www.ohsu.edu/gmsr/amc/amc_price.html

Exemple: Analyse de la réponse à un VLCD

- **Hypothèse:**
  - **L'expression de certains gènes clés varie en réponse au changement de l'environnement nutritionnel ou hormonal.**
- **Matériels:**
  - **Données phénotypiques (investig. cliniques) et d'expression (Puces pangénomiques (44000 ADNc)) au cours d'un régime hypocalorique.**
- **Méthodes:**
  - **Analyse différentielle, exploratoire et prédictive des liens phénotype/ expression.**

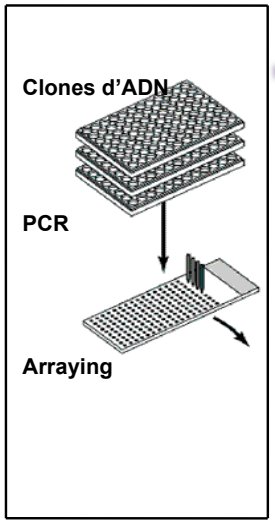


… autres types d'analyse

- **Analyse de l'expression du génome pour caractériser**
  - **Effets de certains médicaments**
  - **Mécanismes de développement de maladie**
  - **Réponses à des facteurs environnementaux**
  - **Diagnostic moléculaire**
  - **Réseaux de régulations de gènes**
- **Détection de variation de séquences**
  - **Typage génétique**
  - **Détection de mutations somatique**
  - **Séquençage direct**

Seule l'imagination des chercheurs et leur budget peut limiter le nombre d'analyses…

Etape 1: De l'artisanat à la fabrication en série de puces



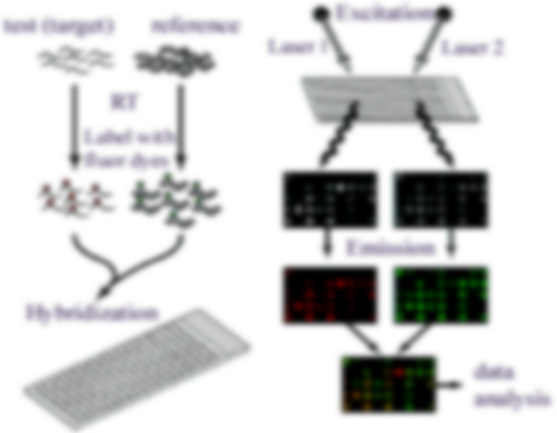
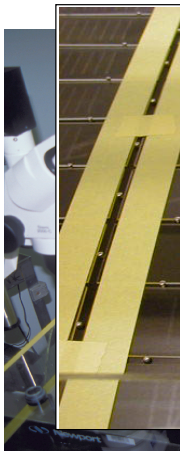
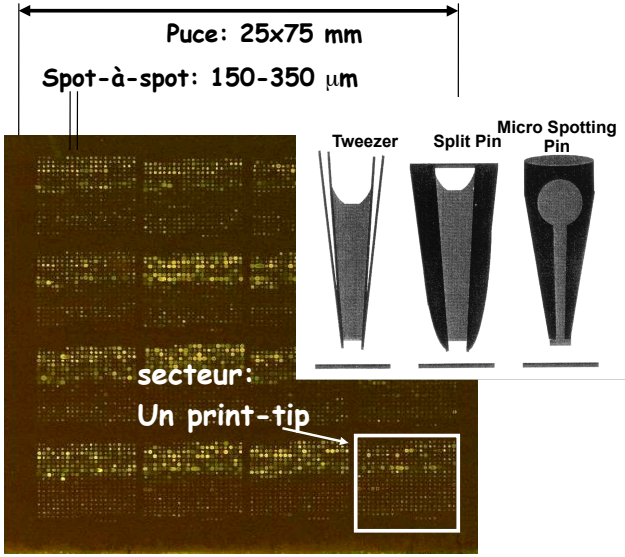


Figure de: David J. Duggan *et al.* (1999) Expression Profiling using cDNA microarrays. *Nature Genetics* **21**: 10-14

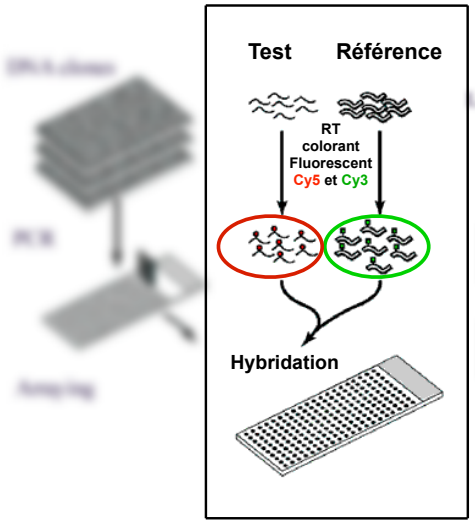
Le calibrage du robot imprimeur



URL: http://cmgm.st

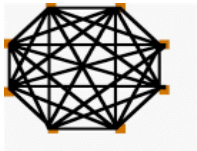


Etape 2: Hybridation



Le choix du plan d'expérience

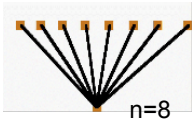
Type I: Test/Référence Condi vs. condj ou avant/après



n=8
nbpuces=28

Type II: Pool de référence d'ARN (tissus-teque)

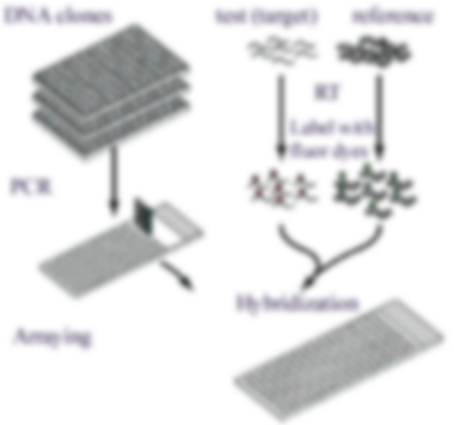
- Tous les ARN de test sont comparés au même "pool" de référence.
- Permet de réduire le nombre d'expérimentations comparatives pour n de 0.5 x (n^2-n) à n
- Le fait d'utiliser deux populations d'ARN en compétition et de mesurer le rapport d'hybridation permet d'éviter les complications liés aux problèmes de cinétique de l'hybridation.
- Permet de comparer des protocoles d'expérimentations



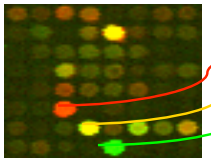
n=8
nbpuces=8

Série temporelles/Données appariées ou non

Etape 3: Le signal des puces

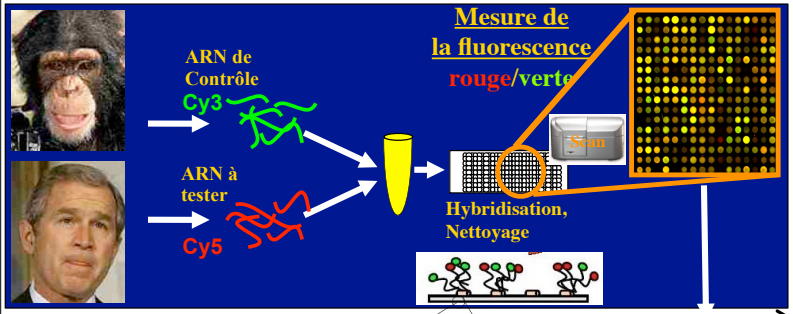


Mesure du ratio



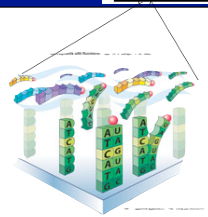
Cy3	Cy5	Cy5/Cy3	$\log_2$(Cy5/Cy3)
200	10000	50.00	5.64
4800	4800	1.00	0.00
9000	300	0.03	-4.91

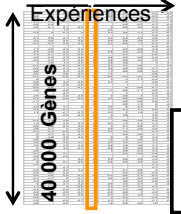
Le principe des puces : photographie de l'expression génique



Key factor of variability/precision:

- **Protocols** (**Type I ou Type II**, "dye-swap", replication)
- **Quality** of RNA used
- **Technology** of the chips (cDNA, oligo, …)
- **Chip batches**
- **Hybrudizatiion** (conditions and technician)
- **Scan** (software and technician)
- **Number of chips**

