Méthodes à noyaux en bioinformatique

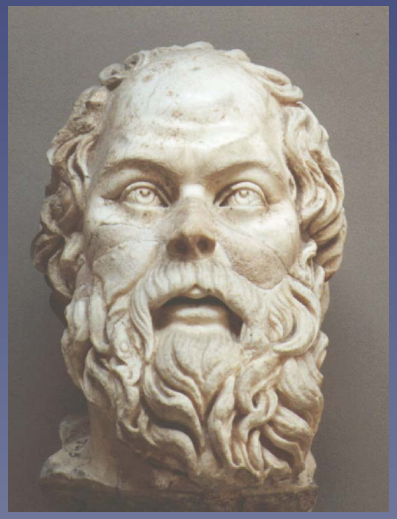
Jean-Philippe.Vert@mines.org

Ecole des Mines de Paris
Groupe bio-informatique

Habilitation à diriger les recherches de l'Université Paris 6
10 décembre 2004.

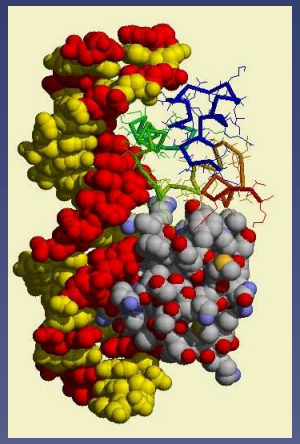
1953-2003: séquencage génome humain

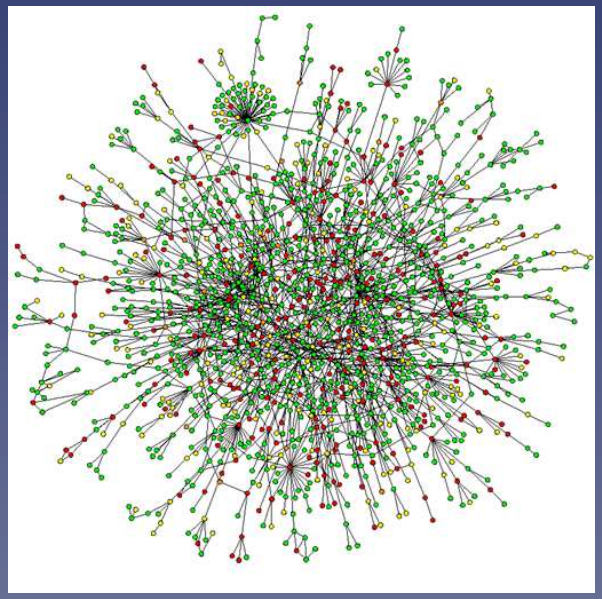
...AATCGATCGCGCATGCTAGCTACTAGCTAGTCGATGCATGAATGTTGCAAA
GTCGATGCATGAATGTTGCAGAAAACGCGCTGATATGTCATCGATGCAGATGAT
GCTAGCTAGCTATCAGCTGATCGATCGATCGATGCATGCAAATTATCGCGGAGT
AGTCATATGTGATACTACTACTCATGACACAAAACGCGGTAGCGCGTAGGCGCG
CAGTCGATCGGCGCGCGCGATAGGTATATATTATCATGATCAT...

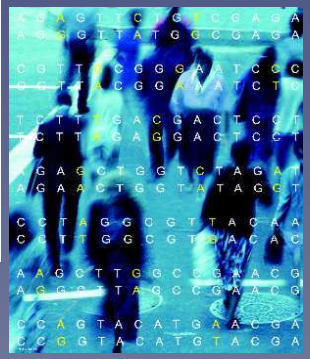


- 6 milliards de lettres
- De moins en moins de gènes (20-25,000 ?)
- Le début de la “post-génomique”

L'iceberg derrière le génome







Et bien d'autres...

Motivations

Imaginer un cadre théorique et des algorithmes pour

- représenter, intégrer les quantités de données
- modéliser, penser les systèmes vivants
- faire de l'inférence

Plan

1. Méthodes à noyaux pour la bio-informatiques
2. Noyaux pour données biologiques
3. Analyse et inférence de réseaux biologiques
4. Conclusion

Partie 1

Méthodes à noyaux et bio-informatique

Les données biologiques

Les données générées sont souvent:

- structurées et hétérogènes : sequences, structures 3D, graphes, réseaux, profiles d'expression, arbres phylogénétiques, SNP, ...
- en grandes quantitées (10^6 séquences de gènes)
- en grande dimension (1 puce à ADN mesure $10^5 \sim 10^6$ gènes)

Une réponse possible: méthodes à noyaux

Les méthodes à noyaux répondent (partiellement) à ces contraintes:

- Noyaux pour données structurées
- Opérations sur les noyaux pour intégrer des données hétérogènes
- Régularisation pour travailler en grande dimension avec peu de données
- Approches statistiques permettant d'extraire de l'information à partir de grandes bases de données

Rappel: noyau defini positif (n.d.p.)

Definition 1. *Un noyau défini positif (n.d.p.) sur l'ensemble $\mathcal{X}$ est une fonction $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symétrique:*

$$\forall (\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2, \quad K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x}),$$

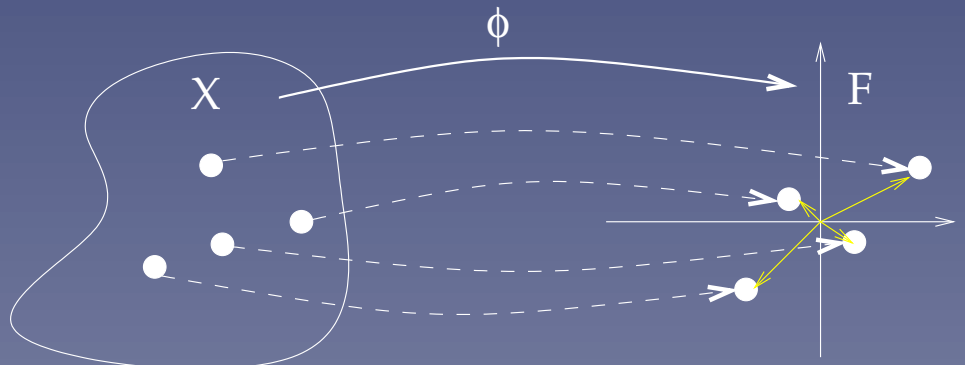
et qui satisfait, pour tout $N \in \mathbb{N}$, $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \in \mathcal{X}^N$ et $(a_1, a_2, \ldots, a_N) \in \mathbb{R}^N$:

$$\sum_{i=1}^{N} a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0.$$

Interprétation géométrique des n.d.p.

Theoreme 2. *(Aronszajn, 1950)* K *est un n.p.d. sur un espace* $\mathcal{X}$ *quelconque ssi il existe un espace de Hilbert* $\mathcal{H}$ *muni du produit scalaire* $\langle ., . \rangle_{\mathcal{H}}$ *et une application* $\Phi : \mathcal{X} \mapsto \mathcal{H}$*, telle que:*

$$\forall (\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2, \quad K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}}.$$



L'astuce noyau

- Tout algorithme pour vecteurs ne faisant intervenir que des produits scalaires peut être effectué implicitement en remplacant le produit scalaire par un n.d.p

L'astuce noyau

- Tout algorithme pour vecteurs ne faisant intervenir que des produits scalaires peut être effectué implicitement en remplacant le produit scalaire par un n.d.p
- Exemple: Support Vector Machines (classification, régression), clustering, PCA, ICA, CCA, régression logistique...= méthodes à noyau

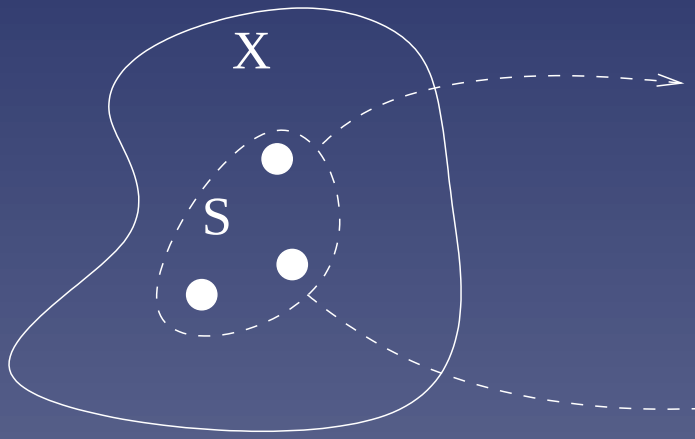
L'astuce noyau

- Tout algorithme pour vecteurs ne faisant intervenir que des produits scalaires peut être effectué implicitement en remplacant le produit scalaire par un n.d.p
- Exemple: Support Vector Machines (classification, régression), clustering, PCA, ICA, CCA, régression logistique...= méthodes à noyau
- Des noyaux "simples" peuvent correspondre à des Φ "complexes"

L'astuce noyau

- Tout algorithme pour vecteurs ne faisant intervenir que des produits scalaires peut être effectué implicitement en remplacant le produit scalaire par un n.d.p
- Exemple: Support Vector Machines (classification, régression), clustering, PCA, ICA, CCA, régression logistique...= méthodes à noyau
- Des noyaux "simples" peuvent correspondre à des Φ "complexes"
- Les objets ne sont pas nécessairement des vecteurs!

