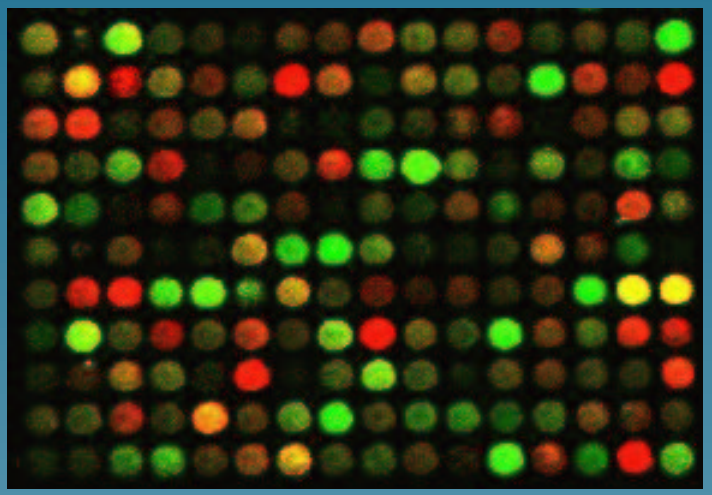
Analyse du transcriptome



Emmanuel Barillot, Franck Rapaport
Jean-Philippe Vert et Andrei Zinovyev

Institut Curie et Ecole des Mines de Paris

Journée ACI IMPBIO, Lyon, France, July 5, 2005.

Plan

1. Introduction à l'analyse du transcriptome
2. Une approche pour l'analyse de voies métaboliques
3. Le projet Kernelchip : cancer et régulation

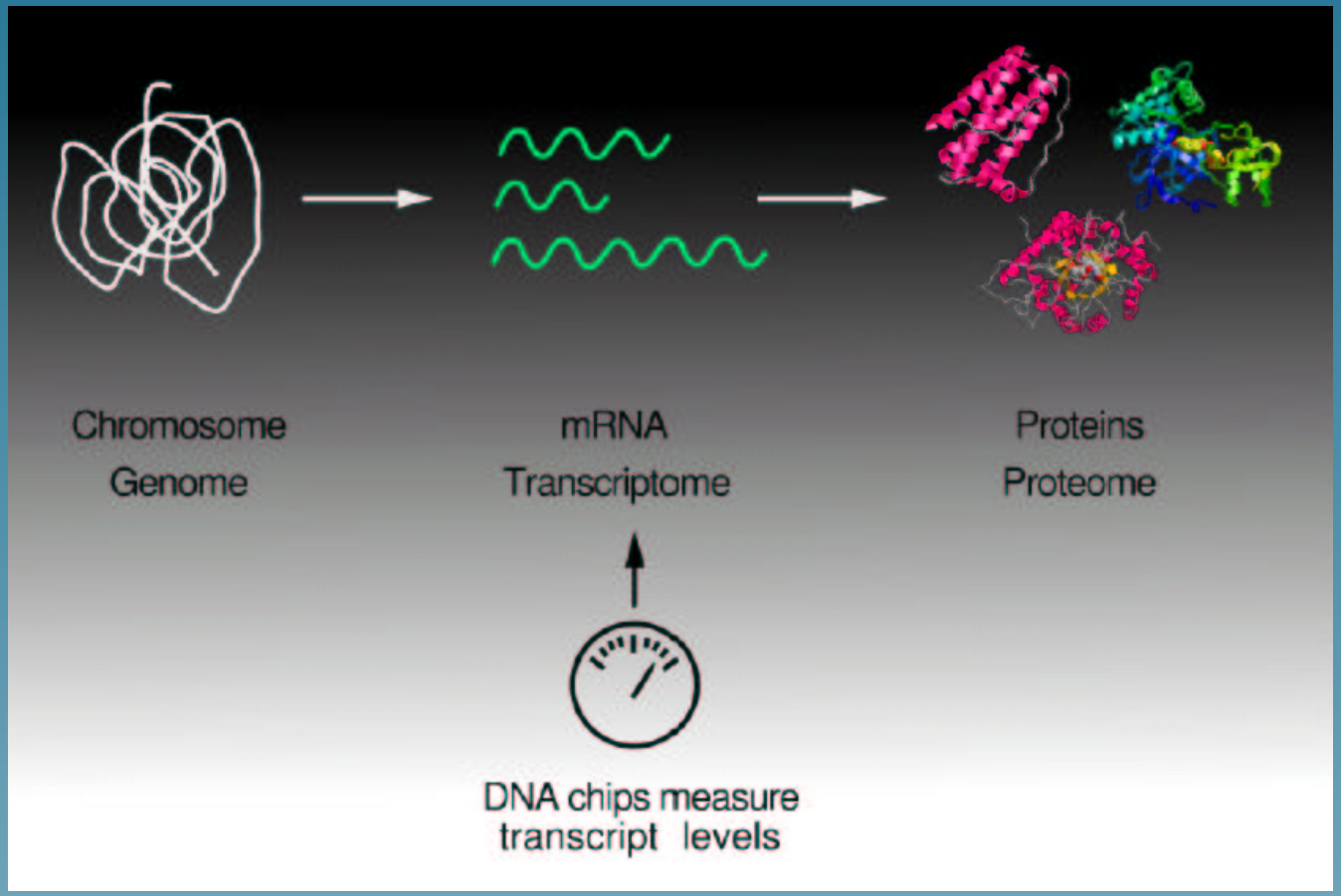
Partie 1

Introduction à l'analyse du transcriptome

En bref...

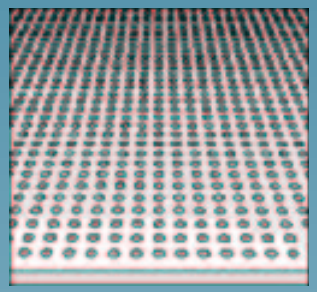
- Le génome humain contient environ 25,000 gènes, codant 100,000+ protéines
- Comprendre la vie = comprendre comment ces protéines interagissent et sont régulées ?
- Les puces à ADN mesurent la quantité d'ARNm (presque les protéines...) pour tous les gènes simultanément, à un instant donné.

Les puces à ADN mesurent l'ARNm



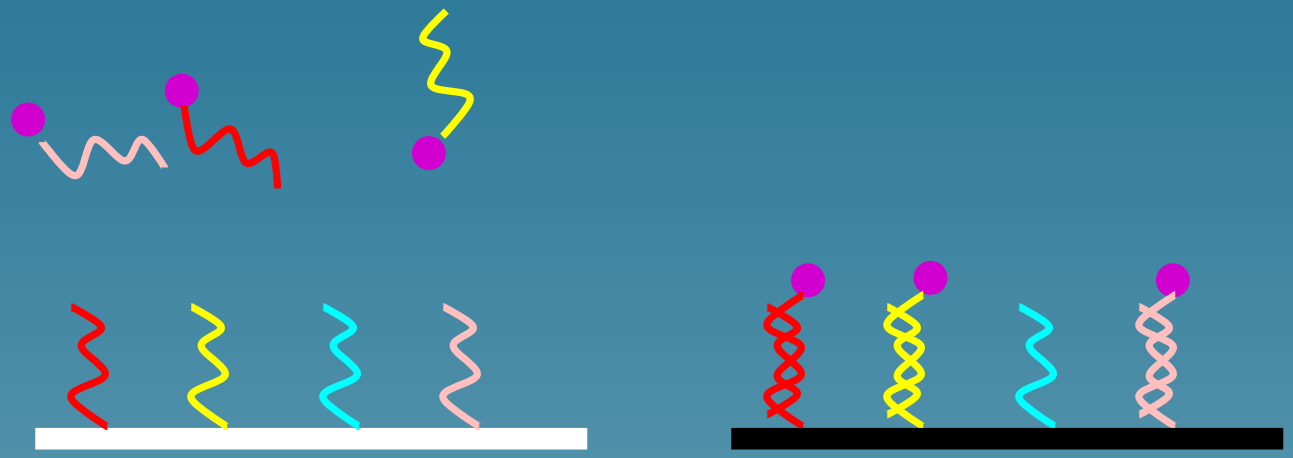
Une puce à ADN concrètement

- Un grand nombre de molécules d'ADN fixées sur un substrat (verre, nylon, ou silicon)
- De 100 à 300,000 spots