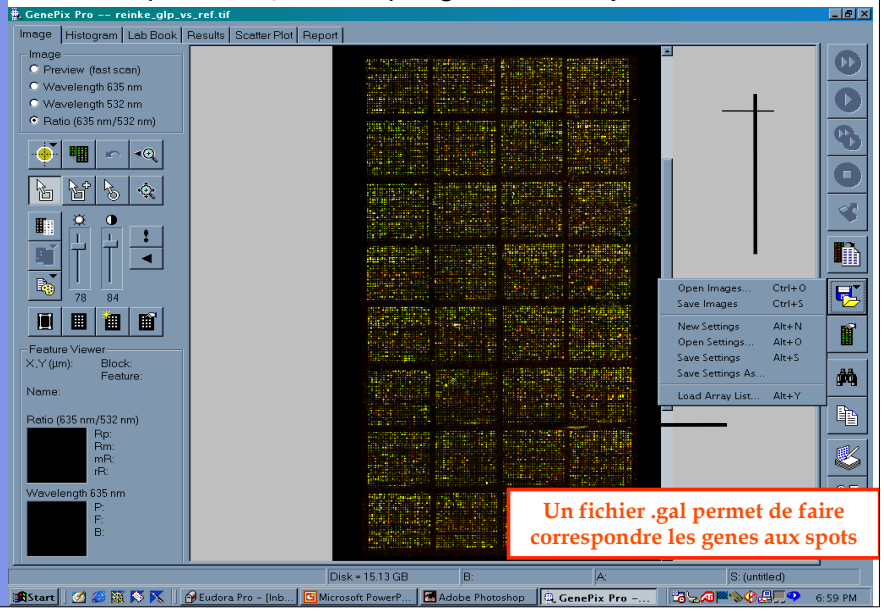




Analyse des profils D'expression

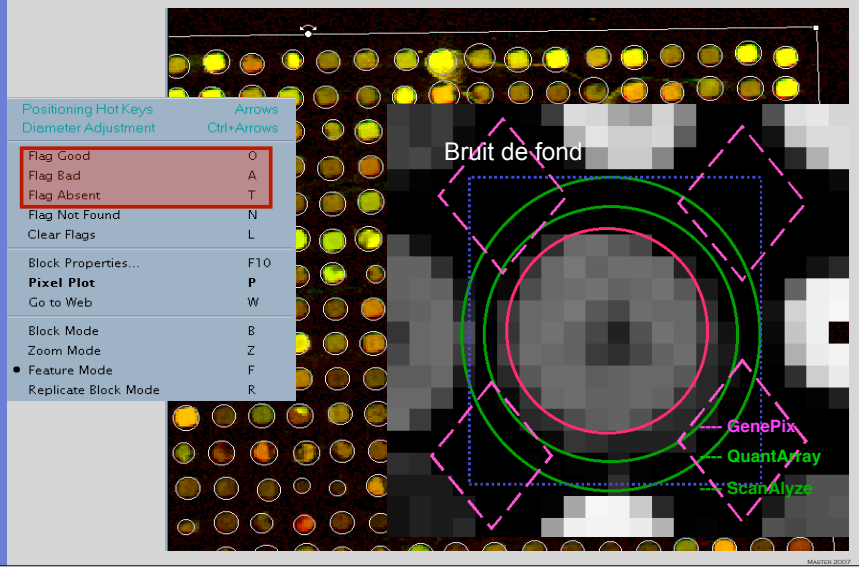
21

Les outils (comment ça marche): alignement des spots



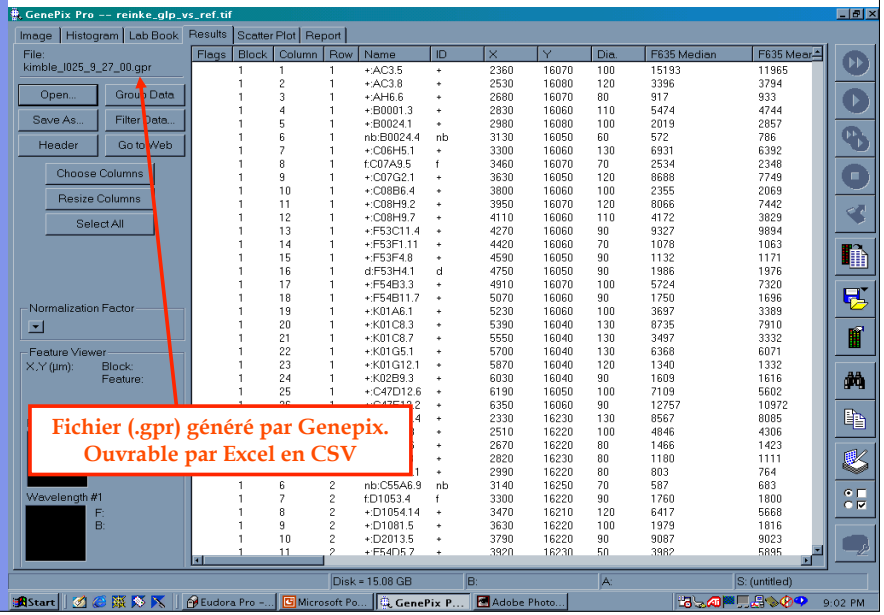
22

Les outils: Etiquettage des spots/ Calcul du bruit de fond



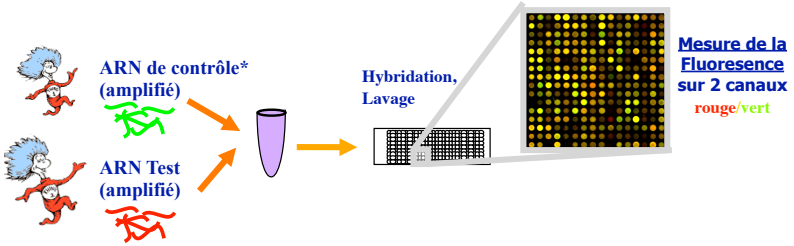
23

Les outils: Quantification des ratios d'expression



24

LES DONNÉES BRUTES: PROCESSUS QUALITÉ NÉCESSAIRE



Facteurs clefs affectant (variabilité/précision) les résultats bruts:

- **Protocole** (Type I ou Type II, "dye-swap", replication)
- **Qualité** des ARNm utilisés
- **Technologie** des puces utilisées (cDNA, oligo (choix des sondes), …)
- Fabrication des puces (même **lot**)
- **Hybridation** (mêmes conditions et même technicien)
- **Scan** des puces (logiciels (version!) et technicien)
- **Nombre** de puces à traiter (~ 50)

Besoins: Stocker les données brutes pour les traiter et les analyser

40 000 genes, 100 patients, 3 temps, 2 replicats, ~50Mo par image~400Ko ratios

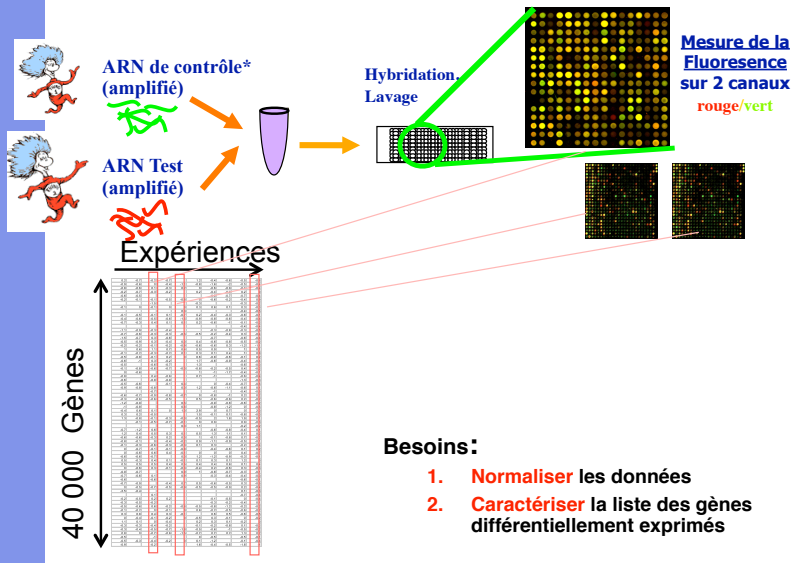
-> 300 Go d'Images 24 Millions de ratios

Taille des Bases de Données

W.P. Kuo et al. / Journal of Biomedical Informatics 37 (2004) 293–303

Specimen	No. of samples	No. of genes	Platform	Author
Adenocarcinomas	279	9376 common	Affymetrix, cDNA	Ramaswamy et al. [63]
Breast cancer	117	~25,000	cDNA	van't Veer et al. [64]
Drosophila melanogaster	66	4028	cDNA	Arbeitman et al. [65]
Prostate cancer	52	~12,600	Affymetrix	Singh et al. [66]
CNS embryonal tumors	99	6817	Affymetrix	Pomeroy et al. [67]
Primary tumors	144	16,063	Affymetrix	Ramaswamy et al. [35]
Small round blue cell tumors	63	6567	cDNA	Khan et al. [3]
Lung carcinomas	186	12,600	Affymetrix	Bhattacharjee et al. [21]
Ex: VLCD	49	40000	cDNA	Clement, et al.

Données standardisées: une colonne = une biopuce

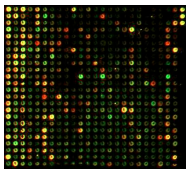


Besoins:

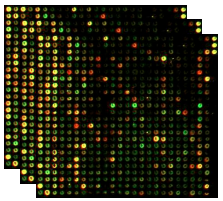
1. **Normaliser** les données
2. **Caractériser** la liste des gènes différentiellement exprimés

Normalisation des données:

utiliser hypothèses pour retirer le signal « non biologique »

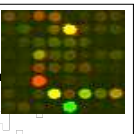


- **Sur une même biopuce**
  - Pour normaliser les intensités totales
    - »Marqueur fluorescent
    - »Quantité d'ARNm
    - »Paramètres du scanner
  - Gènes de références
  - Gènes dupliqués
  - Blocs
  - Aiguilles (printip)



- **Entre puces appariées**
  - Prendre en compte les dye-swap
- **Entre biopuces**
  - Pour corriger un biais systématique
  - Prendre en compte les données répliquées

Normalisation Simple



Red	Green	Difference	Ratio (G/R)	Log_2 Ratio	Centered R
16500	15104	–1396	0.915	–0.128	–0.048
357	158	–199	0.443	–1.175	–1.095
8250	8025	–225	0.973	–0.039	0.040
978	836	–142	0.855	–0.226	–0.146
65	89	24	1.369	0.453	0.533
684	1368	529	2.000	1.000	1.080
13772	11209	–2563	0.814	–0.297	–0.217
856	731	–125	0.854	–0.228	–0.148

- **Hypothèse: l'ARN total utilisé est le même dans les deux.**
- **Normalisation de l'intensité tot '**

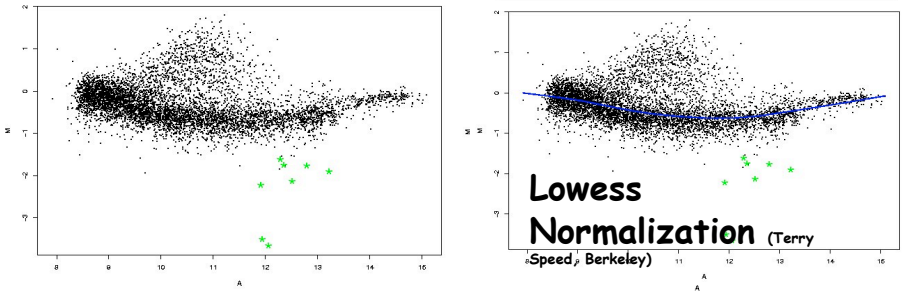
Normalization Factor: $N = \frac{\sum_{k=1}^{Narray} R_k}{\sum_{k=1}^{Narray} G_k}$

Normalization: $G'_k = NG_k$ and $R'_k = R_k$

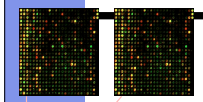
Normalization moins simple

(Terry Speed, Berkeley)

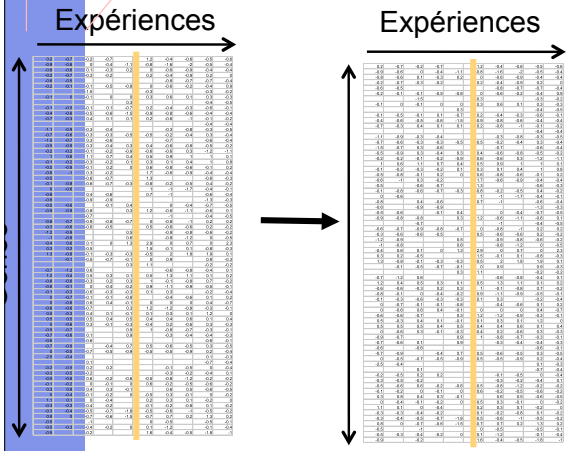
- **Ratio-Intensity plot**
- **$M = \log_2(R/G)$**
- **$A = \log_2(R*G)/2$**
- **M vs. A**