Représentation des données par noyaux



ϕ(S)=(aatcgagtcac,atggacgtct,tgcactact)

$$K = \begin{pmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 1 & 0.6 \\ 0.3 & 0.6 & 1 \end{pmatrix}$$

- Chaque ensemble de données est une matrice/fontion d.p.
- Les algorithmes traîtent des matrices semi-définie positives

Partie 2

Noyaux pour la bio-informatique

Qu'est-ce qu'un "bon" noyau $K(x, y)$?

- Contraintes mathématiques: symétrique, défini positif

Qu'est-ce qu'un "bon" noyau $K(x, y)$?

- Contraintes mathématiques: symétrique, défini positif
- Le problème de classification doit être "plus facile" dans l'espace image

$$\hat{f} = \arg\min_{f \in H_K} \left\{ R_{emp}(f) + \lambda ||f||^2_{H_K} \right\}$$

Qu'est-ce qu'un "bon" noyau $K(x, y)$?

- Contraintes mathématiques: symétrique, défini positif
- Le problème de classification doit être "plus facile" dans l'espace image

$$\hat{f} = \arg\min_{f \in H_K} \left\{ R_{emp}(f) + \lambda ||f||^2_{H_K} \right\}$$

- Le noyau doit être rapide à calculer

Quelques exemples de noyaux

- Noyau interpolé pour séquences de longueur fixe ($PSB'02$)

Quelques exemples de noyaux

- Noyau interpolé pour séquences de longueur fixe (*PSB'02*)
- Noyau pour profiles phylogénétiques (*ISMB'02*)

Quelques exemples de noyaux

- Noyau interpolé pour séquences de longueur fixe (*PSB'02*)
- Noyau pour profiles phylogénétiques (*ISMB'02*)
- Noyau pour structures 2D de molécules (*ICML'04*)

Quelques exemples de noyaux

- Noyau interpolé pour séquences de longueur fixe ($PSB'02$)
- Noyau pour profiles phylogénétiques ($ISMB'02$)
- Noyau pour structures 2D de molécules ($ICML'04$)
- Noyau d'information mutuel pour séquences ($IJCNN'04$)

Quelques exemples de noyaux

- Noyau interpolé pour séquences de longueur fixe ($PSB'02$)
- Noyau pour profiles phylogénétiques ($ISMB'02$)
- Noyau pour structures 2D de molécules ($ICML'04$)
- Noyau d'information mutuel pour séquences ($IJCNN'04$)
- Noyau d'alignement local pour séquences ($Bioinformatics\ 04$)

Quelques exemples de noyaux

- Noyau interpolé pour séquences de longueur fixe (*PSB'02*)
- Noyau pour profiles phylogénétiques (*ISMB'02*)
- Noyau pour structures 2D de molécules (*ICML'04*)
- Noyau d'information mutuel pour séquences (*IJCNN'04*)
- Noyau d'alignement local pour séquences (*Bioinformatics 04*)
- Noyau pour ensembles de points (*NIPS'04*)

Applications

En combinaison avec des SVM, ces noyaux ont été appliqués à:

- détection de peptides signaux dans des séquences biologiques

Applications

En combinaison avec des SVM, ces noyaux ont été appliqués à:

- détection de peptides signaux dans des séquences biologiques
- prédiction de fonctions de gènes

Applications

En combinaison avec des SVM, ces noyaux ont été appliqués à:

- détection de peptides signaux dans des séquences biologiques
- prédiction de fonctions de gènes
- criblage virtuels de molécules en recherche pharmaceutique

Applications

En combinaison avec des SVM, ces noyaux ont été appliqués à:

- détection de peptides signaux dans des séquences biologiques
- prédiction de fonctions de gènes
- criblage virtuels de molécules en recherche pharmaceutique
- détection d'homologie entre séquences biologiques

Applications

En combinaison avec des SVM, ces noyaux ont été appliqués à:

- détection de peptides signaux dans des séquences biologiques
- prédiction de fonctions de gènes
- criblage virtuels de molécules en recherche pharmaceutique
- détection d'homologie entre séquences biologiques
- classification d'images

Exemple 1: profiles phylogénétiques ($ISMB\ 02$)

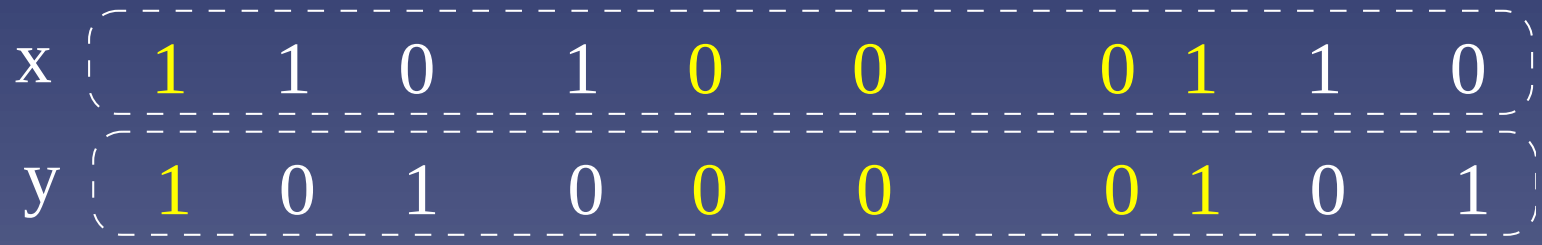
Un vecteur de bits $0/1$ indiquant la présence ou l'absence du gène dans un ensemble de génomes séquencés

Gene	human	yeast	. . .	HIV	E. coli
YAL001C	1	1	. . .	0	0
YAB002W	0	0	. . .	0	1
⋮	⋮	⋮	⋮	⋮	⋮

- Peut être calculé $in\ silico$
- Utile pour prédire la fonction des gènes

Approche naive

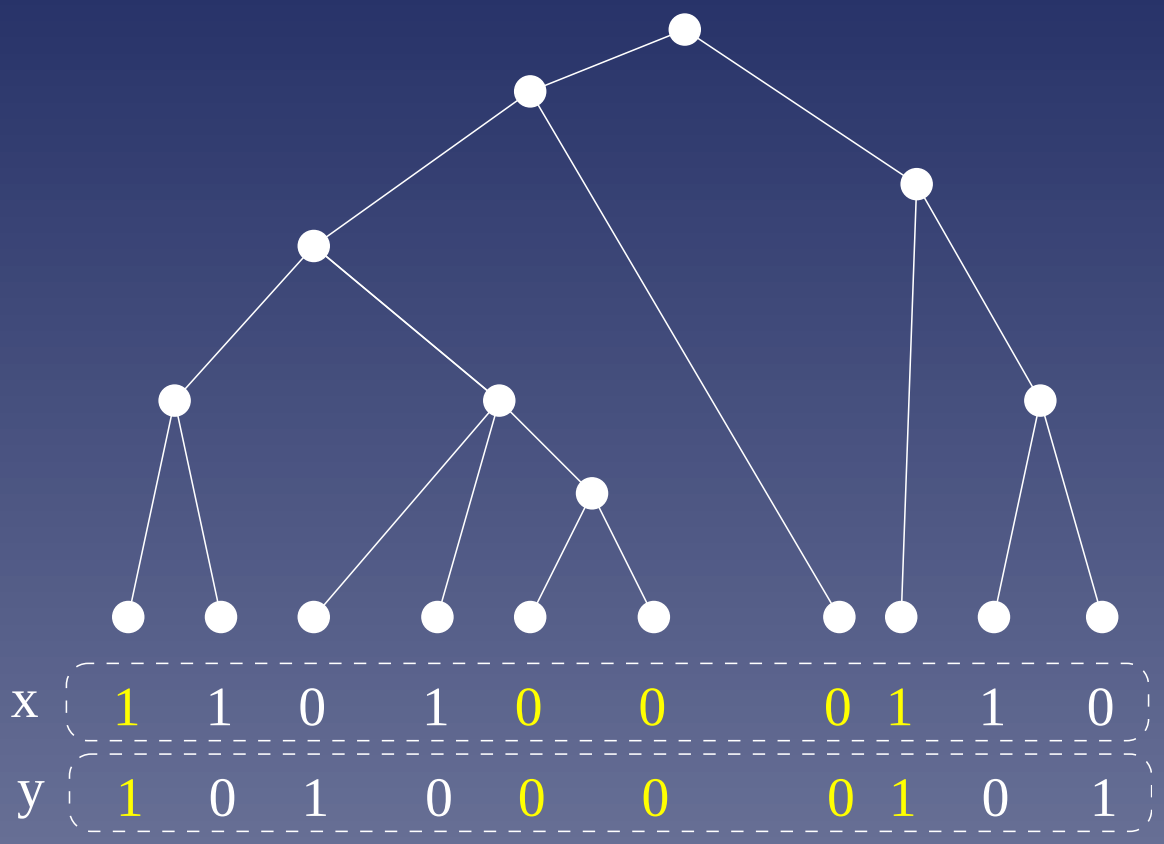
- Compter le nombre de bits communs:



$$s(x, y) = 5$$

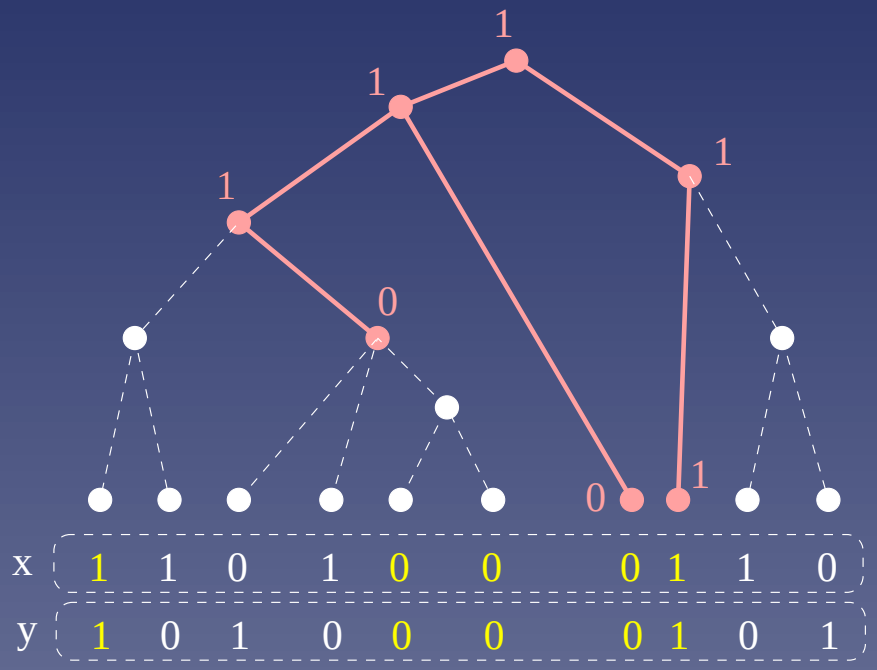
- Clustering ou k-NN pour prédire la fonction des gènes à partir de cette mesure de similarité (Pellegrini et al., 1999)

Ce qui n'est pas utilisé dans l'approche naive



La connaissance de l'arbre phylogénétique

Noyau “phylogénétique”



$$K(x,y) = \sum_e P(e)P(x|e)P(y|e)$$

Implémentation

- Il faut effectuer une somme sur un nombre exponentiel de termes
- Deux astuces peuvent être combinées: factorisation des probabilités le long des branches de l'arbres (message-passing algorithm), et somme sur les sous-arbres (Context-Tree Weighting)
- $K(x, y)$ est calculable avec 1 traversée post-order de l'arbre
- La complexité est donc linéaire avec la longueur des profiles.

Classification de gènes par fonction (ROC 50)

Functional class	Naive kernel	Tree kernel	Difference
Amino-acid transporters	0.74	0.81	**+ 9%**
Fermentation	0.68	0.73	**+ 7%**
ABC transporters	0.64	0.87	**+ 36%**
C-compound transport	0.59	0.68	**+ 15%**
Amino-acid biosynthesis	0.37	0.46	**+ 24%**
Amino-acid metabolism	0.35	0.32	*- 9%*
Tricarboxylic-acid pathway	0.33	0.48	**+ 45%**
Transport Facilitation	0.33	0.28	*- 15%*

Extensions

- $X_1, \ldots, X_n$ v.a. discrètes
- $I_1, \ldots, I_v \subset \{1, \ldots, n\}$ une famille de sous-ensembles
- Noyau interpolé:

$$K(x, y) = \frac{1}{v} \sum_{i=1}^{v} p(x_{I_i}) \delta(x_{I_i}, y_{I_i}) \times p(x_{I_i^c} | x_{I_i}) p(y_{I_i^c} | y_{I_i})$$

Propriété 1

Ce noyau interpole entre le noyau diagonal:

$$K_{diag}(x, y) = p(x)\delta(x, y)$$

et le noyau produit:

$$K_{prod}(x, y) = p(x)p(y).$$

Propriété 2

Deux éléments x et y sont similaires si ils ont des parties rares communes:

$$K(x,y) = K_{prod}(x,y) \times \frac{1}{v} \sum_{i=1}^{v} \frac{\delta(x_{I_i}, y_{I_i})}{p(x_{I_i})}$$

Implementations en temps linéaire

- v.a. i.i.d., tous les sous-ensembles possibles ($PSB\ 02$):

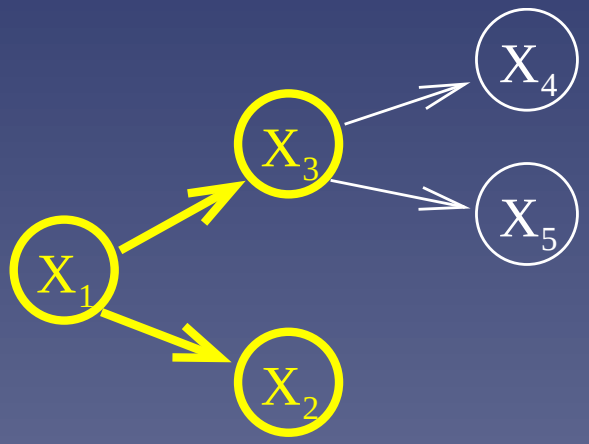
Implementations en temps linéaire

- v.a. i.i.d., tous les sous-ensembles possibles ($PSB\ 02$):

- Modèle de Markov, blocs contigus

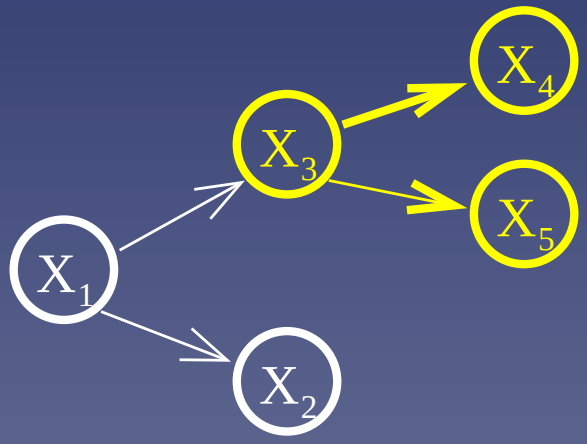
Implementations en temps linéaire

- Modèle graphique d'arbre, sous-arbres contenants la racine



Implementations en temps linéaire

- Modèle graphique d'arbre, sous-arbres quelconque



Exemple 2: noyau d'alignement local (*Bioinformatics 04*)

- Collaboration avec H. Saigo et T. Akutsu
- But = développer un noyau pour séquences biologiques (protéines, AND)
- Il existe déjà des mesures de similarité pertinentes; comment en faire des noyaux?

Alignement local

- Pour deux séquences x et y, un alignement local π avec gaps est:

```
ABCD EF---G-HI JKL
     | |   |  |
  MNO EFPORGS-I TUVWX
```

- Le score est:

$$s(x, y, \pi) = s(E, E) + s(F, F) + s(G, G) + s(I, I) - s(gaps)$$

Score de Smith-Waterman (SW)

$$SW(x,y) = \max_{\pi \in \Pi(x,y)} s(x,y,\pi)$$

- Calculé par programmation dynamique en $O(|x| \cdot |y|)$
- Pas défini positif en général

Noyau d'alignement local (LA kernel)

$$K_{LA}^{(\beta)}(x,y) = \sum_{\pi \in \Pi(x,y)} \exp\left(\beta s(x,y,\pi)\right),$$

Theoreme 3. *Si la matrice de substitution est conditionnellement définie positive, alors* $K_{LA}^{(\beta)}$ *est un noyau défini positif sur l'espace des séquences de longueur variable. Il vérifie:*

$$\lim_{\beta \to +\infty} \frac{1}{\beta} \ln K_{LA}^{(\beta)}(x,y) = SW(x,y).$$