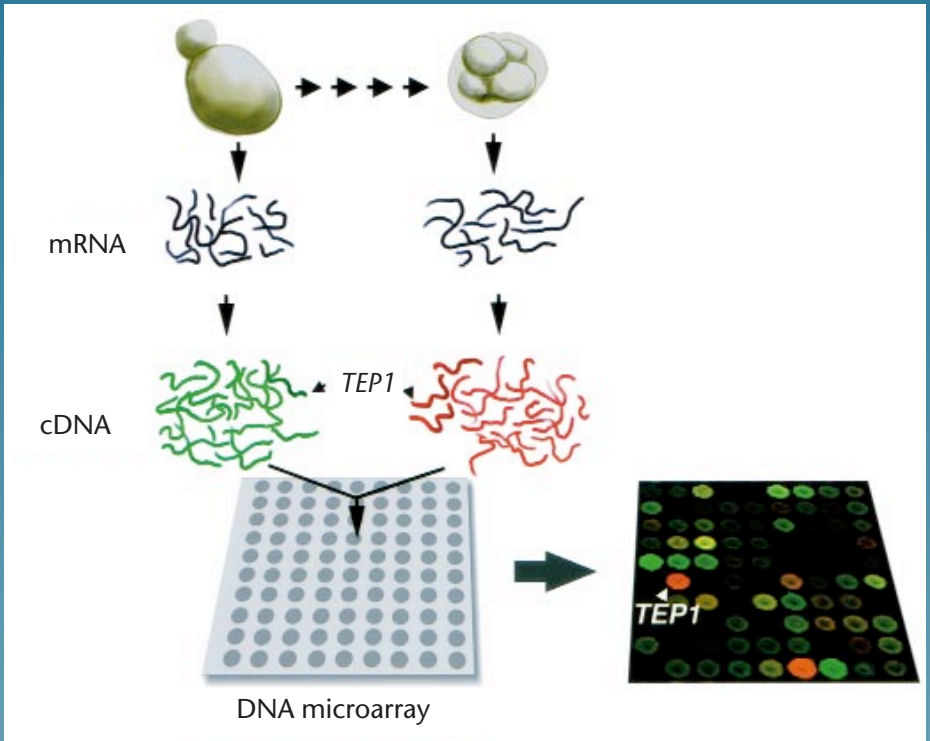


Affymetrix GeneChip® probe array. Image courtesy of Affymetrix.

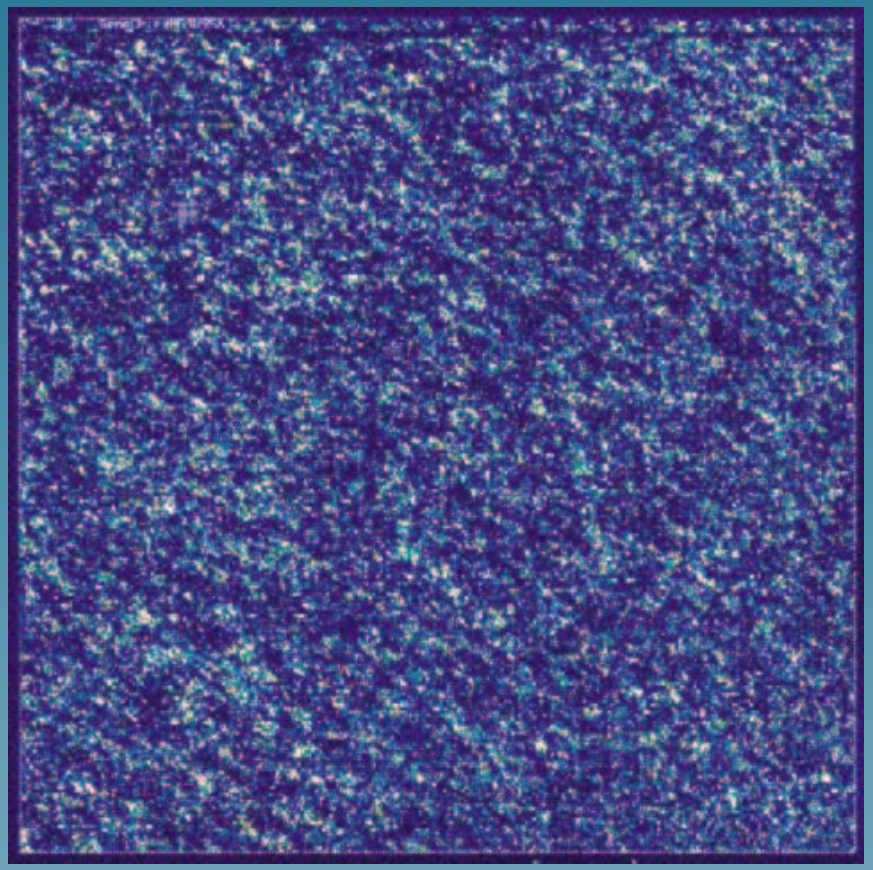
Le principe : hybridation

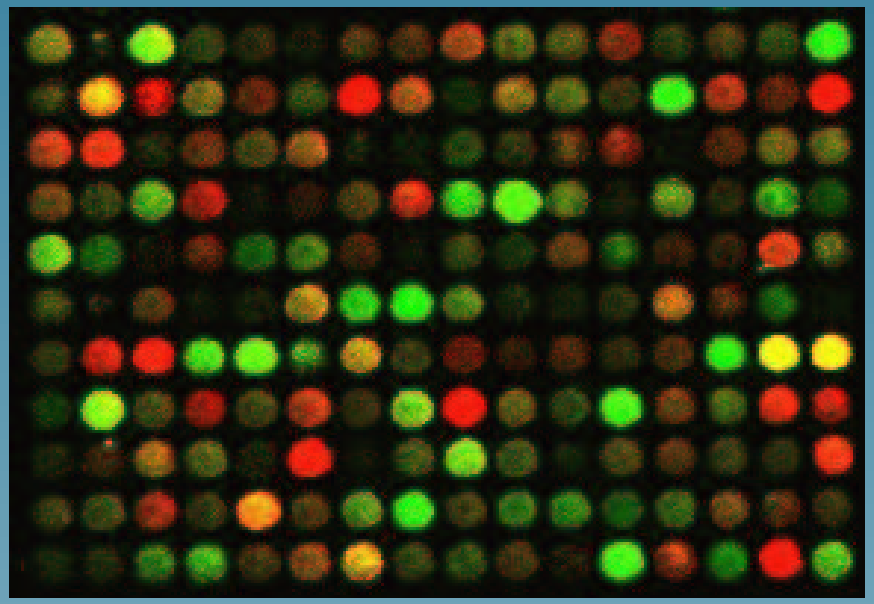


Exemple: hybridation comparative



Le résultat





Le transcriptome

Le transcriptome reflète

- la source du tissue, l'organe, le type de cellules
- l'activité et l'état du tissue:
  - ⋆ état de développement, croissance, mort
  - ⋆ cycle cellulaire
  - ⋆ malade / sain
  - ⋆ réponse à des thérapie