Processus de transformation des données



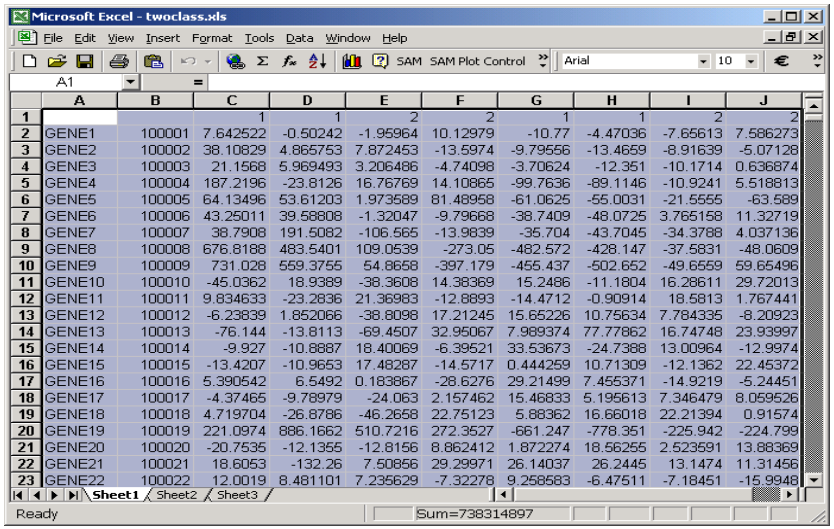
Signal analogique



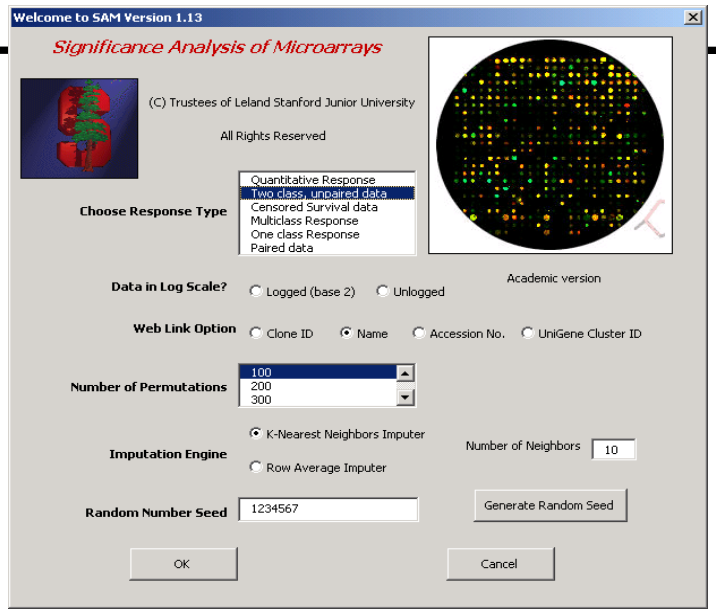
Gènes différentiellement exprimés et évaluation du taux d'erreur.

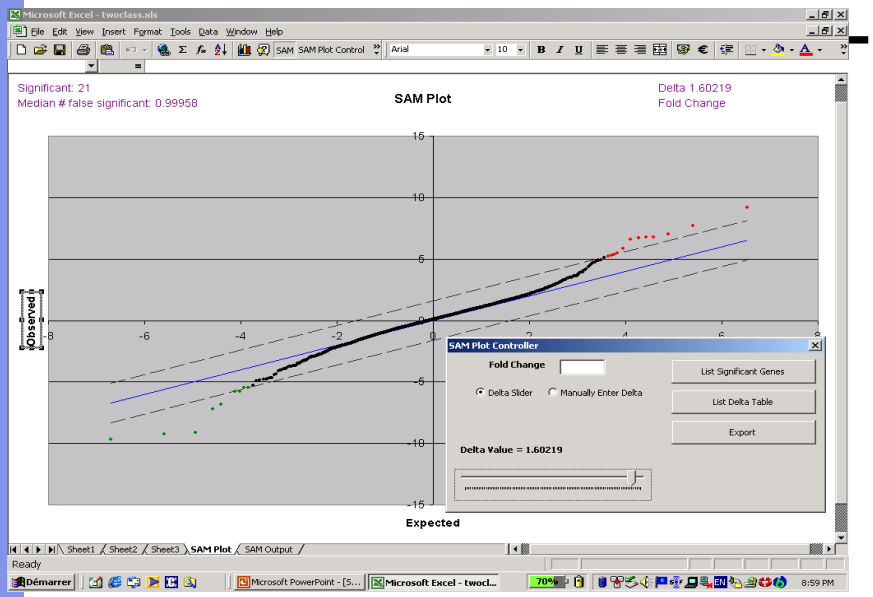
- **FWER: Peu de gènes, très peu de faux positifs mais beaucoup de fau négatifs**
- **FDR: Plus de gènes et de faux positifs mais moins de faux négatifs.**

Outils(comment cela marche): SAM plugin Excel / Package R



	A	B	C	D	E	F	G	H	I	J
1			1	1	2	2	1	1	2	2
2	GENE1	100001	7.642522	-0.50242	-1.95964	10.12979	-10.77	-4.47036	-7.65613	7.586273
3	GENE2	100002	38.10829	4.865753	7.872453	-13.5974	-9.79556	-13.4659	-8.91639	-5.07128
4	GENE3	100003	21.1568	5.969493	3.206486	-4.74098	-3.70624	-12.351	-10.1714	0.636874
5	GENE4	100004	187.2196	-23.8126	16.76769	14.10865	-99.7636	-89.1146	-10.9241	5.518813
6	GENE5	100005	64.13496	53.61203	1.973589	81.48958	-61.0625	-55.0031	-21.5555	-63.589
7	GENE6	100006	43.25011	39.58808	-1.32047	-9.79668	-38.7409	-48.0725	3.765158	11.32719
8	GENE7	100007	38.7908	191.5082	-106.565	-13.9839	-35.704	-43.7045	-34.3788	4.037136
9	GENE8	100008	676.8188	483.5401	109.0539	-273.05	-482.572	-428.147	-37.5831	-48.0609
10	GENE9	100009	731.028	559.3755	54.8658	-397.179	-455.437	-502.652	-49.6559	59.65496
11	GENE10	100010	-45.0362	18.9389	-38.3608	14.38369	15.2486	-11.1804	16.28611	29.72013
12	GENE11	100011	9.834633	-23.2836	21.36983	-12.8893	-14.4712	-0.90914	18.5813	1.767441
13	GENE12	100012	-6.23839	1.852066	-38.8098	17.21245	15.65226	10.75634	7.784335	-8.20923
14	GENE13	100013	-76.144	-13.8113	-69.4507	32.95067	7.989374	77.77862	16.74748	23.93997
15	GENE14	100014	-9.927	-10.8887	18.40069	-6.39521	33.53673	-24.7388	13.00964	-12.9974
16	GENE15	100015	-13.4207	-10.9653	17.48287	-14.5717	0.444259	10.71309	-12.1362	22.45372
17	GENE16	100016	5.390542	6.5492	0.183867	-28.6276	29.21499	7.455371	-14.9219	-5.24451
18	GENE17	100017	-4.37465	-9.78979	-24.063	2.157462	15.46833	5.195613	7.346479	8.059526
19	GENE18	100018	4.719704	-26.8786	-46.2658	22.75123	5.88362	16.66018	22.21394	0.91574
20	GENE19	100019	221.0974	886.1662	510.7216	272.3527	-661.247	-778.351	-225.942	-224.799
21	GENE20	100020	-20.7535	-12.1355	-12.8156	8.862412	1.872274	18.56255	2.523591	13.88369
22	GENE21	100021	18.6053	-132.26	7.50856	29.29971	26.14037	26.2445	13.1474	11.31456
23	GENE22	100022	12.0019	8.481101	7.235629	-7.32278	9.258583	-6.47511	-7.18451	-15.9948





Microsoft Excel - twoclass.xls

B17 = False Significant Number (Median, 90 percentile)

Significant Genes List

Input Parameters

Imputation Engine	10-Nearest Neighbor Imputer
Data Type	Two Class, unpaired data
Data in log scale?	TRUE
Number of Permutations	100
Blocked Permutation?	FALSE
RNG Seed	1234567
(Delta, Fold Change)	(1.38245,)
(Upper Cutoff, Lower Cutoff)	(4.83839, -5.30624)

Computed Quantities

Computed Exchangeability Factor S	3.461219987
S0 percentile	0.01
False Significant Number (Median,	(1.99916, 3.99832)
False Discovery Rate (Median, 90 p	(7.68907, 15.37814)
Pi0Hat	0.99958

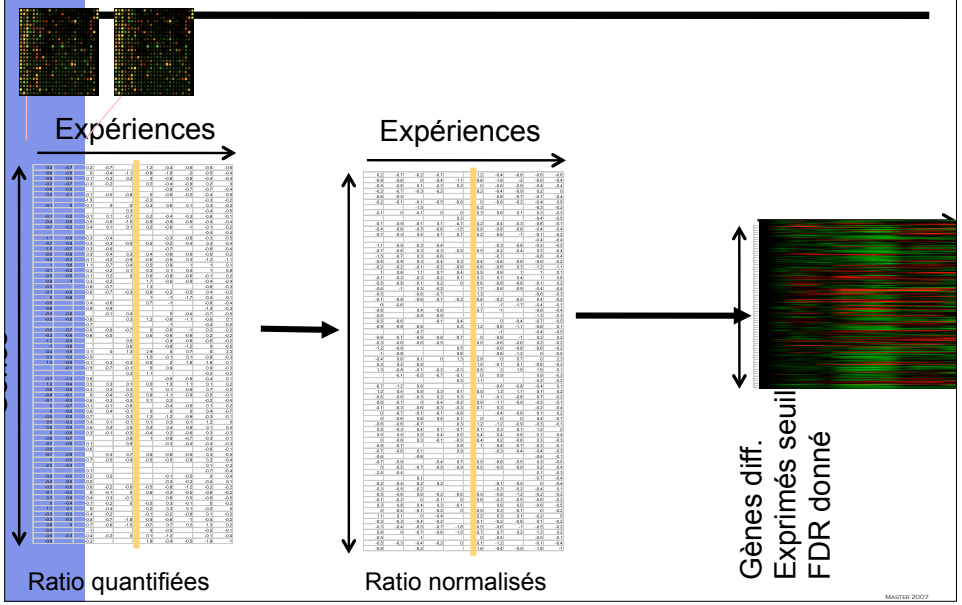
16 Positive Significant Genes

Row	Gene Name	Gene ID	Score(d)	Numerator(r)	Denominator(s+s0)	Fold Change	q-value (%)
2716	GENE2715	102715	9.171384011	1485.912975	162.0162206	∞	4.7599009
6685	GENE6684	106684	7.673095301	225.6693025	29.41046522	0,368,547.75807	4.7599009
3271	GENE3270	103270	7.030539419	117.107345	16.6569502	0,368,547.75807	4.7599009
2864	GENE2863	102863	6.780016457	420.70773	62.05113699	0,368,547.75807	4.7599009
145	GENE144	100144	6.767894012	83.27397	12.30426627	0,368,547.75807	4.7599009
3284	GENE3283	103283	6.710258078	228.3957075	34.03679931	0,368,547.75807	4.7599009
5470	GENE5469	105469	6.604019641	307.17931	46.51399098	0,368,547.75807	4.7599009
1884	GENE1883	101883	5.834754829	101.1551125	17.33665175	0,368,547.75807	4.7599009
277	GENE276	100276	5.452035862	538.21359	98.71791082	0,368,547.75807	4.7599009
4364	GENE4363	104363	5.356665087	427.42526	79.79316479	0,368,547.75807	4.7599009
3149	GENE3148	103148	5.307632206	232.6684675	43.83658446	0,368,547.75807	4.7599009
4599	GENE4598	104598	5.243130841	535.0823025	102.0539671	0,368,547.75807	4.7599009
6090	GENE6089	106089	5.101054836	312.65621	61.29246206	0,368,547.75807	7.6890706