Preuve

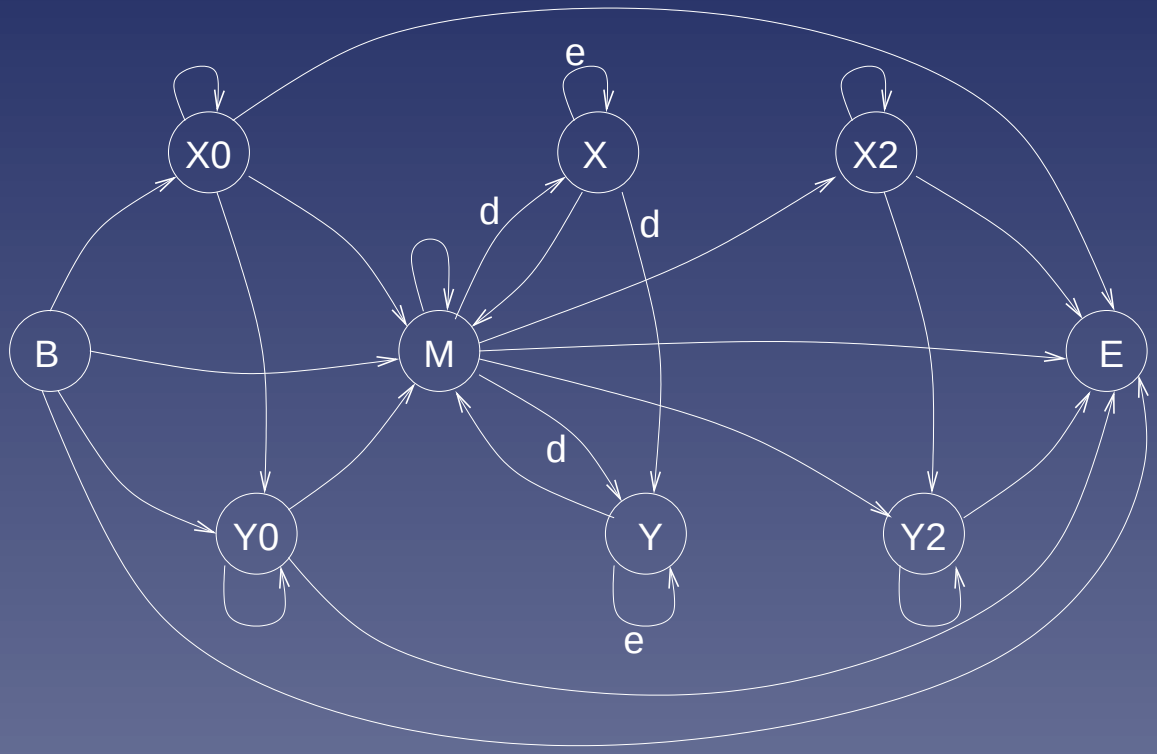
$K_{LA}^{(\beta)}$ s'écrit comme une convolution de noyaux plus simples (Haussler, 1999):

$$K_{LA}^{(\beta)} = \sum_{n=0}^{\infty} K_0 \star \left(K_a^{(\beta)} \star K_g^{(\beta)} \right)^{(n-1)} \star K_a^{(\beta)} \star K_0.$$

avec

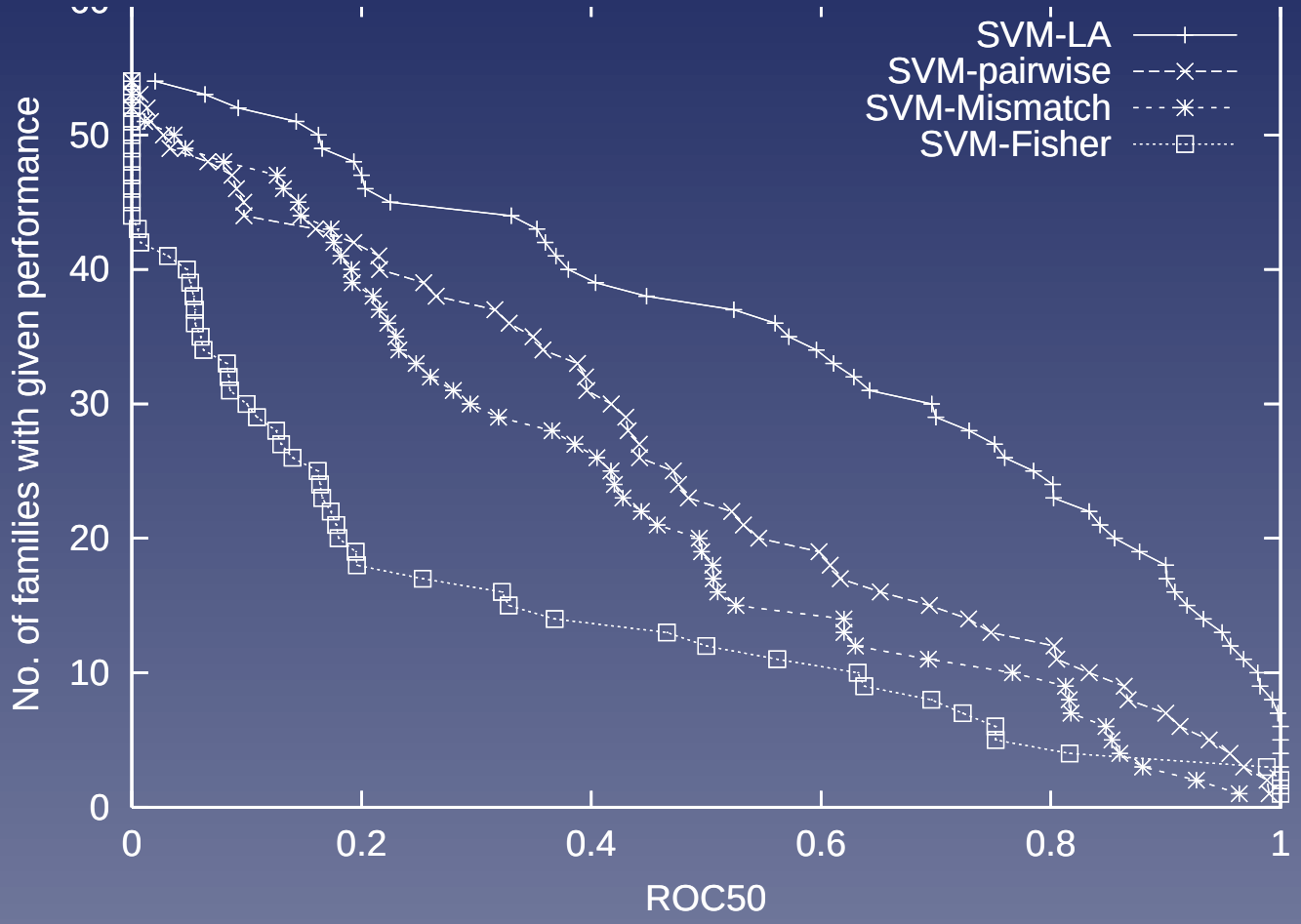
$$K_1 \star K_2(x,y) = \sum_{x_1 x_2 = x, y_1 y_2 = y} K_1(x_1,y_1) K_2(x_2,y_2)$$

Calcul du noyaux



Programmation dynamique, $O(|x| \cdot |y|)$

Résultats du benchmark SCOP



Exemple 3: noyau d'information mutuel ($IJCNN'04$)

- Travail de thèse de Marco Cuturi 
- Motivations:
  - ⋆ Faire un noyau pour séquences de longueur variables performant et rapide (temps linéaire)
  - ⋆ Etudier les liens entre noyaux et théorie de l'information / compression / codage de séquences

De la compression à la comparaison

- L'algorithme context tree weigthing (CTW) calcule un probabilité de mélange d'une séquence en temps linéaire :

$$P_w(x) = \sum_{m \in \mathcal{M}} \int_{\theta_m} P_{m,\theta_m}(x)$$

- Noyau d'information mutuelle entre séquences:

$$K(x,y) = \sum_{m \in \mathcal{M}} \int_{\theta_m} P_{m,\theta_m}(x) P_{m,\theta_m}(y)$$
$$= \exp\left[l(xy)\right]$$

Application

- Implémentation linéaire d'un triple mélange
  - ⋆ modèles de Markov de longueur variable
  - ⋆ mélange de Dirichlet
- Résultats encourageants pour la détection d'homologie lointaine
- Approche pertinente pour d'autres algorithmes de compression?

Example 4: noyaux pour ensembles et mesures (NIPS'04)

- Travail de thèse de Marco Cuturi 
- Motivations:
  - ⋆ généraliser le travail précédent
  - ⋆ Approfondir le lien entre structure algébrique (semi-groupe) et noyaux

Noyaux pour mesures

On étudie les noyaux sur $\mathcal{X} = \mathcal{M}_+^b(\mathcal{U})$ ($\mathcal{U}$ Hausdorff) de la forme

$$K_f(x, x') = f(x + x').$$

Theoreme 4. *Pour f continue sur $\mathcal{X}$ muni de la topologie faible, K_f est un noyau défini positif si et seulement si f s'écrit:*

$$f(x) = \int_{\mathcal{C}(\mathbb{R}^{\mathcal{U}})} e^{x[h]} d\nu(h)$$

où ν est une mesure de Radon positive sur $\mathcal{C}(\mathbb{R}^{\mathcal{U}})$,

Exemples

- Si $\mathcal{X}$ est un ensemble de densité d'entropie finie, alors

$$K(x, x') = -h\left(\frac{x + x'}{2}\right) = \int \left(\frac{x + x'}{2}\right) \log\left(\frac{x + x'}{2}\right)$$

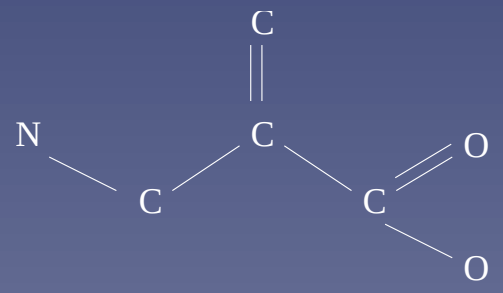
est conditionnellement d.p. ($\exp(\beta K)$ est d.p. $\forall \beta$).