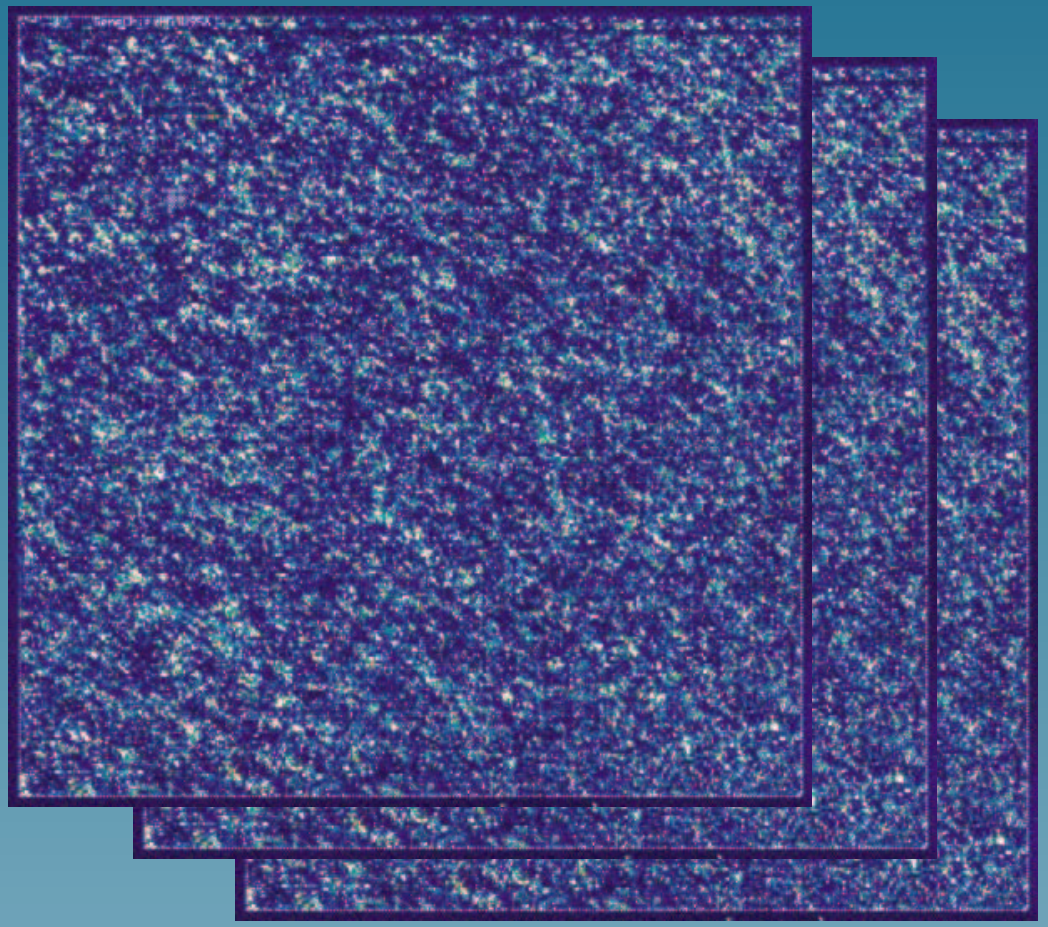
Les espoirs

- découverte de cibles thérapeutiques
- diagnostic et pronostic médical
- pharmacogénomique
- biologie des systèmes etc...

Analyse typique du transcriptome

- Analyse d'image, normalisation
- Détection de gènes différentiellement exprimés
- Analyse exploratoire, clustering
- Analyse discriminante
- Reconstruction de réseaux génétiques