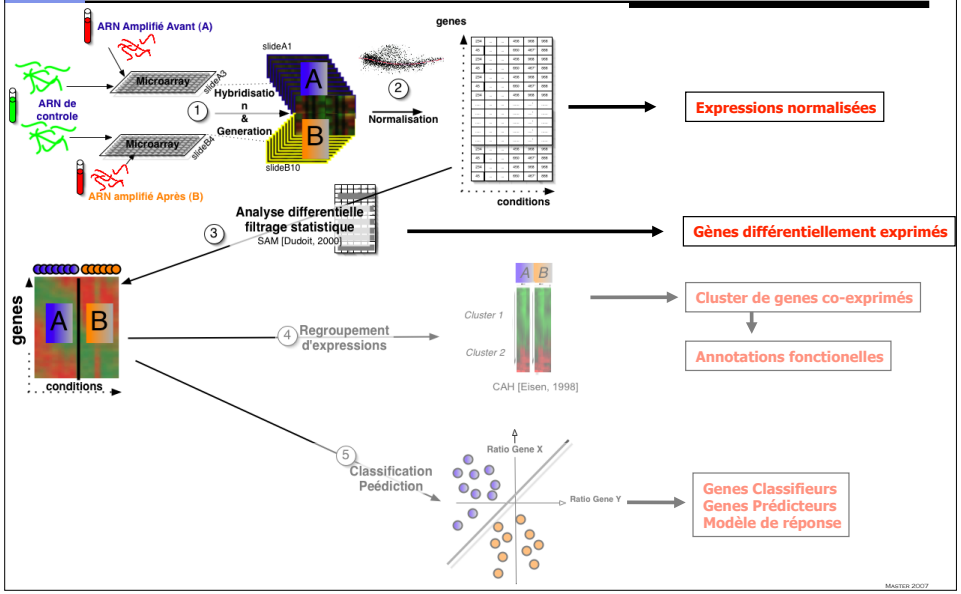
Sheet1 Sheet2 Sheet3 SAM Plot SAM Output

Ready

Processus de transformation des données



Processus d'analyse Données Puces



PLAN

I. A) Généralités B) petite intro à la bioinformatique

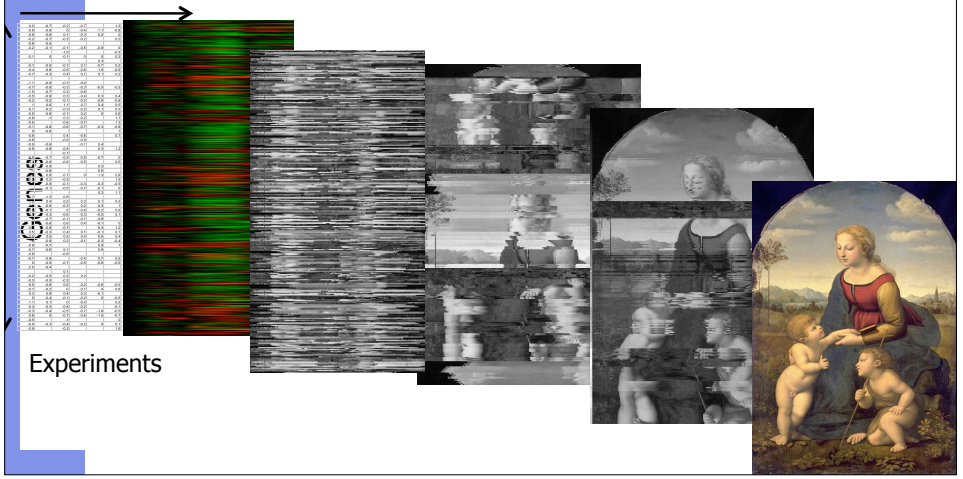
II. Une source de données: les BioPuces

III. La fouille de données BioPuces

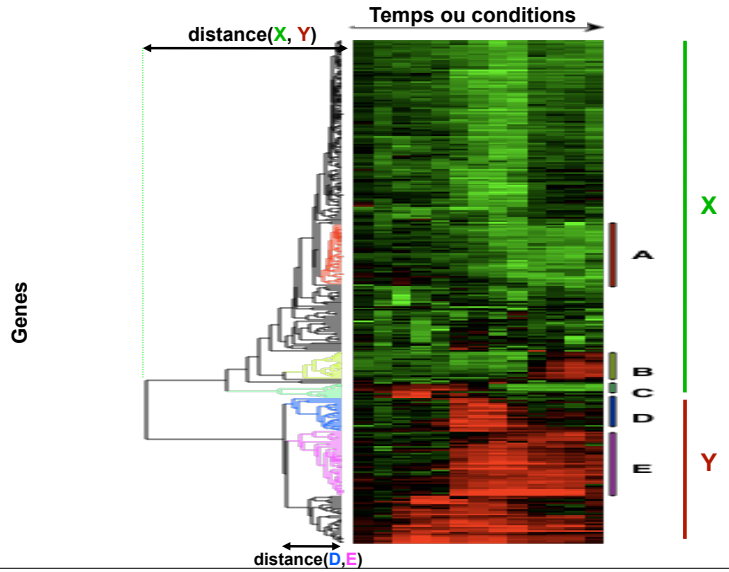
- Clustering
- Classification/Prediction
- Feature Selection

IV. Conclusion, Projets...

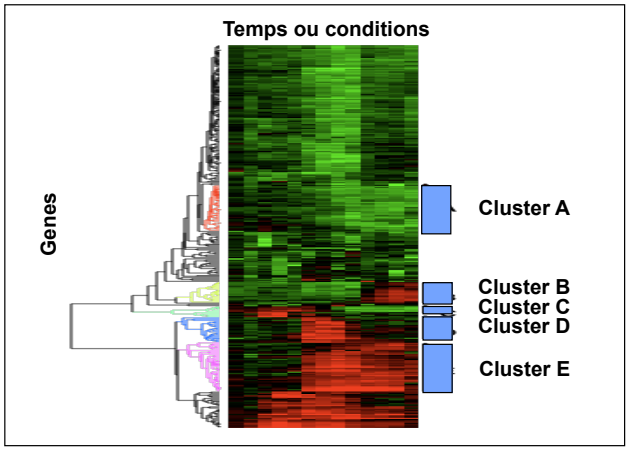
Regrouper des gènes ayant le meme profil d'expression



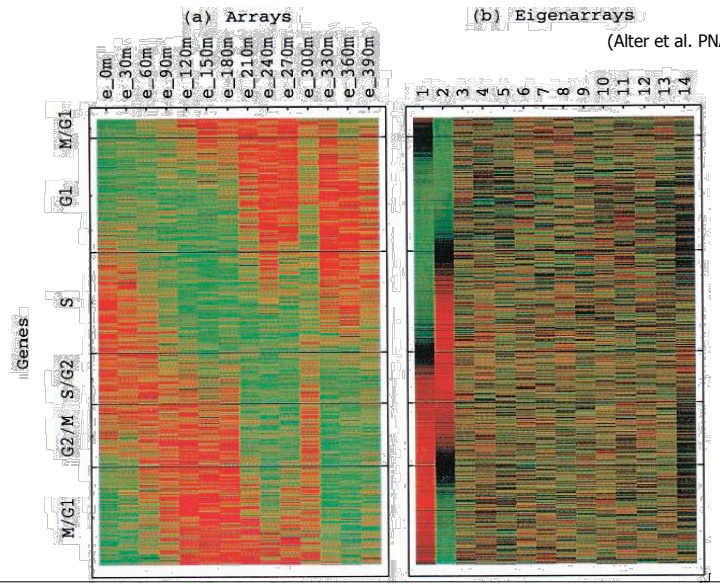
Classification Ascendante Hiérarchique (Cluster [Eisen,98])



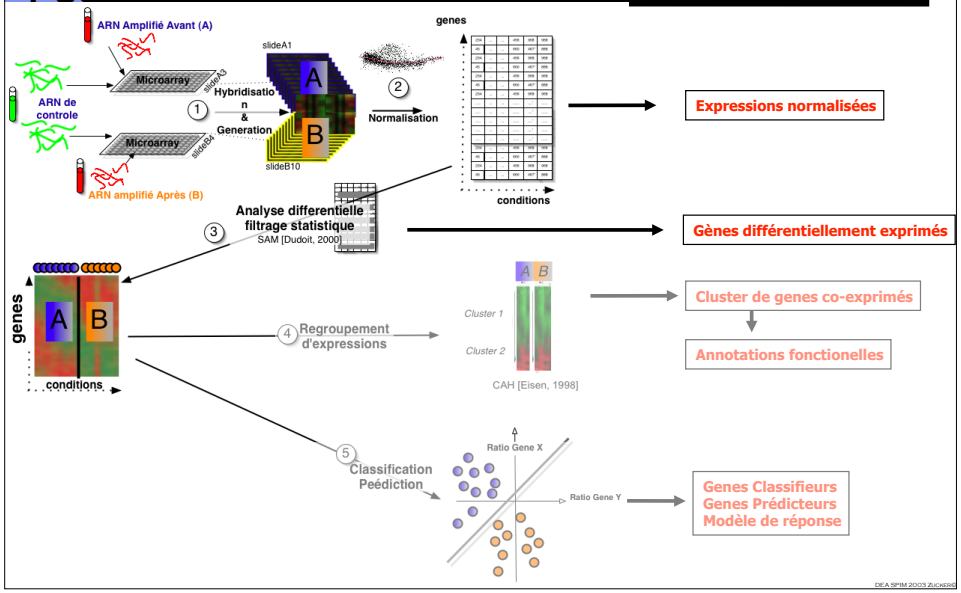
Regrouper des gènes ayant le meme profil d'expression (II)