- Si $\mathcal{X}$ est l'ensemble des ensembles finis de points, alors le noyau suivant est d.p. (kernelisation possible)

$$K(x, x') = \frac{1}{|\Sigma_{x+x'}|},$$

Exemple 4: Noyaux pour molécules *(ICML'04)*

- Travail de thèse de Pierrer Mahé
- Développer et améliorer des noyaux pour petites molécules



- Applications: criblage virtuel, prédiction de pharmacocinétique...

Noyaux pour molécules

- (Kashima et al. 2003) Soient deux graphes G_1 et G_2, H_1 et H_2 des chemins aléatoires indépendants sur G_1 et G_2. Un noyau valide est:

$$K(G_1, G_2) = P(label(H_1) = label(H_2))$$

- Calcul par inversion d'une matrice $|G| \times |G|$ avec $G = G_1 \times G_2$

Deux améliorations

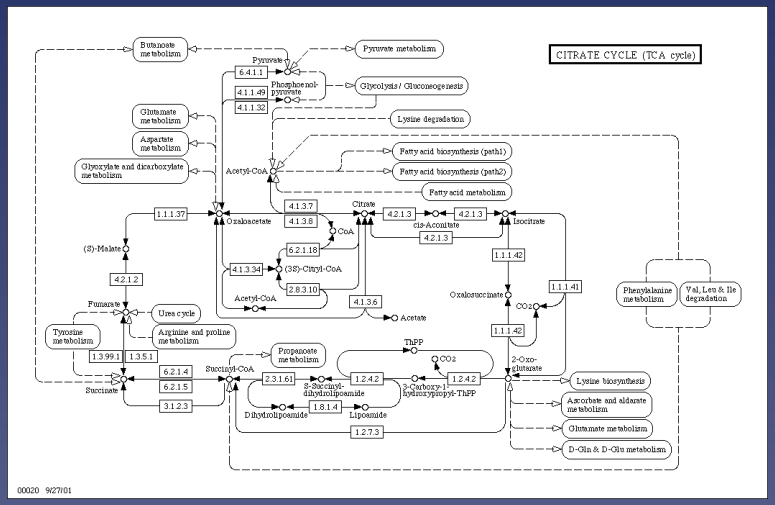
- Augmenter la spécificité des labels
  - ⋆ Diminue $|G_1 \times G_2|$, vitesse $\times 10^2$.
  - ⋆ augmente la spécificité du feature space, performance augmente
- Filtrer les chemins revenant sur leurs pas: il existe une transformation f telle que:

$$K'(G_1, G_2) = K\left(f(G_1), f(G_2)\right)$$

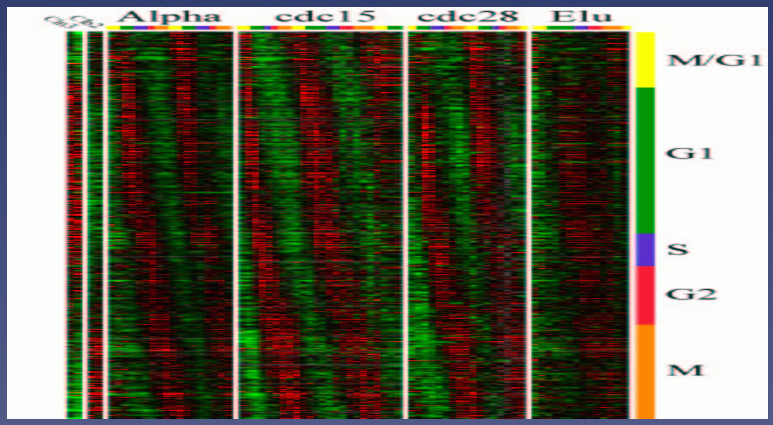
Partie 3

Analyse et inférence de graphes

Comparaisons de données hétérogènes (NIPS'02)



VS



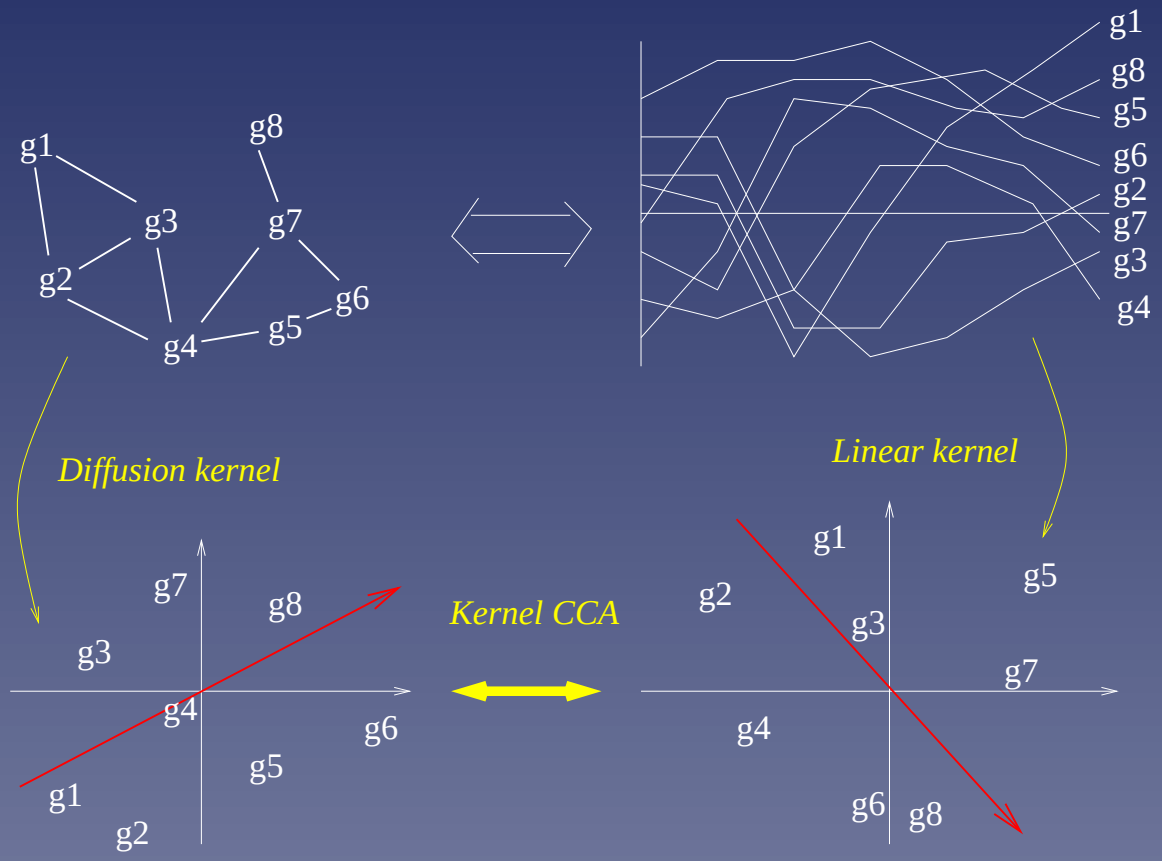
Détection de chemins "actifs"? Débruitage des données?
Inférence de nouvelles arêtes?
Existe-t-il des "corrélations"?

Rappel : ACC

- L'analyse de corrélations canoniques (ACC) permets de détecter des corrélations entre deux vecteurs aléatoires
- Les directions de corrélations sont définies par:

$$\max_{w_1,w_2} \frac{w_1^\top X^\top Y w_2}{\left(w_1^\top X^\top X w_1 + \lambda_1 w_1^\top w_1\right)^{\frac{1}{2}} \left(w_2^\top Y^\top Y w_2 + \lambda_2 w_2^\top w_2\right)^{\frac{1}{2}}}$$