Analyse d'image, normalisation

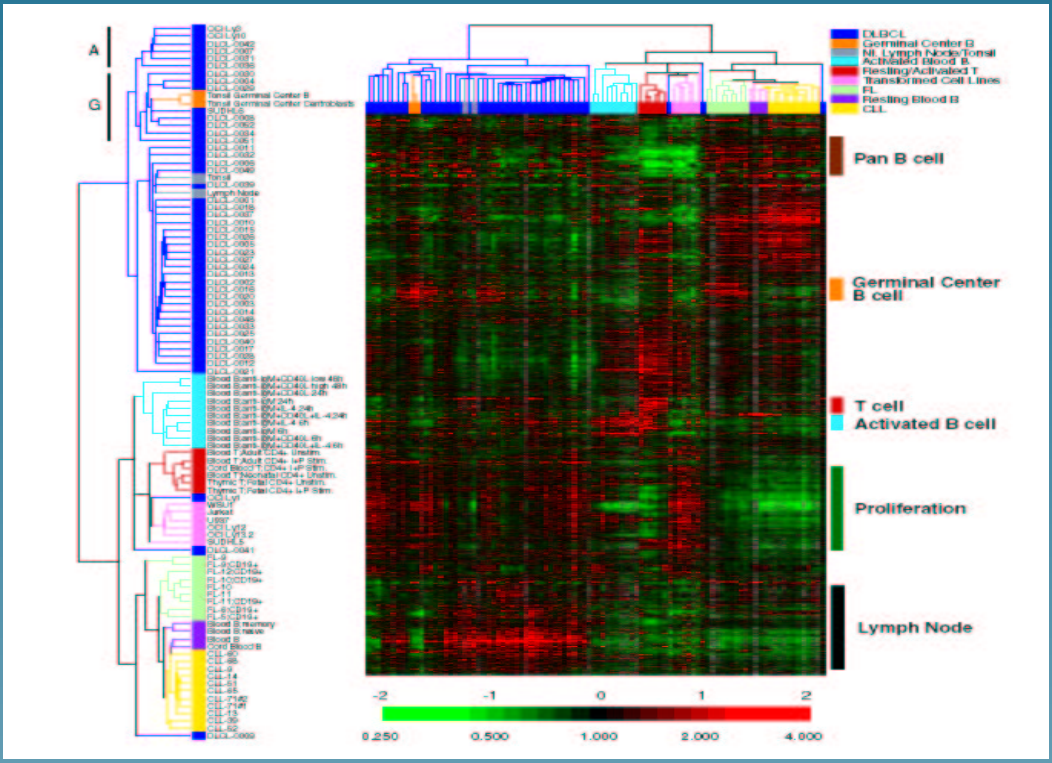


Experiments

Genes

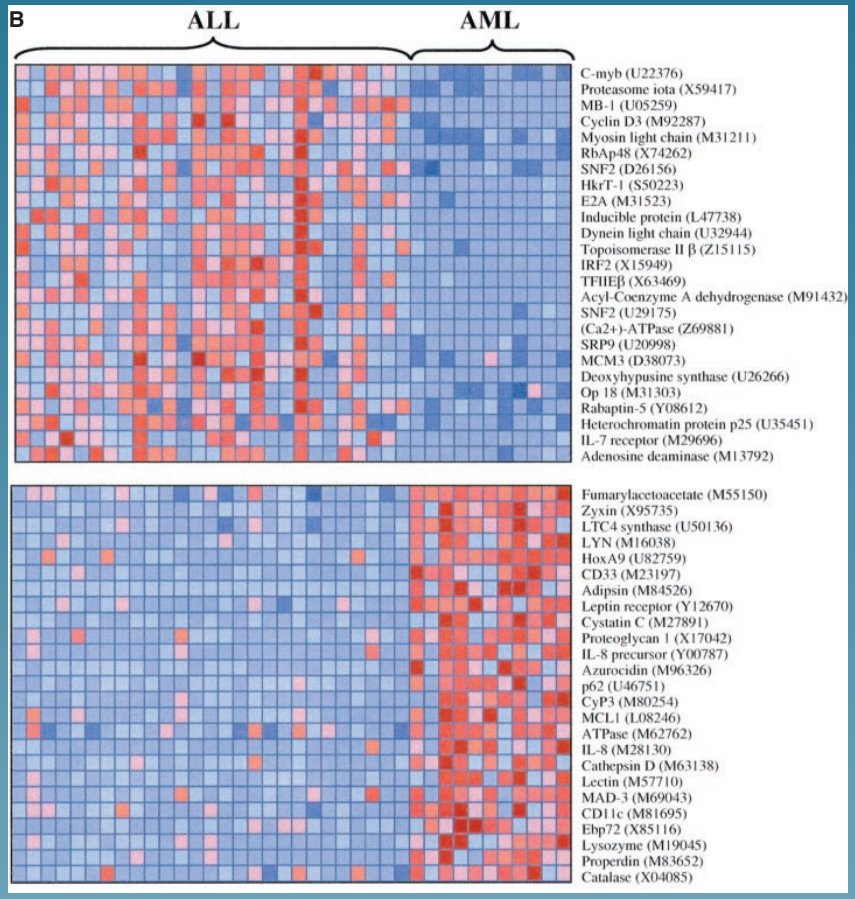
1.5	−2	0.2	3.4	−2.1	...
−4	2.1	0.5	1.1	0.9	...
...	...	...	...		

Exemple d'analyse exploratoire



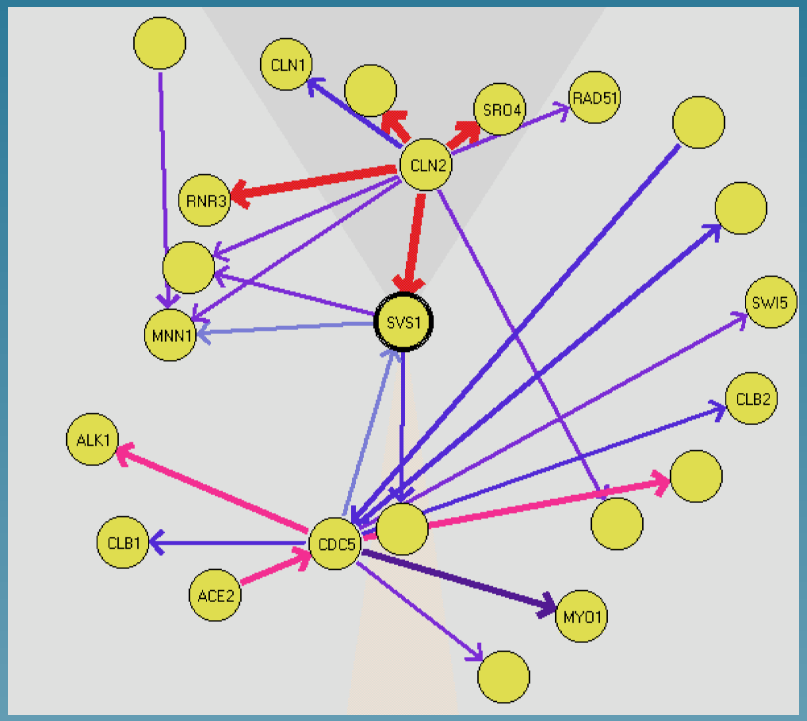
(Alizadeth et al., 2000)

Exemple d'analyse discriminante



(Golub et al., 1999)

Exemple de reconstruction de réseau

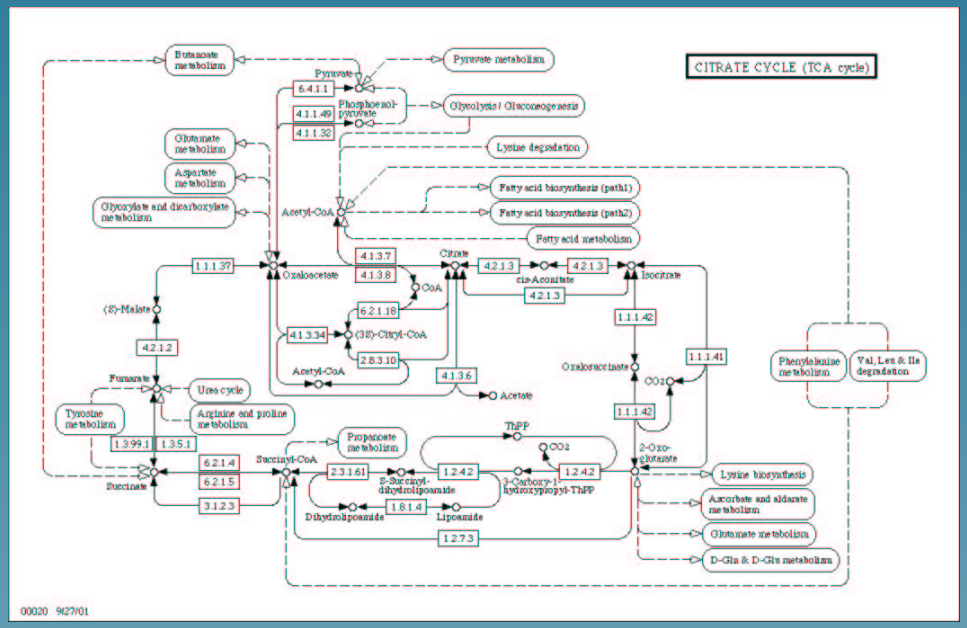


(Friedman et al., 2000)

Partie 2

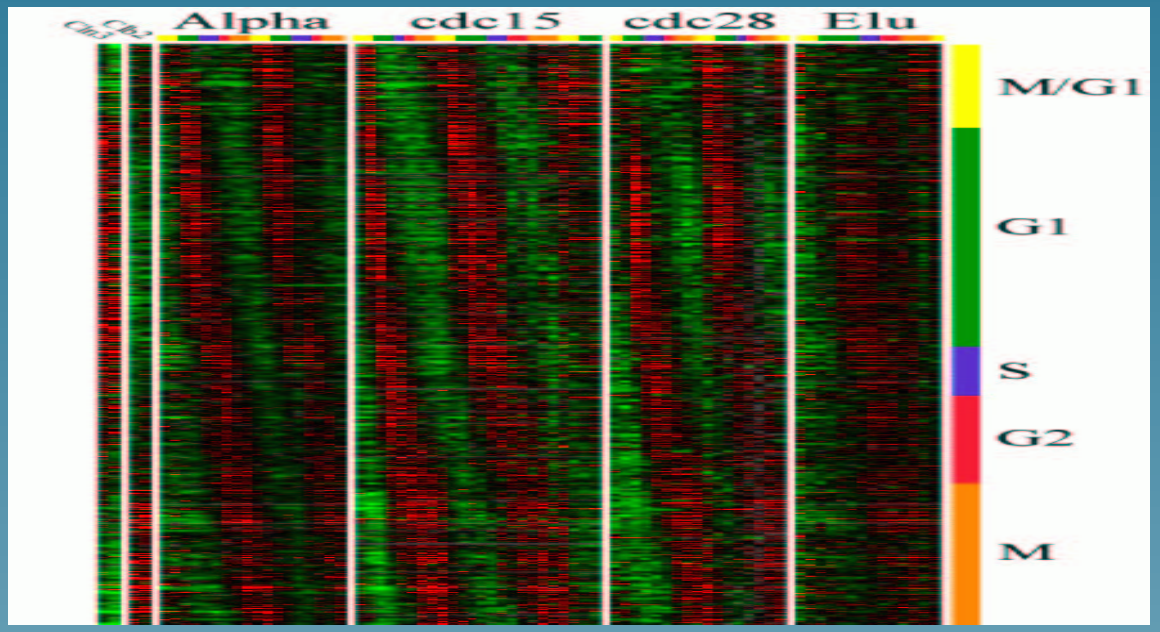
Une approche pour l'analyse de voies métaboliques