Analyse en Composante principale: données puces



Processus d'analyse Données Puces



K-Means et R

```
# Suppression des variables préalablement créées
rm(list=ls(all=TRUE))
# Création d'une matrice de données artificielles contenant deux sous-populations
c1 <- matrix(rnorm(100, sd = 0.3), ncol = 2)
c2 <- matrix(rnorm(100, mean = 1, sd = 0.3), ncol = 2)
mat <- rbind(c1, c2)


# Affichage des points de c1 et c2
plot(c1, col="blue", pch=16, xlim=range(mat[,1]), ylim=range(mat[,2]))
points(c2, col="green", pch=16)


# Application de l'algorithme des kmeans
cl <- kmeans(mat, 2, 20)
# Affichage du résultat de kmeans
cl
```


Classification hiérarchique et R

```
# Suppression des variables préalablement créées
rm(list=ls(all=TRUE))
# Chargement des données USArrests
data(USArrests)
# Sélection d'une partie de la base
mat <- USArrests[-c(20:50),]


# Calcul de la matrice de distances
dd <- dist(mat)
# Affichage de la matrice de distances
dd


# Application de l'algorithme de classification hiérarchique
```

Les cartes auto-organisatrices [Kohonen 82]

- Self-Organizing Maps ou SOM, souvent appelés "réseaux de Kohonen"
- Très proches des algorithmes de centres mobiles
  - représentent les données par un nombre réduit de prototypes (appelés neurones dans les SOM)
  - méthode d'apprentissage compétitive :
    - » le "gagnant" = le prototype le plus proche de l'objet donné
- Différence SOM & K-means :
  - préservation de la topologie
  - les objets les plus semblables seront plus proches sur la carte => visualisation claire des groupements

SOM : principes

- Les relations de voisinage entre les neurones définissent une topologie et donc un nouvel espace. 2 espaces :
  - Un espace des entrées dans lequel peuvent être représentés les données et les vecteurs poids des neurones
  - Un espace de sortie (ou carte à une ou deux dimensions) qui contient l'ensembles des neurones et sur lequel une topologie a été définie.
- But de la cartographie associative :
  - Associer chaque vecteur d'entrée à un neurone de la carte
  - On espère que 2 vecteurs proches dans l'espace des entrées exciteront 2 neurones proches sur la carte.

Carte Auto-Organisatrice: Self-organization maps (SOM)

(Kohonen, 1982)

$m_i(t+1)=m_i(t)+h_{ci}(t)[x(t)-m_i(t)]$

where, t is the time index,

$h_{ci}(t)=f(\|r_c-r_i\|,t)$,

r_i and r_c are the location of node c and I.

(Tamayo et al. PNAS, 1999)

Cartes de Kohonen

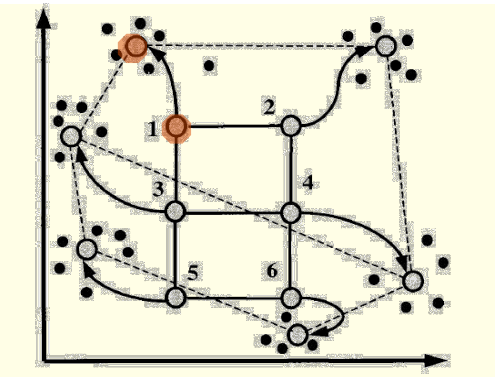
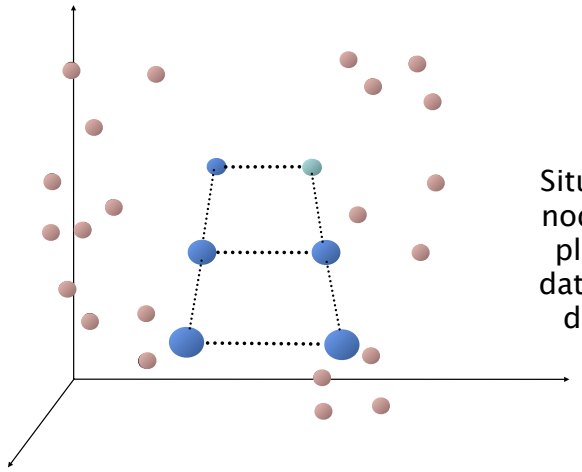


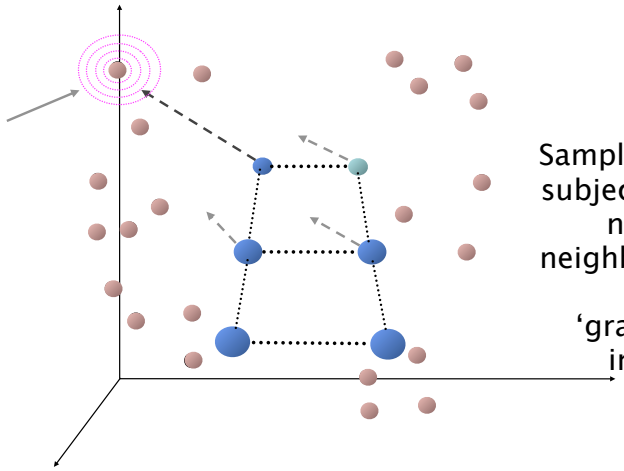
FIG. 1. Principle of SOMs. Initial geometry of nodes in 3 × 2 rectangular grid is indicated by solid lines connecting the nodes. Hypothetical trajectories of nodes as they migrate to fit data during successive iterations of SOM algorithm are shown. Data points are represented by black dots, six nodes of SOM by large circles, and trajectories by arrows.

Self-Organizing Maps



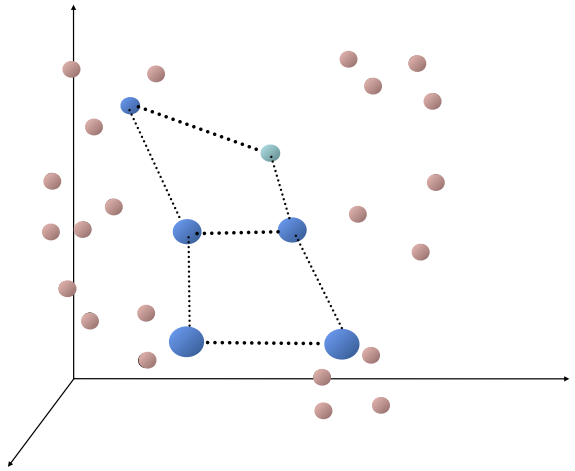
Situate grid of nodes along a plane where datapoints are distributed

Self-Organizing Maps

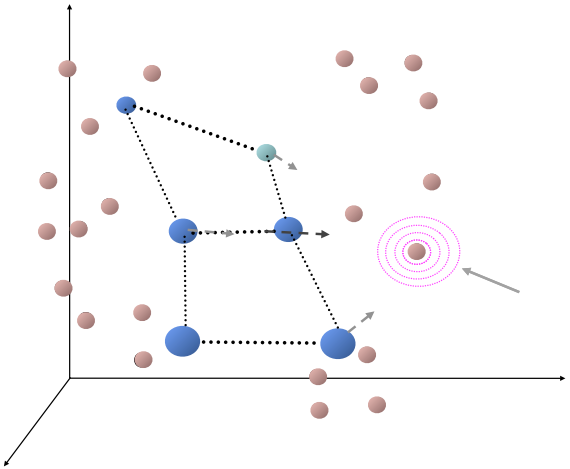


Sample a gene and subject the closest node and neighboring nodes to its 'gravitational' influence

Self-Organizing Maps

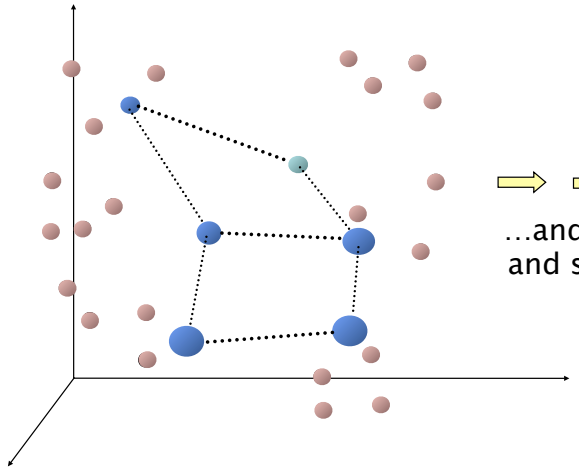


Self-Organizing Maps



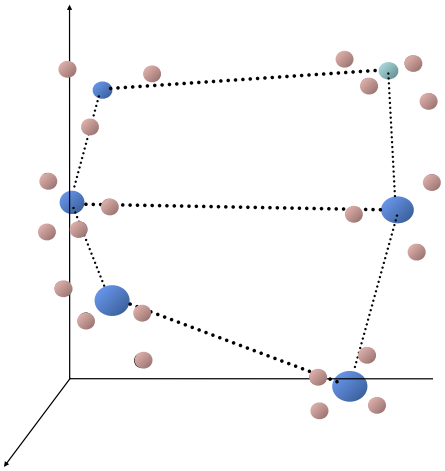
Sample another gene…

Self-Organizing Maps



…and so on, and so on…

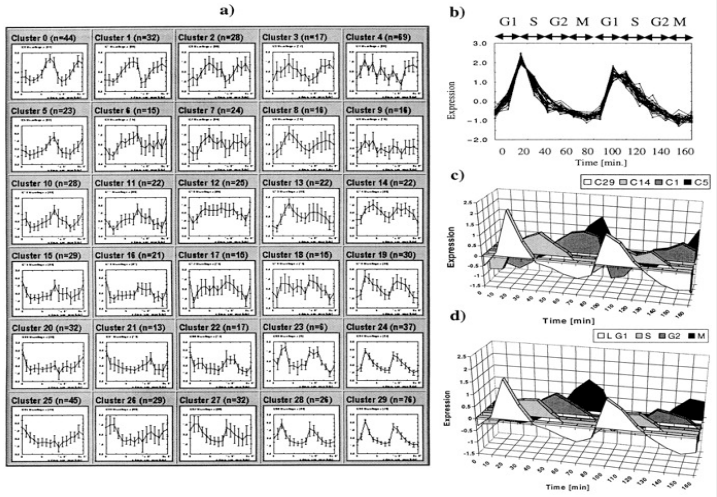
Self-Organizing Maps



…until all genes have been sampled several times over. Each cluster is defined with reference to a node, specifically comprised by those genes for which it represents the closest node.