Détection de corrélations par kernel-ACC



Interprétation

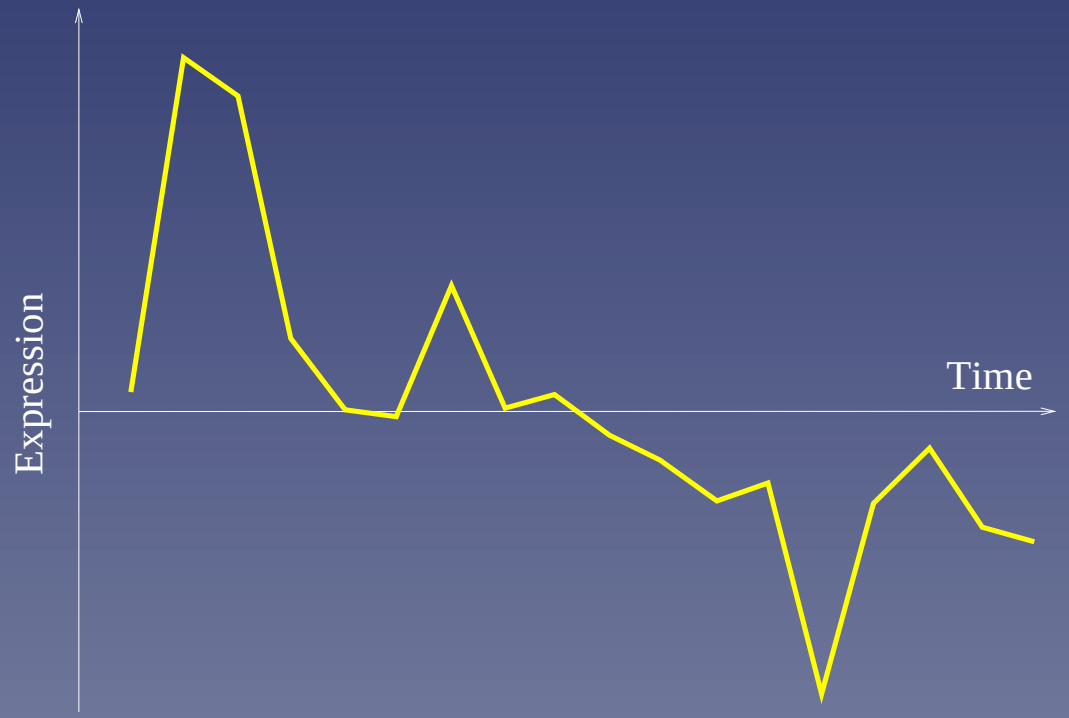
L'ACC à noyau résoud:

$$\max_{(f_1,f_2)\in\mathcal{H}_1\times\mathcal{H}_2} \hat{corr}(f_1,f_2) \times \left(1+\frac{\| f_1 \|^2_{\mathcal{H}_1}}{\| f_1 \|^2_2}\right)^{-\frac{1}{2}} \times \left(1+\frac{\| f_2 \|^2_{\mathcal{H}_2}}{\| f_2 \|^2_2}\right)^{-\frac{1}{2}}$$

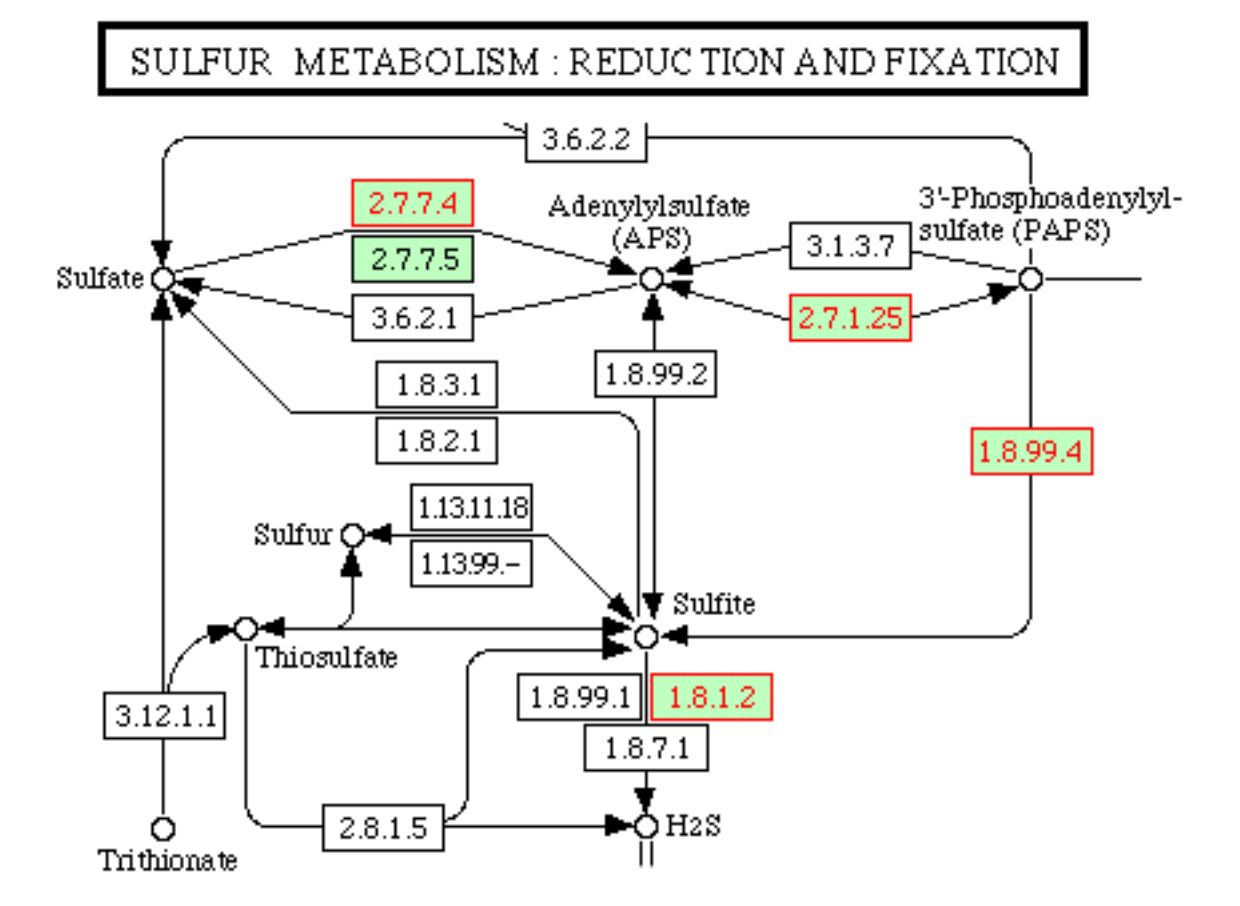
- f_1 est smooth sur le graphe
- f_2 capture de la variance dans les données

Applications ($ECCB'03$)

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



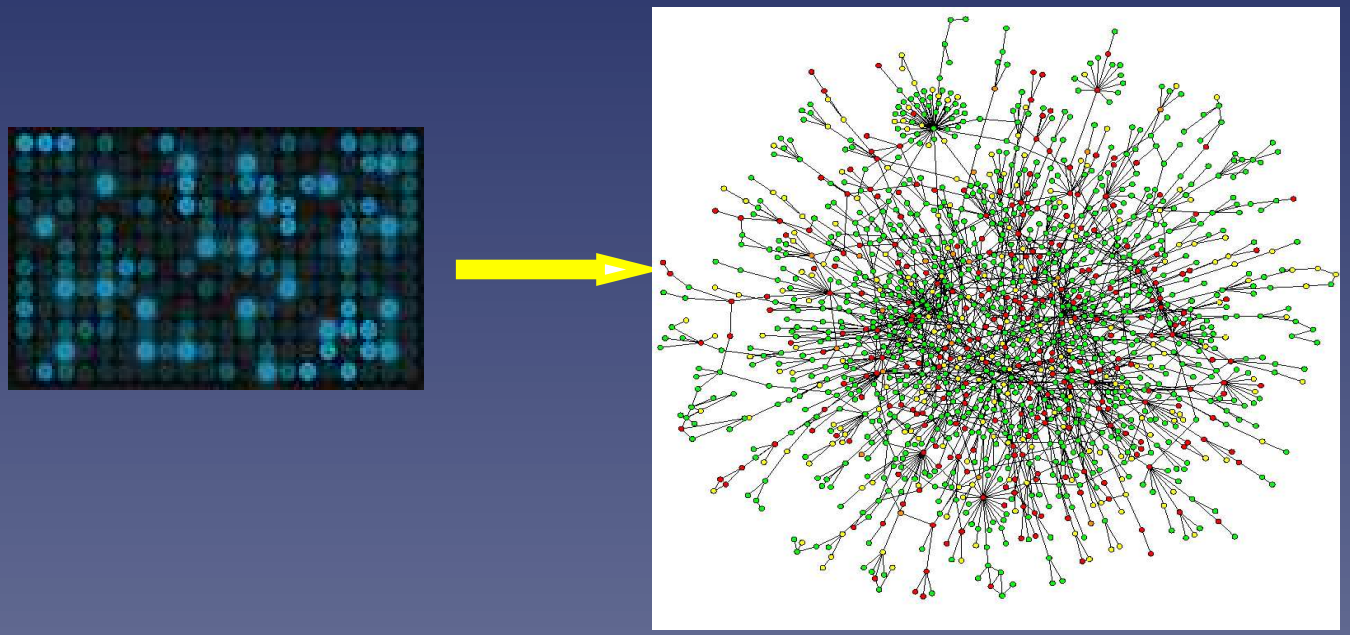
Exemple de gènes positivement corrélés



Extensions

- 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 ($ISMB'03$)

Inférence supervisée de graphe



Réseaux Bayésiens (Friedman et al., 2001), systèmes dynamiques (Akutsu, 2000), graphes de similarité (Marcotte et al., 1999)...