Motivation : de nombreuses voies métaboliques sont connues



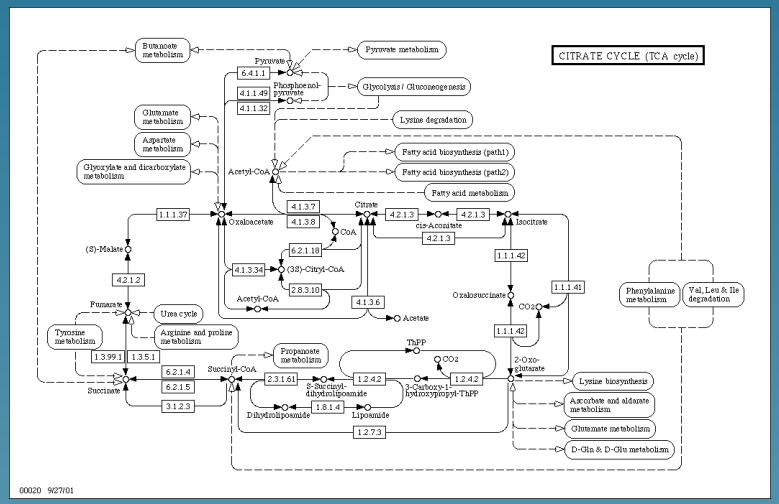
From http://www.genome.ad.jp/kegg/pathway

Les puces à ADN mesurent la dynamique de l'expression

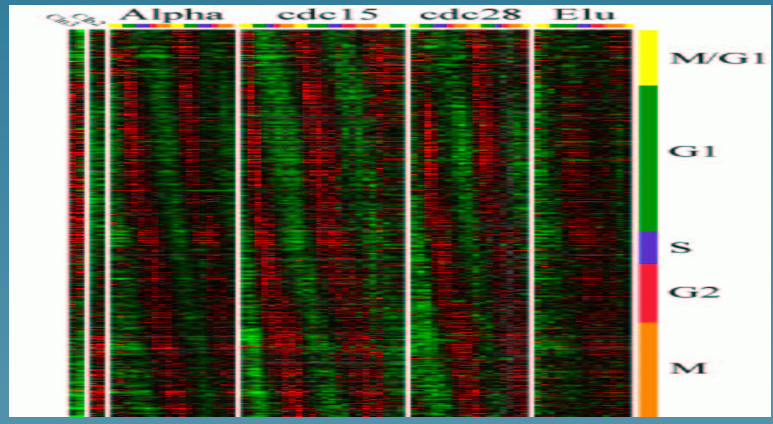


(From Spellman et al., 1998)

Question : comment les comparer?



VS



Détecter l'activité des voies? Trouver de nouvelles voies?

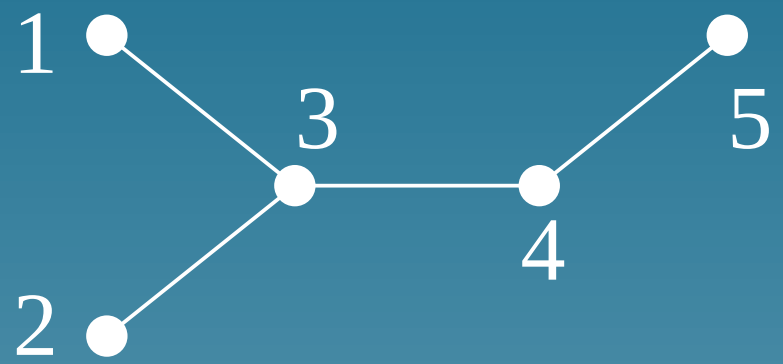
Astuce mathématique ($NIPS'02$)

- n gènes
- Expressions : $f = (f_1, \ldots, f_n)^\top \in \mathbb{R}^n$
- Un graphe G de gènes définit une nouvelle géométrie Euclidienne sur les profiles d'expression par la formule:

$$||f||_G^2 = f^\top L_G f,$$

où L_G est le Laplacien du graphe.

Laplacien du graphe



$$L_G = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

Propriétés du Laplacien

- L est une matrice symétrique, à valeur propres positives, donc $||f||_G$ est bien une norme Euclidienne.
- Elle vérifie:

$$||f||_G^2 = f^\top L_G f = \sum_{i \sim j} (f(i) - f(j))^2$$

donc f a une petite norme si f varie lentement le long des arètes du graphes.

Pourquoi $||f||_G$?

- Les voies métaboliques sont des composantes connexes du graphe
- Contrôler $||f||_G$ assure que f varie peu au sein de voies métaboliques potentielles
- Plusieurs problèmes se formulent naturellement à partir de cette norme

Exemple 1 : Régression régularisée

- Supposons qu'à chaque donnée d'expression $x_i \in \mathbb{R}^n$ soit associée une covariable $y_i \in \mathbb{R}$ (âge, développement d'une tumeur, niveau de pollution, ...)
- La régression par moindre carrés classique cherche un vecteur $\hat{w} \in \mathbb{R}^n$ qui minimise:

$$\min_{w \in \mathbb{R}^n} \sum_{i=1}^{p} \left(w^\top x_i - y_i \right)^2 .$$

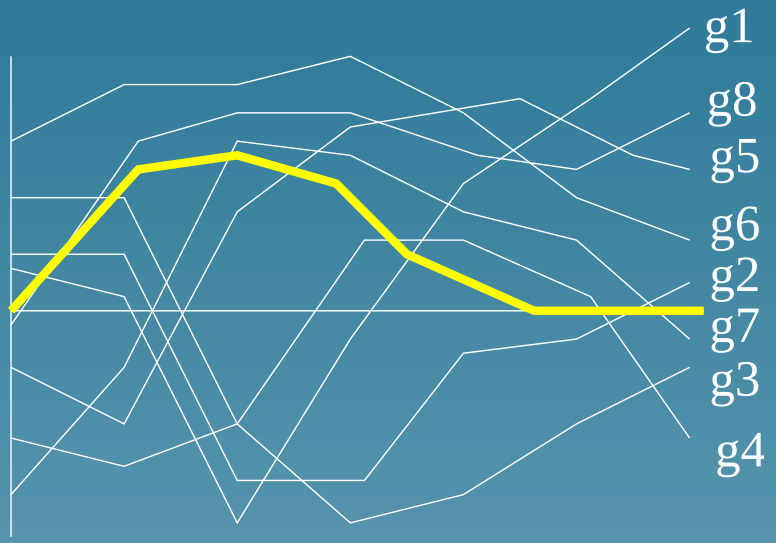
Exemple 1 : Régression régularisée (cont.)