Self-organizing maps

from Tamayo et al. 1999 (yeast cell cycle data)



SOM : bilan

- **Robuste aux données incomplètes**
- **Ajustable aux très grandes bases de données :**
  - **établissement d'une carte des 78 000 protéines de Swiss-Prot en 2000**
  - **regroupement d'environ 7 millions de documents (demandes de brevets)**
- **Essentiellement un outil d'analyse exploratoire et de visualisation.**
- **Tentatives pour utiliser les SOM à des fin prédictives non concluantes. [Michie et al. 94]**

SOM et R : Statmethods.net

```
# Suppression des variables préalablement créées
rm(list=ls(all=TRUE))
# Chargement de la librairie SOM
library(som)
# Chargement des données yeast
data(yeast)
# Nettoyage des données
yeast <- yeast[, -c(1, 11)]
yeast.f <- filtering(yeast)
yeast.f.n <- normalize(yeast.f)

# Application de SOM avec une carte 5x6
res <- som (yeast.f.n, xdim=5, ydim=6)
# Visualisation du résultat
plot(res)
```

PL AN

A. Les données BioPuces: du signal aux données brutes

B. Stockage et standardisation des données

C. Le traitement statistique des données

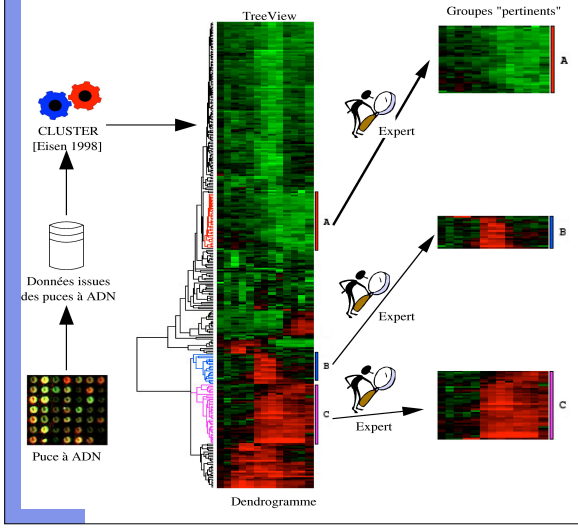
D. Analyse et Fouille de données: Clustering

E. Analyse et Fouille de données: Annotations

F. Analyse et Fouille de données: Prédiction

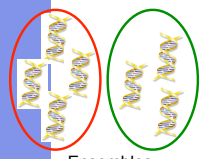
G. V. Conclusions

Faire sens des listes de gènes



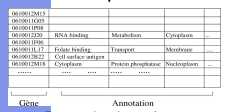
- Identifier les groupes de gènes ayant mêmes profils d'expression
  - Identification par l'expert des groupes « pertinents »
  - Cas particulier : deux groupes de gènes
    - les « up »
    - les « down »

Annotation des gènes



Ensembles de gènes *(un ou deux)*

Annotation



Base de données de gènes annotés *(annotations GO et localisation chromosomique)*

- Annotation manuelle

  - ☹ Aller-retour constant à la littérature
  - ☹ Données en constante évolution
  - ☹ Choix biaisé par l'expert et l'expérimentation
  - ☹ Différents niveaux de généralité des annotations

⇒ Automatisation des annotations

- Annotation automatique

Utilisation des informations disponibles sur le site de SOURCE :

  - Annotations GO
  - Localisation chromosomique

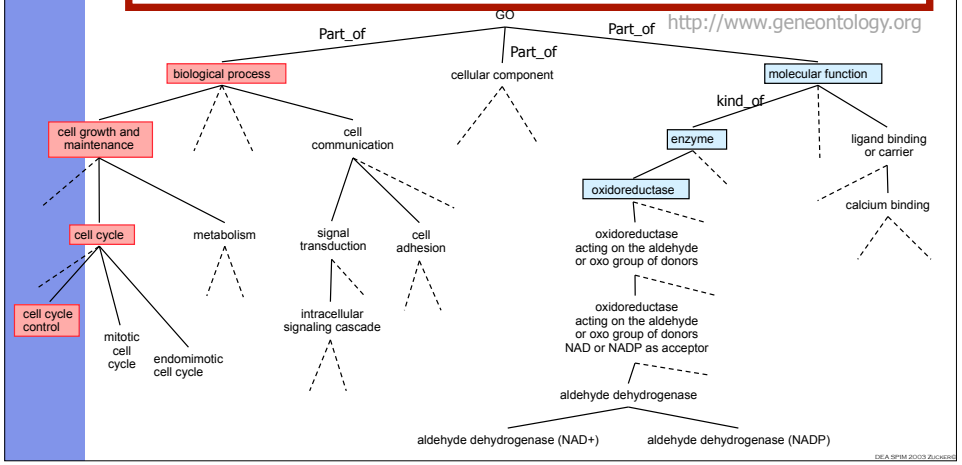
Des annotations longues et souvent biaisées

Réference	% gènes mobilisés	Validation (N =)	Normalisation	Test de multiplicité	Annotation
Nadler 2000	~ 10 %	No	No	No	Manual
Dudoit 2000	~ 10 %	No	Yes	Yes	Manual
Jagoe, 2002	~ 6 %	Yes	No	No	Manual
Suzuki, 2002	~ 1 %	Yes (N=7)	No	No	Manual
Sreekumar,2002	~ 4 %	Yes (N=3)	No	No	Manual
Wada, 2002	~ 10 %	No	No	No	Manual
Lopez, 2003	~ 15 %	Yes (N=3)	No	No	Manual
Ferrante, 2000	~3 %	No	Yes	Yes (FPR)	Manual
Soukas, 2000	~25 %	~20	Yes	No	Manual
Liang, 2001		~50	No	No	Manual

L'ONTOLOGIE GO

The Gene Ontology Consortium. 2000. Nat Genet 25:25-29.

« Biological process »: biological **objective** to which the gene or gene product contributes
« Molecular function »: biochemical **activity** of gene product
« Cellular component »: **place** in the cell where a gene product is active



GENE ONTOLOGY *

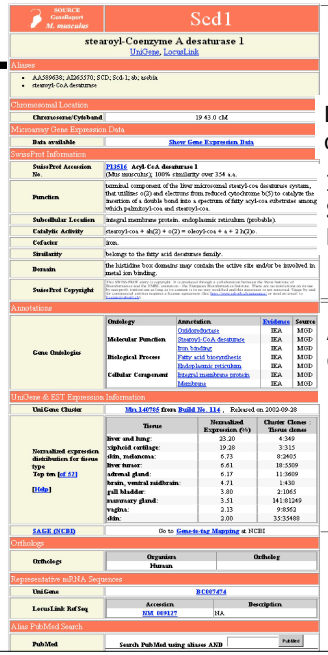
GO:0003673 : Gene_Ontology (149784)
GO:0008150 : biological_process (99849)
GO:0009987 : cellular process (32926)
GO:0007154 : cell communication (9155)
GO:0007267 : cell-cell signaling (1416)
GO:0046879 : hormone secretion (61)
GO:0030072 : peptide hormone secretion (32)
GO:0030073 : insulin secretion (21)
GO:0050796 : regulation of insulin secretion (3)
GO:0046883 : regulation of hormone secretion (22)
GO:0050796 : regulation of insulin secretion (3)
GO:0007582 : physiological process (62723)
GO:0050791 : regulation of physiological process (7543)
GO:0051046 : regulation of secretion (54)
GO:0046883 : regulation of hormone secretion (22)
GO:0050796 : regulation of insulin secretion (3)
GO:0046903 : secretion (749)
GO:0046879 : hormone secretion (61)
GO:0030072 : peptide hormone secretion (32)
GO:0030073 : insulin secretion (21)
GO:0050796 : regulation of insulin secretion (3)
GO:0046883 : regulation of hormone secretion (22)
GO:0050796 : regulation of insulin secretion (3)
GO:0051046 : regulation of secretion (54)
GO:0046883 : regulation of hormone secretion (22)
GO:0050796 : regulation of insulin secretion (3)
GO:0050789 : regulation of biological process (12540)
GO:0050791 : regulation of physiological process (7543)
GO:0051046 : regulation of secretion (54)
GO:0046883 : regulation of hormone secretion (22)
GO:0050796 : regulation of insulin secretion (3)

DES GÈNES À GO

http://source.stanford.edu/cgi-bin/sourceSearch

- Permet un accès simultané à différentes sources d'informations sur les gènes
- Interrogation en-ligne pour plusieurs gènes simultanément

Source - HTML

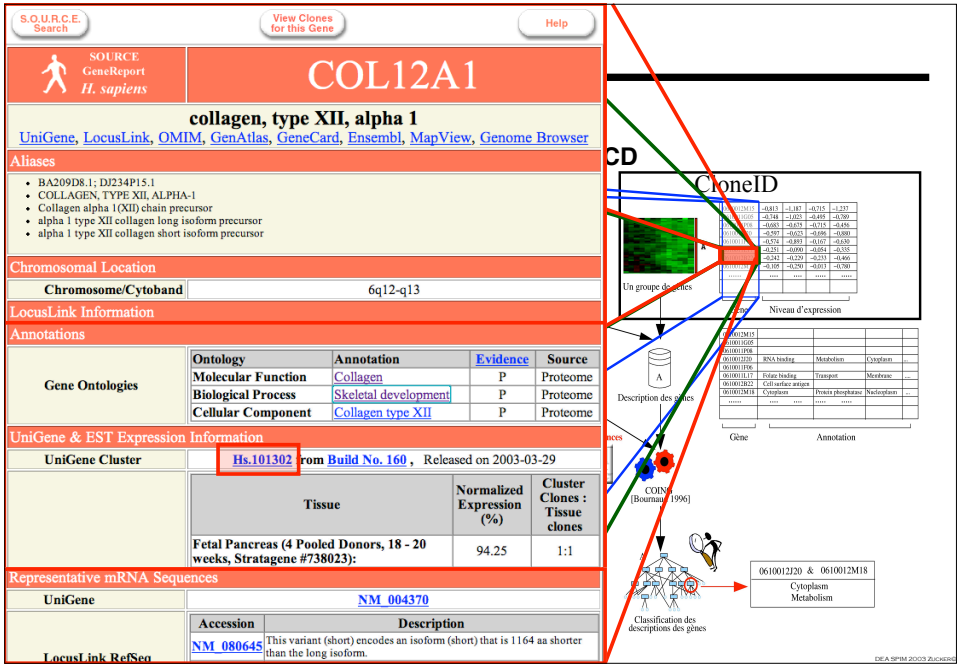


Le nom du gène

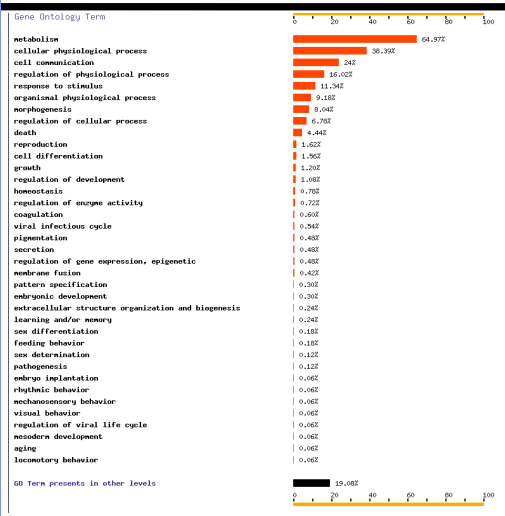
Localisation du gène

Informations Sur les protéines liées au gène

Annotation GO du gène
Ex : membrane



FatiGO *



- Approche mono-variable → annotations GO
- Analyse centrée sur un niveau de profondeur ontologique prédéfini
- Le résultat → listes d'annotations GO ordonnées selon la couverture génique

*Al-Shahrour, F. et al. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **2004**.