Deux idées

Collaboration avec Yoshihiro Yamanishi 

- L'apprentissage supervisé de graphe est plus performant et souvent plus adapté au problème

Deux idées

Collaboration avec Yoshihiro Yamanishi 

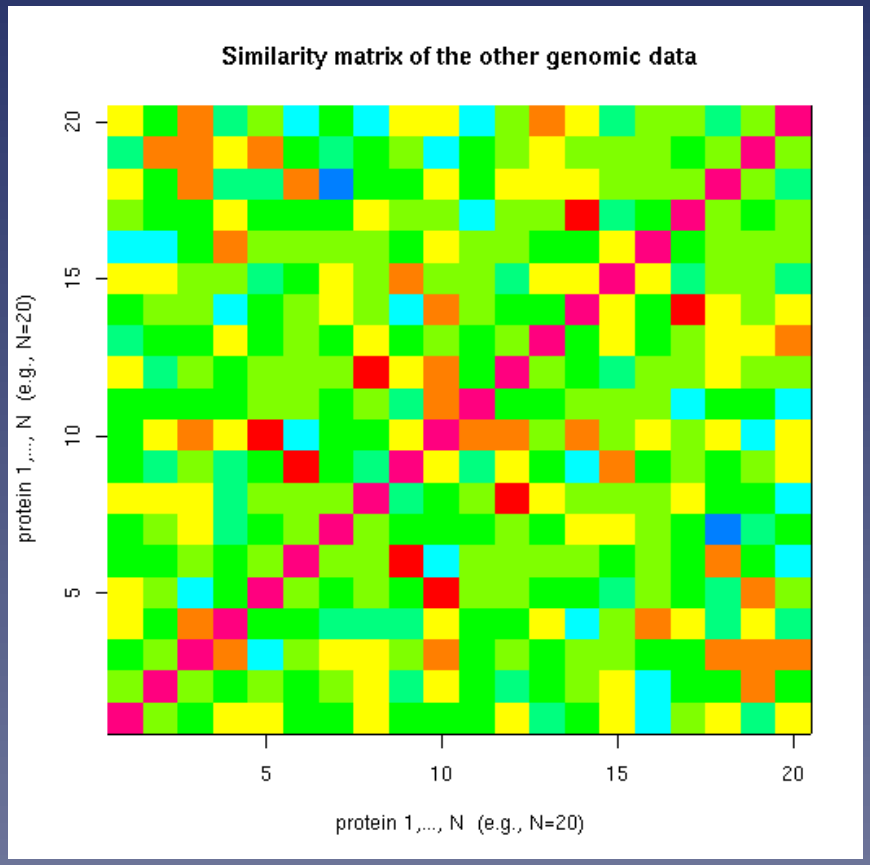
- L'apprentissage supervisé de graphe est plus performant et souvent plus adapté au problème
- L'apprentissage supervisé de graphe peut être formulé en termes d'apprentissage de métrique

Deux idées

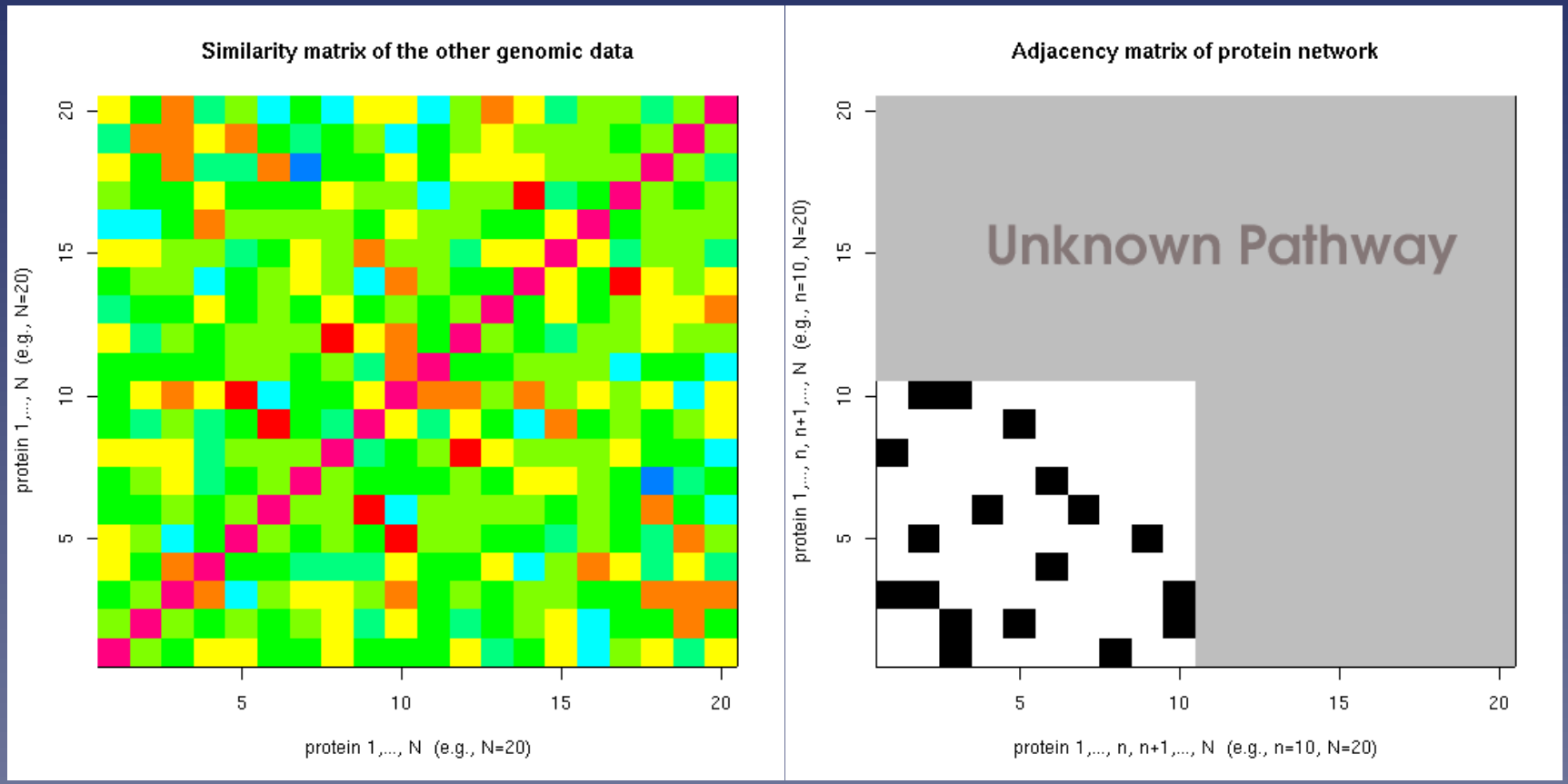
Collaboration avec Yoshihiro Yamanishi 

- L'apprentissage supervisé de graphe est plus performant et souvent plus adapté au problème
- L'apprentissage supervisé de graphe peut être formulé en termes d'apprentissage de métrique
- Les méthodes à noyaux permettent d'intégrer facilement des données hétérogènes