- Si on sait que w doit mettre en valeur certaines voies métaboliques, il est plus efficace de minimiser:

$$\min_{w \in \mathbb{R}^n} \sum_{i=1}^{p} \left(w^\top x_i - y_i \right)^2 + \lambda ||w||_G^2.$$

- Avantages : meilleures propriétés statistiques, meilleure interprétabilité

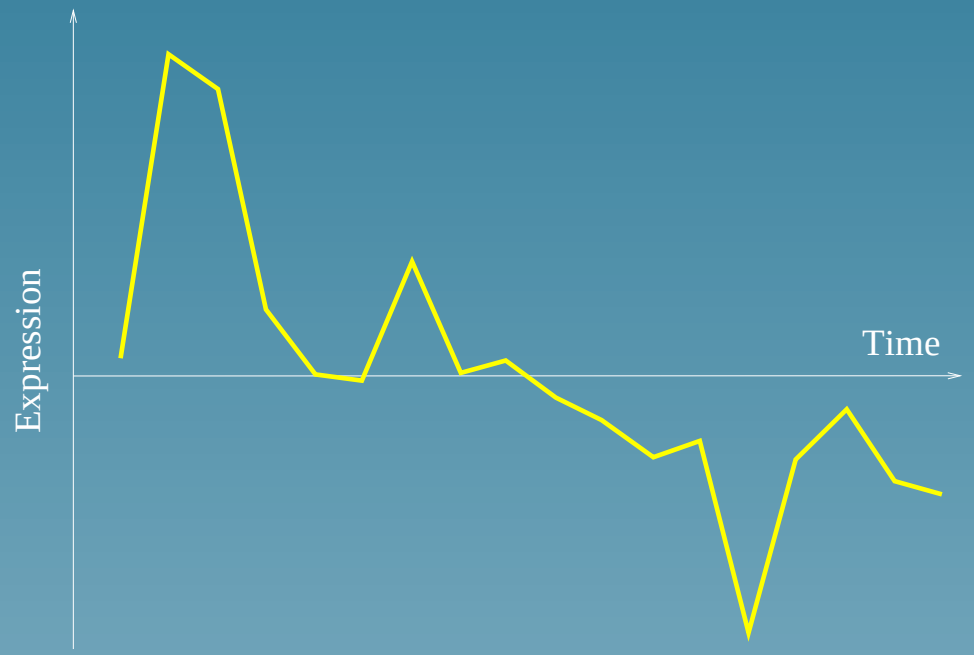
Exemple 2 : Extraction d'activité de voies



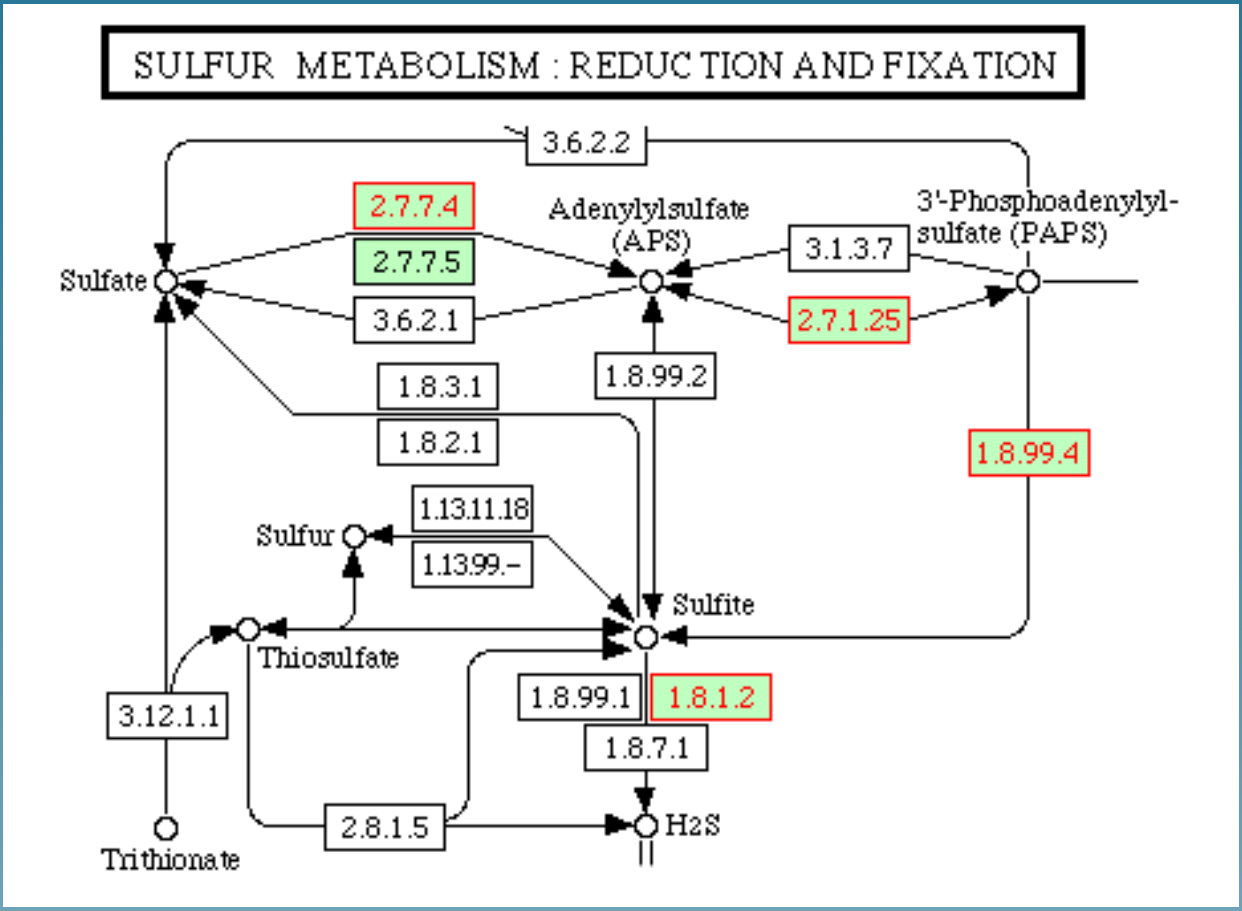
Trouver $w \in \mathbb{R}^p$ tel que $||f||_G$ soit petit, avec $f_i = w^\top x_i$.