Autres Ressources disponibles

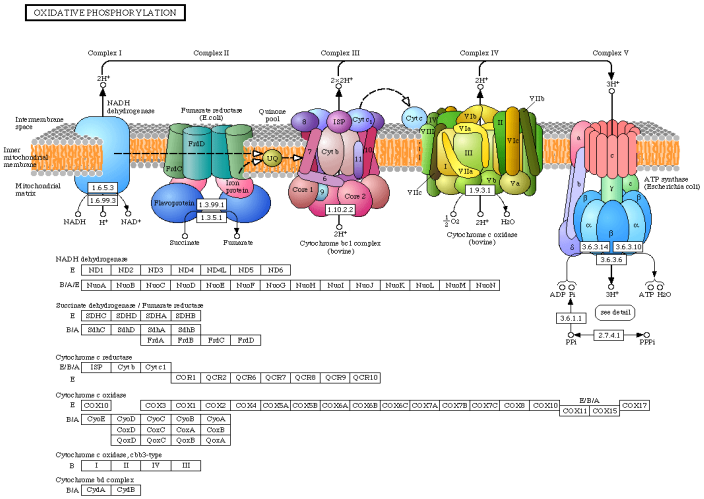
- Représentation formelle de la connaissance biologique: Gene Ontology, SwissProt Keywords, KEGG Pathways
- Bases d'annotations génomiques
  - « Bas niveau » basées sur terminologies / ontologies
    - Gene Ontology (EBI, NCBI – LocusLink, Unigene, …)
    - SwissProt (EBI – UniProt / InterProt) …
  - « Haut niveau » - modélisation des réseaux de régulation biologique (KEGG, BBID, GenMAPP, BioCarta)
- Autres annotations
  - positionnement au sein du génome (cartes QTL, etc.)
  - structure nucléotidique
  - structure (et fonctions) des protéines représentatives
  - niveau d'action (cellulaire, tissues, organes) …

KEGG (1)

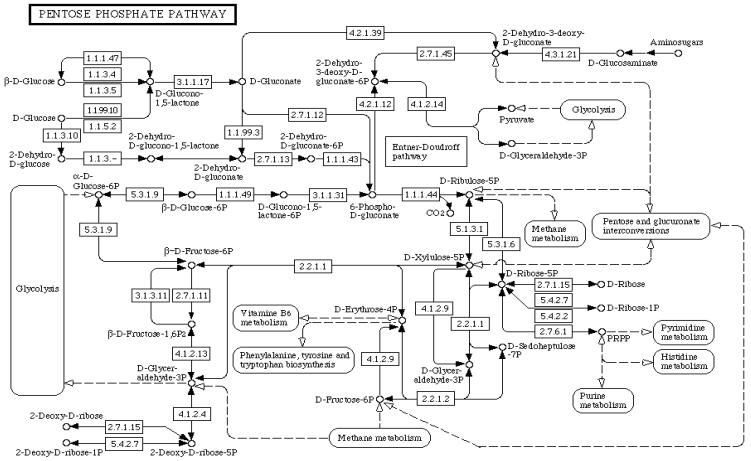
- Base d'annotations fonctionnelles génomiques
- Informations sur le rôle des gènes (enzymes, protéines) dans les réseaux de régulation biologiques
  → modèles des voies métaboliques et de régulation
  - Avantages: précision et richesse de l'information modélisée
  - Limites: modélisation difficile, couverture limitée

*Kanehisa, M. et al. The KEGG databases at GenomeNet. *Nucleic Acids Res* **2002**.

KEGG (2)



KEGG (3)



PLAN

A. Les données BioPuces: du signal aux données brutes
B. Stockage et standardisation des données
C. Le traitement statistique des données
D. Analyse et Fouille de données: Clustering
E. Analyse et Fouille de données: Annotations
F. Analyse et Fouille de données: Prédiction
G. Conclusions

BioCarta (2)

BioCarta (3)

PLAN

I. A) Généralités B) petite intro à la bioinformatique

II. Une source de données: les BioPuces

III. La fouille de données BioPuces

- Clustering
- Classification/Prediction
- Feature Selection

IV. Conclusion, Projets...

Classer et Prédire: algorithmes

- ***But :*** **utiliser les données d'expression pour construire des modèles prédictifs ou des classeurs (par ex: arbre de décisions,, réseaux de neurones)**
- ***Difficultés:*** **peu d'exemples (conditions), souvent < 100, beaucoup d'attributs (gènes), souvent > 1,000**
- **Problématique d'apprentissage et statistique [MLJ, 2003]:**

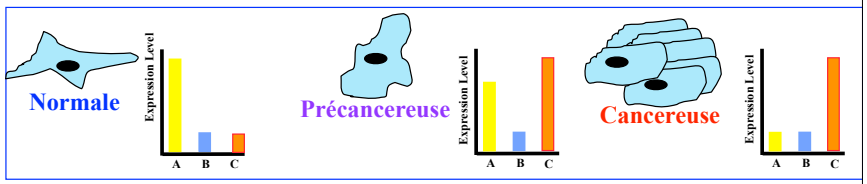
  réduction de dimensions (sélection de genes)
  adapter les algorithmes

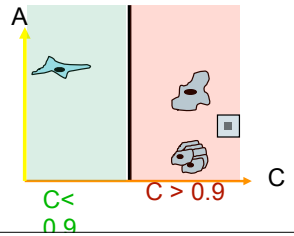
  - **Réseaux de neurones**
  - **Machine à vecteurs de support (SVM)**
  - **Arbre de décision**
  - **Random Forest**
  - **Arbre de régression**
  - **K plus proche voisins**

Environnement: R, Bioconductor, Clémentine©, SAS©, MeV, BRB, TIGR©

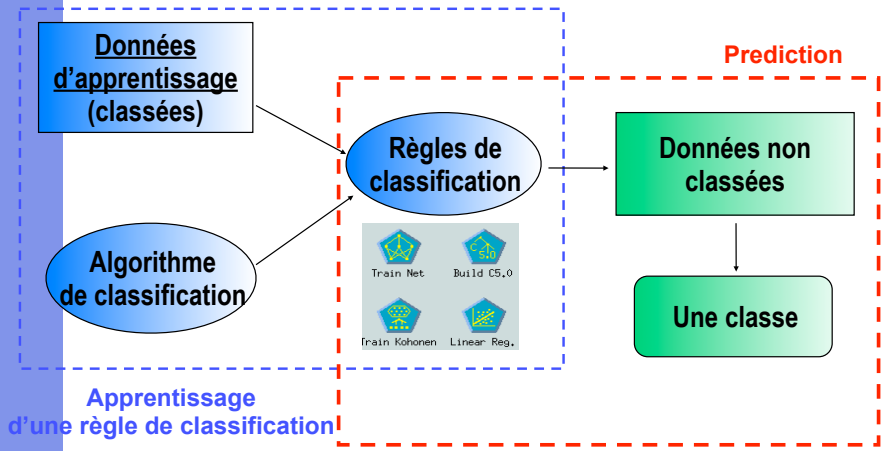
Analyse pour la Classification et

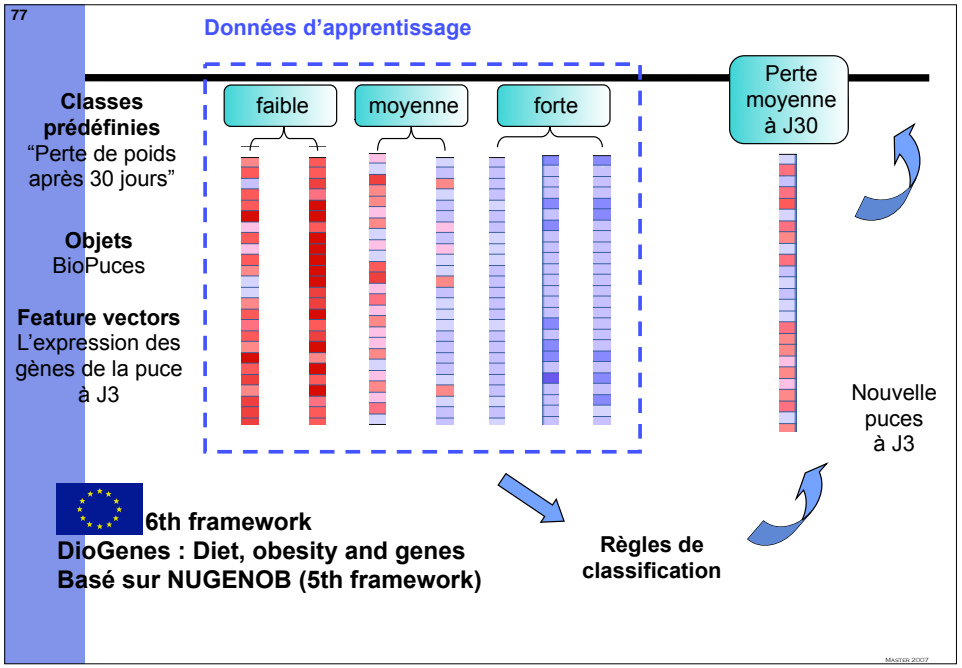
- Problème: **Construire des modèles pour la prédiction et classification à partir de données d'expression.**
- Utilisation : Identification de gènes cibles (prédicteurs), Modèle prédictifs (d'un risque), Modèle de classification (traitement)





Construction automatique de classfieurs



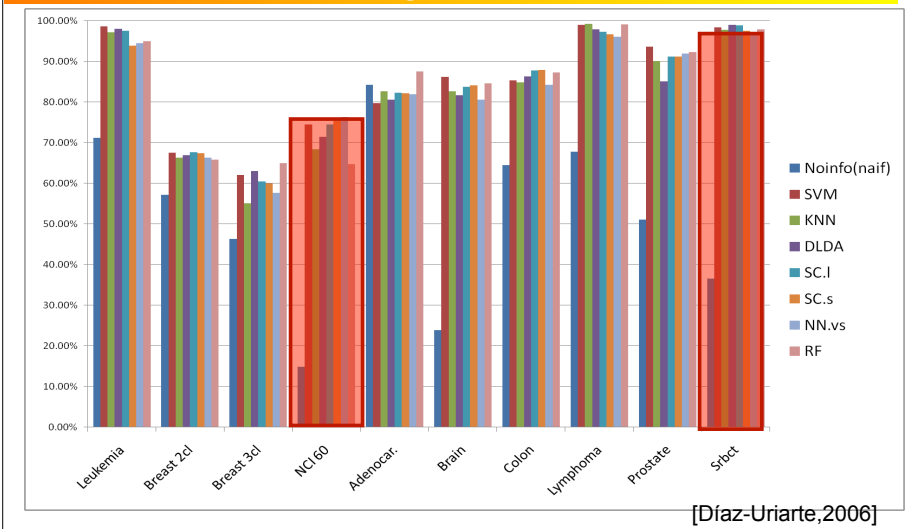


Bases d'apprentissage de la littérature

Nombre d'attributs = O([NB exemples]^2)

Dataset	Original Ref.	AttributsGenes	Exemples Patients	Classes
Leukaemia	[44]	3051	38	2
Breast	[9]	4869	78	2
Breast	[9]	4869	96	3
NCI 60	[61]	5244	61	8
Adenocarcinoma	[62]	9868	76	2
Brain	[63]	5597	42	5
Colon	[64]	2000	62	2
Lymphoma	[65]	4026	62	3
Prostate	[66]	6033	102	2
Srbct	[67]	2308	63	4

Résultats d'apprentissage sur ces bases



[Díaz-Uriarte,2006]

Problématique pour l'apprentissage

- **Peu d'exemple** (50-100 biopuces)
- **Nombreux d'attributs** (2,000-40,000 gènes)
  Une majorité d'attributs non pertinents

- « Malédiction de la dimension »
  - Performances des classeurs dégradées
  - Modèles très complexes
  - Temps de calcul importants
  - Interprétation biologique difficile

- Réduction de dimension nécessaire ((ROC [Mamitzuka, 2006, Pattern Recognition], Incremental Wrapper [Ruiz et al. 2006], Revue [Guyon & Elisseeff, JMLR2003])
  3 types de méthodes dans le contexte des biopuces (Filter, Wrapper, Reformulation)
- Probleme des estimateurs...

PLAN

I. A) GÉNÉRALITÉS B) PETITE INTRO À LA BIOINFORMATIQUE

II. UNE SOURCE DE DONNÉES: LES BIOPUCES

III. LA FOUILLE DE DONNÉES BIOPUCES

- CLUSTERING
- CLASSIFICATION/PREDICTION
- LES RÉSEAUX

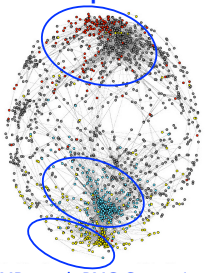
IV. CONCLUSION...

Réseaux d'interaction / co-expression

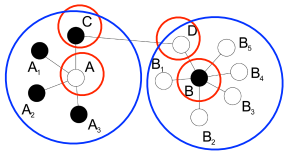
- **Illustration des interactions moléculaires** dans la cellule
- **Hautement dynamiques** et responsives à des stimuli environnementaux variés
- Conditionnés par une **multitude de facteurs** :
  - transcriptionnels
  - épigénétiques
  - évolutionnaires …
- **Modèles structurés** d'interaction biologique

Architecture des réseaux génomiques

***S. Cerevisiae* – Réponse au stress**



Carlson MR, et al. BMC Genomics 2006

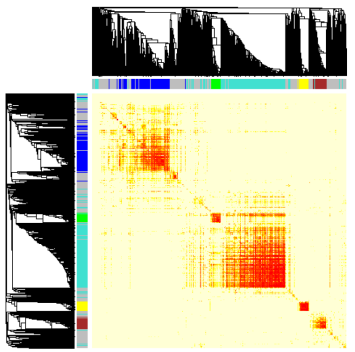


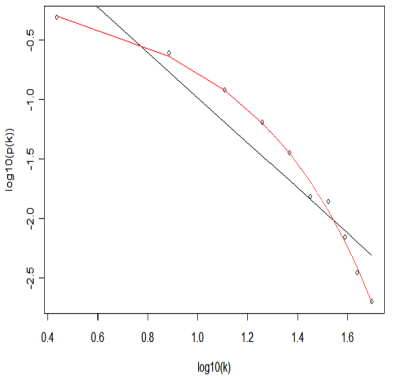
Guimerà R, et al. Nature 2005

- **Architecture modulaire « free scale »** [Jeong H, *et al.* Nature 2001 ; Ravasz E, *et al.* Science 2002]
  - modules d'interaction transcriptomique
  - petit nombre de hubs hautement interconnectés [Guimerà R, *et al.* Nature 2005]
    - hubs locaux (intra-modulaires)
    - hubs globaux (inter-modulaires)

Architecture « scale free »

Modularité **Connectivité**





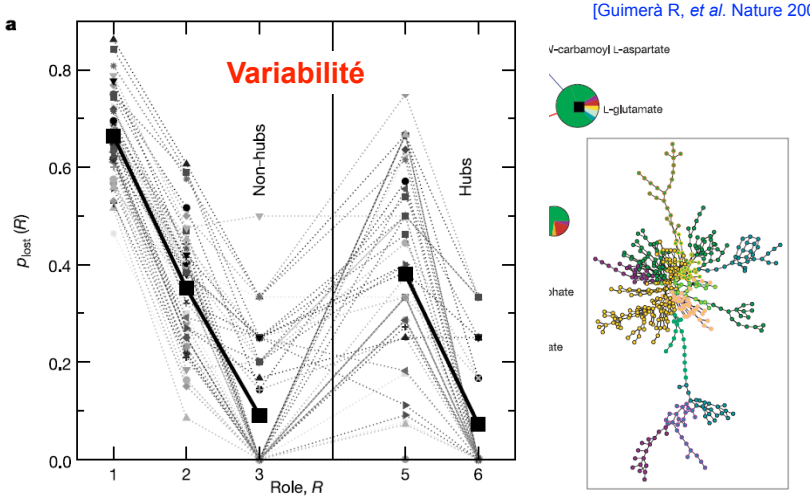
Zhang B, Horvath S. Stat. App Gen. Mol. Biol. 2005

Signification biologique

- **Architecture modulaire** ⇒ avantage évolutif [Hartwell LH, et al. Nature 1999]
  - redondance fonctionnelle ⇒ robustesse
  - adaptabilité environnementale
  - résultat de la sélection naturelle (trial & error)
- **Hubs transcriptionnels** ⇒ signification évolutionnaire distincte [Guimerà R, et al. Nature 2005]
  - hubs inter-modulaires (globaux) ⇒ conservation phylogénique : gènes développementaux, facteurs de transcription, protéines clé
  - hubs intra-modulaires (locaux) ⇒ hautement variables (potentiel évolutif)

Variabilité & conservation phylogénique

[Guimerà R, *et al.* Nature 2005]



Signification biologique

- **Architecture modulaire** ⇒ avantage évolutif [Hartwell LH, et al. Nature 1999]
  - redondance fonctionnelle ⇒ robustesse
  - adaptabilité environnementale

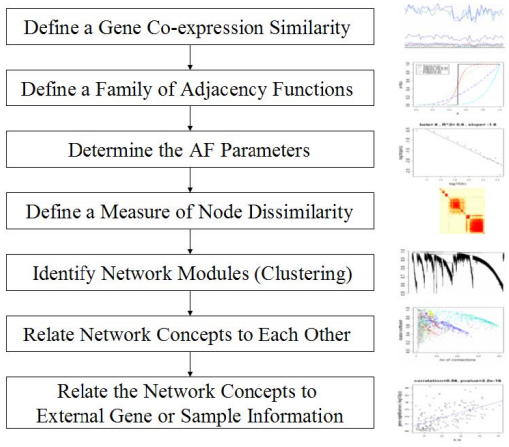
  ⇒ résultat de la sélection naturelle (trial & error)
- **Hubs transcriptionnels** ⇒ signification évolutionnaire dichotomique [Guimerà R, et al. Nature 2005]
  - hubs inter-modulaires (globaux) ⇒ conservés dans la phylogénie: gènes développementaux, facteurs de transcription, protéines clé
  - hubs intra-modulaires (locaux) ⇒ hautement variables (potentiel évolutif)

Approche conventionnelle

Deux étapes :

1. Inférence à partir des données d'expression

2. Analyse biologique



Zhang B, Horvath S, Stat. App. Gen. Mol. Biol. 2005

Inférence des réseaux

- Mesures des **interactions transcriptomiques**
  - similarité des profils d'expression ⇒ fonction d'adjacence
    - corrélation des profiles d'expression
    - analyse de l'information mutuelle
  - discrètes : estimation d'un seuil de significativité
  - continues : dépendantes du modèle de réseau « scale free »
- Mesures de la **connectivité** en réseau (dis/similarité des nœuds)
  - similarité topologique des interactions en réseau

⇒définition des concepts du réseau : modules d'interaction & hubs

Signification biologique

- **Annotation fonctionnelle** & comparaison à des données de la littérature
  - ⇒ analyse fonctionnelle des modules
  - ⇒ investigation du rôle biologique des hubs
- Analyse **phylogénique** des séquences géniques
  - ⇒ scores Blast [Featherstone DE, *et al*. Bioessays 2002]
- **Modèles expérimentaux** de validation biologique
  - ⇒ e. g. knock-out/down génique chez la levure/cultures cellulaires

PLAN

Génomique Fonctionnelle/transcriptomique

• **Une hypothèse: l'approche « pangénomique » plutot que « gène candidat »**

• **Le défis biotechnologique: mesurer simultanément l'expression de milliers de genes (ou tous les gènes) —> les puces à ADN**

• **Les défis biologiques et médicaux:**
  - **Découvrir les fonctions des gènes d'après l'expressions.**
  - **Elucider les voies métabolique à partir de l'expression des gènes.**
  - **… aider au diagnostique**

• **Les directions de recherche en bioinformatique du transcriptome:**
  - **Analyse de données différentielles. Tests multiples (FDR)**
  - **Représentation des données d'expression et normalisation.**
  - **Outils de visualisation.**
  - **Fouille de données. Analyse de données exploratoires.**
  - **Intégration de bases de données, de connaissances, d'ontologies, …**