Application ($ECCB'03$)

Comparaison du graphe des voies métaboliques et de données d'expression du cycle cellulaire de la levure



Exemple de gènes positivement corrélés



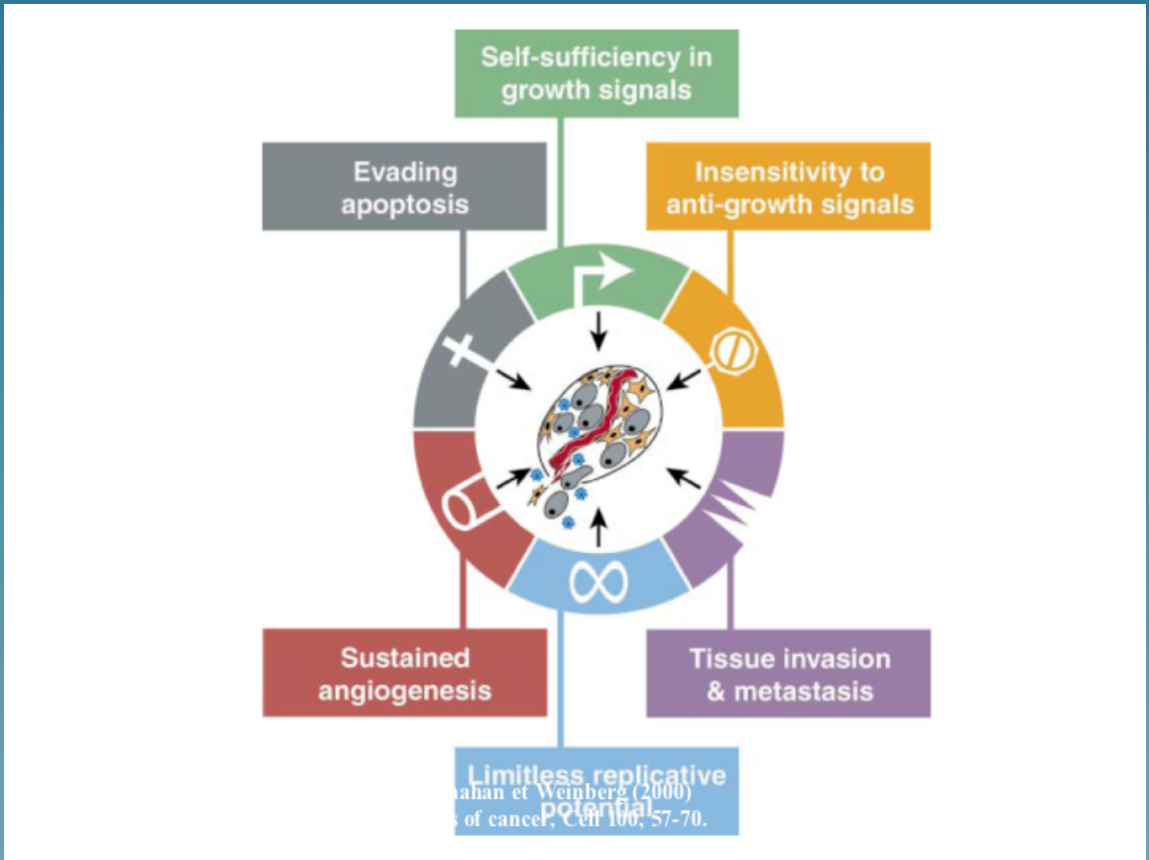
Autres applications

- Extraction de features pour la classification supervisée de gènes (*NIPS'02*)
- Extraction de features pour la classification non supervisée et la détection d'opérons dans les génomes bactériens (*ISMB'03*)
- Reconstruction de réseaux génétiques (*ISMB'04, NIPS'04, ISMB'05*).

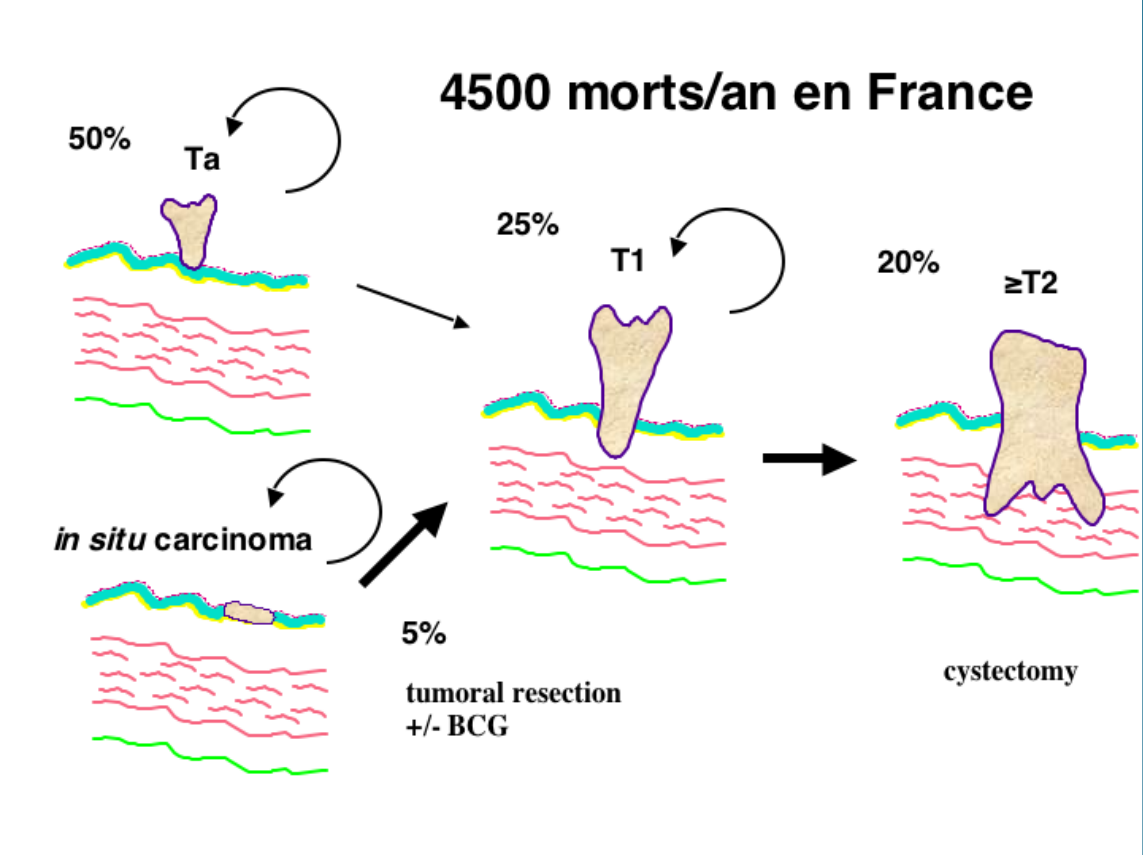
Partie 3

Le projet Kernelchip (2004-07): cancer et régulation

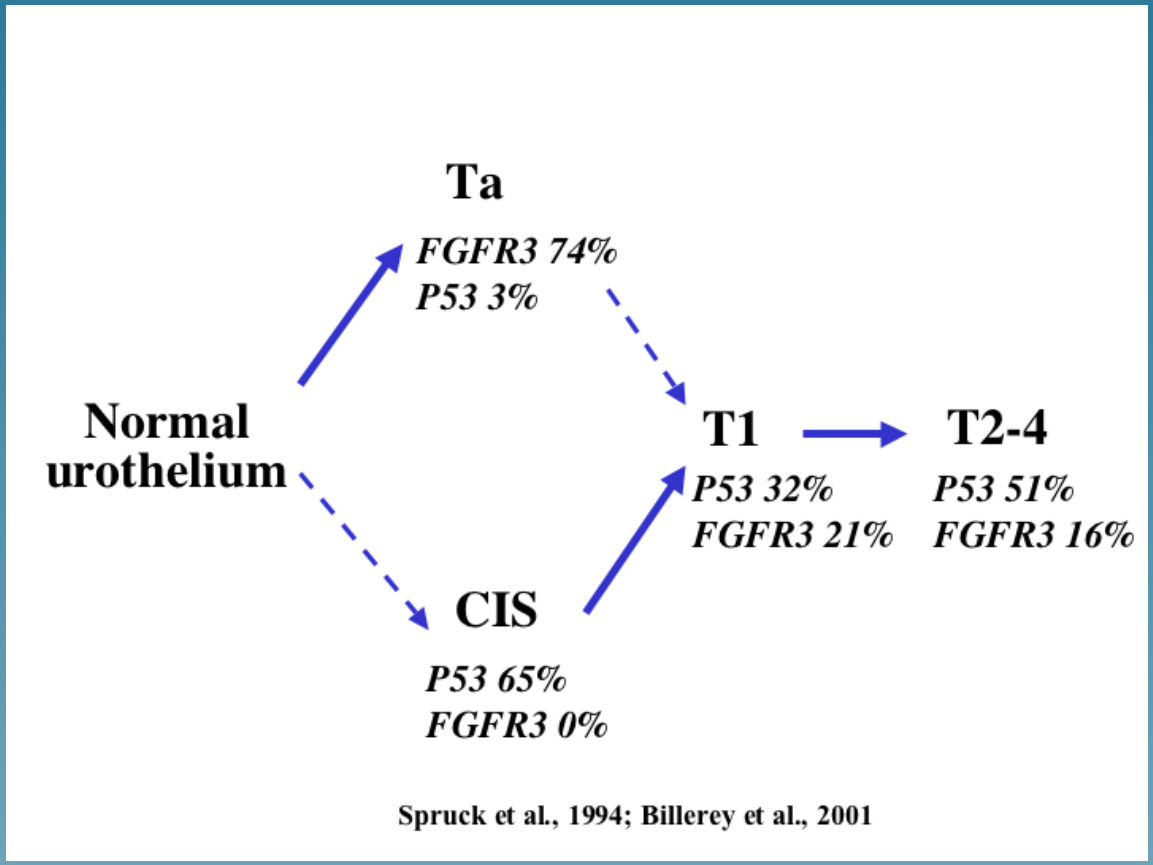
Caractéristiques d'une tumeur



Modèle 1: Cancer de la vessie (carcinomes)

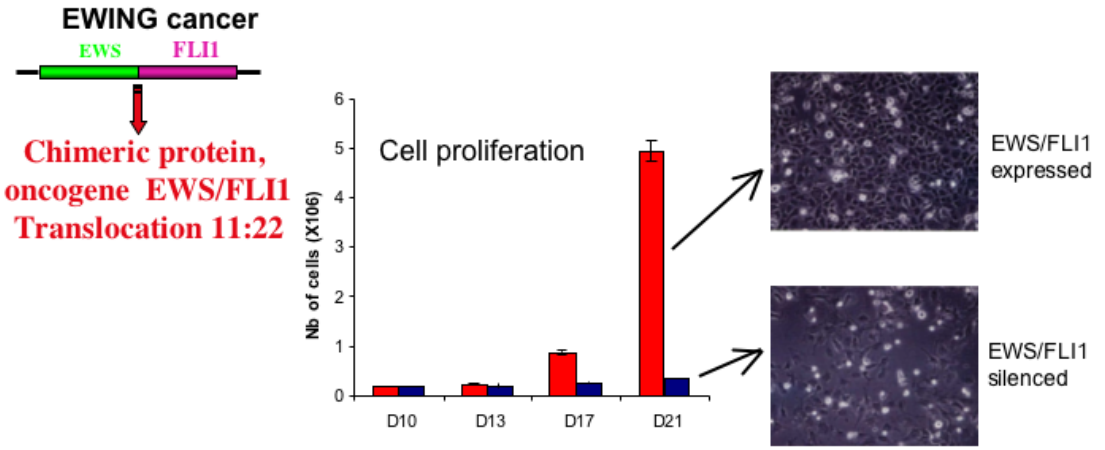


Modèle 1: Cancer de la vessie (cont.)



Modèle 2: Tumeur d'Ewing

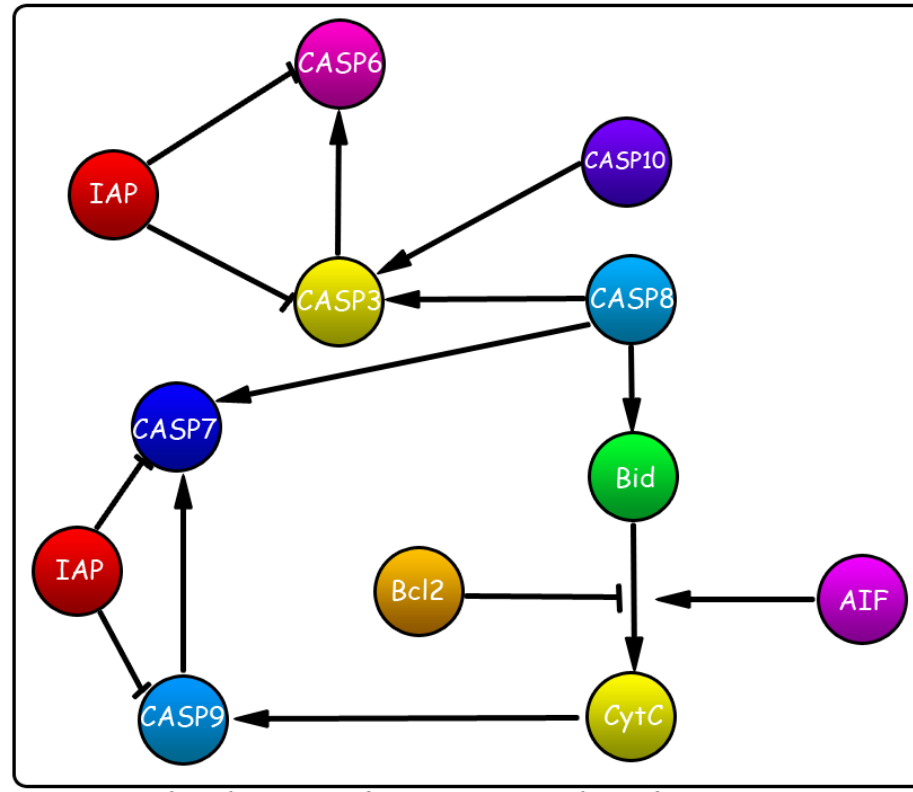
Cell culture with inducible expression of EWS/FLI1
(Olivier Delattre team, Institut Curie)



- **Temporal analysis** of the gene expression profile after inhibition then re-induction of EWS-FLI1
- **Transcriptome profile and genome profile** of 70 tumors

→ Identification and modeling of the gene regulation networks and signalling cascades involved

Exemple de réseau de régulation



Exemple de représentation de réseau : extrait de voies liées à l'apoptose (homo sapiens)

Données

- Transcriptome :
  - ⋆ Cancer de la vessie: 84 patients, puces Affymetrix HGU95AV2, 8,797 gènes par puce (F. Radvanyi)
  - ⋆ Tumeur d'Ewing: 70 patients et séries temporelles, puces Affy U133, 22,000 gènes (O. Delattre)
- Réseau de régulation : SHARP (97 gènes), KEGG (484 gènes)

Le projet

- Graphe dirigé, inhibition, relations indirectes
  $\longrightarrow$ fonctionnelle $||f||^2_G$ modifiée

- Validation de l'approche
  $\longrightarrow$ comparaison des différentes métriques
  $\longrightarrow$ prédiction de phénotype / type de tumeur
  $\longrightarrow$ détection de voies critiques
  $\longrightarrow$ validation biologique

Conclusion

Conclusion

- Richesse et quantité des données du transcriptome
- Nécessité d'intégrer et de croiser ces données avec d'autres sources d'information
- Les données de graphes (interaction, régulation...) nécessitent des développements mathématiques particuliers