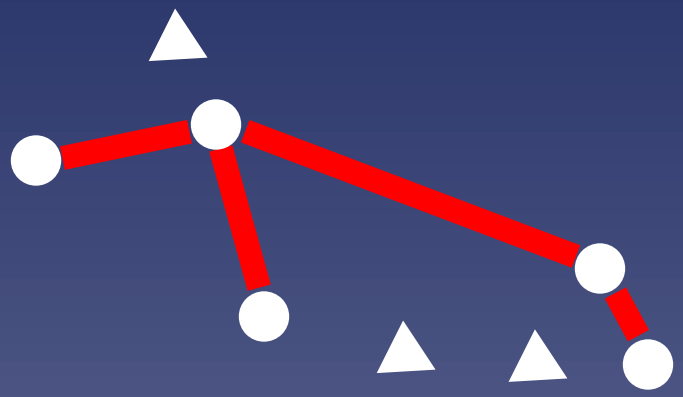
Inférence supervisée de graphe



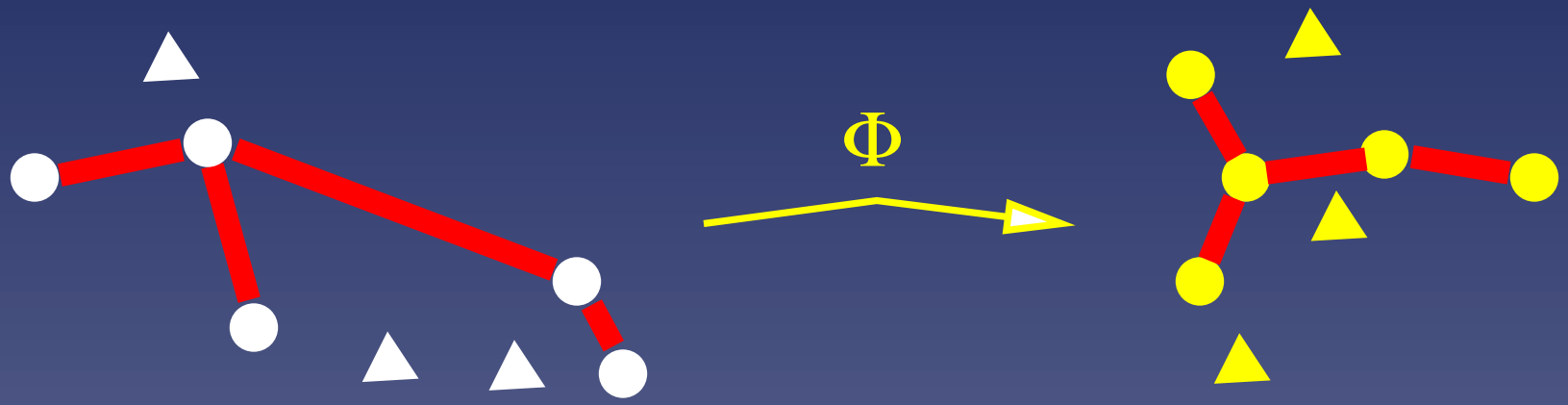
Inférence supervisée de graphe



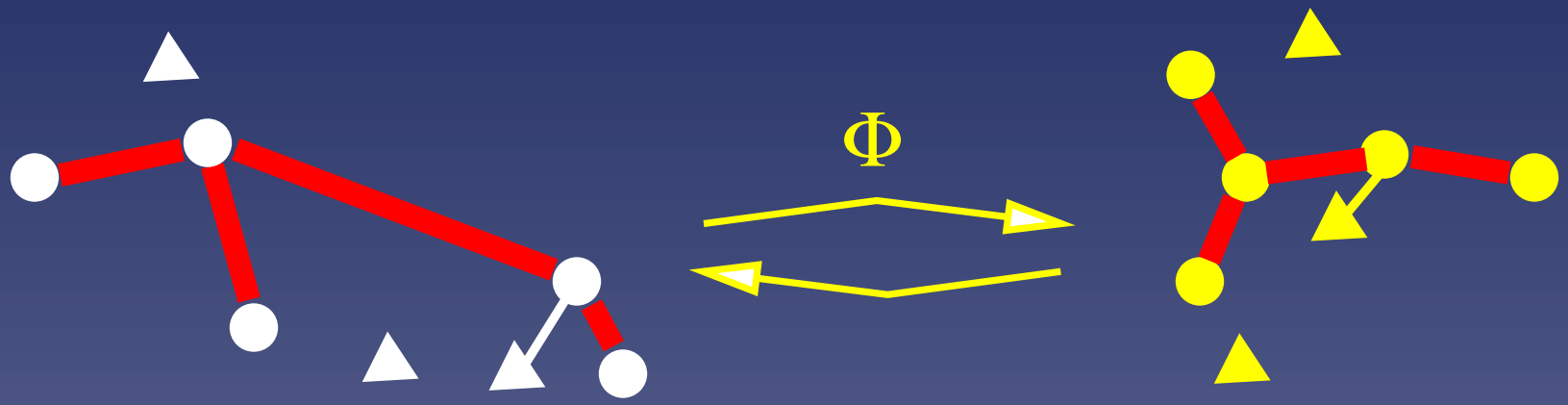
Inférence de graphe par apprentissage de métrique



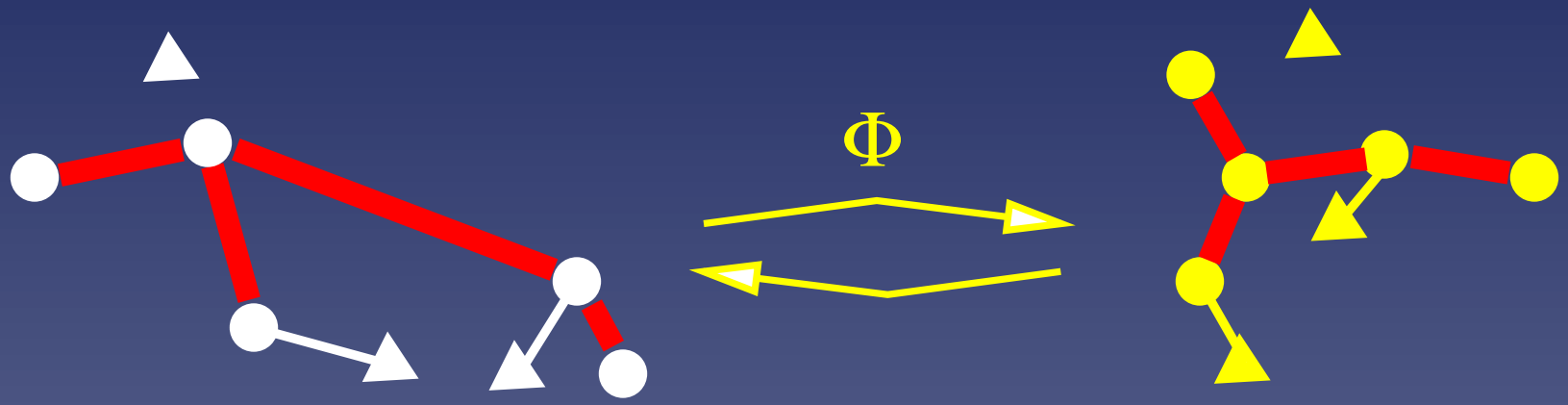
Inférence de graphe par apprentissage de métrique



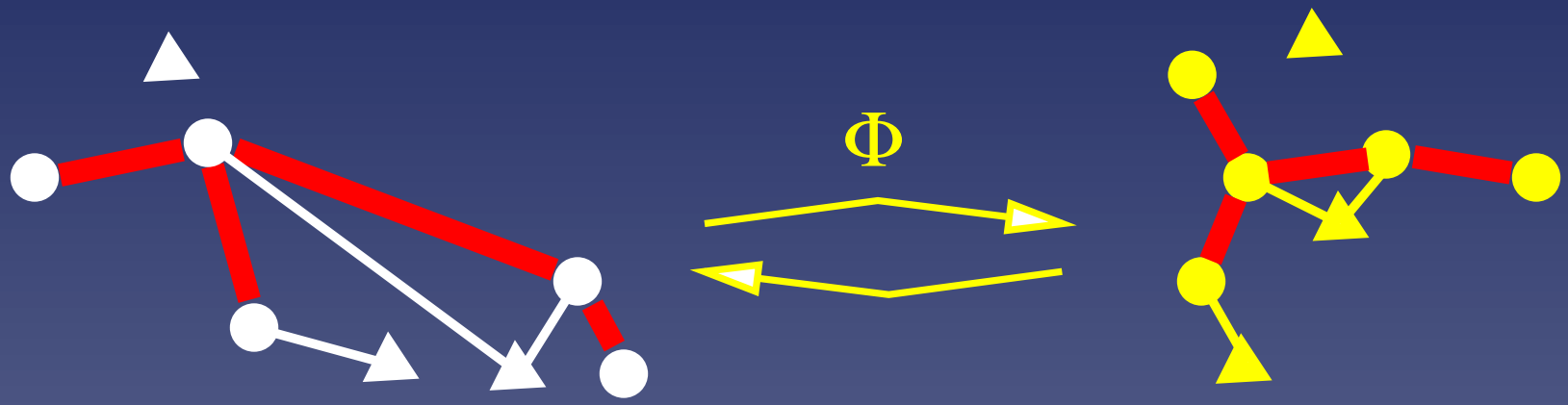
Inférence de graphe par apprentissage de métrique



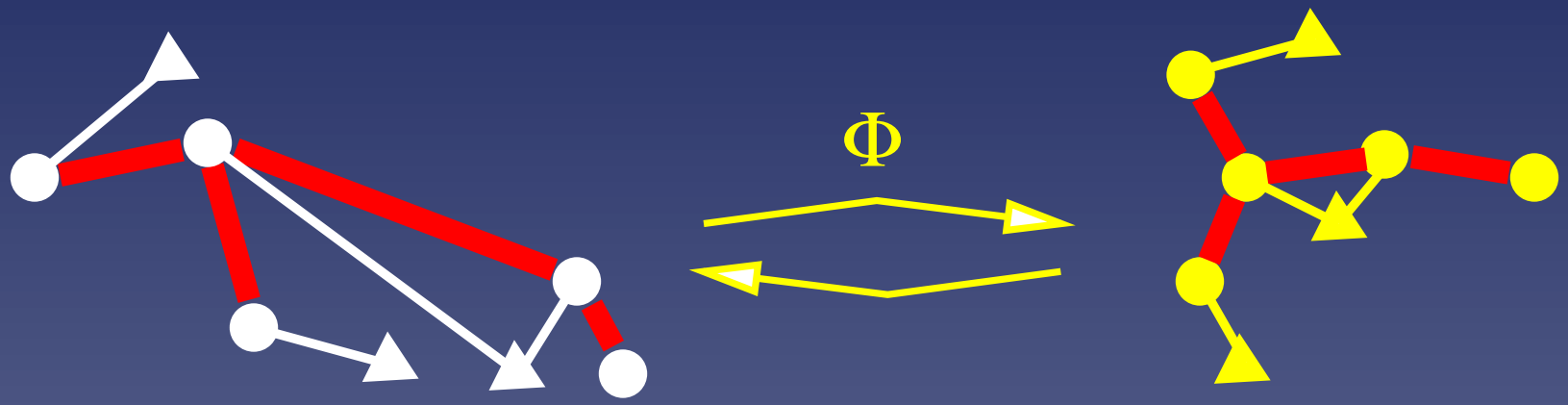
Inférence de graphe par apprentissage de métrique



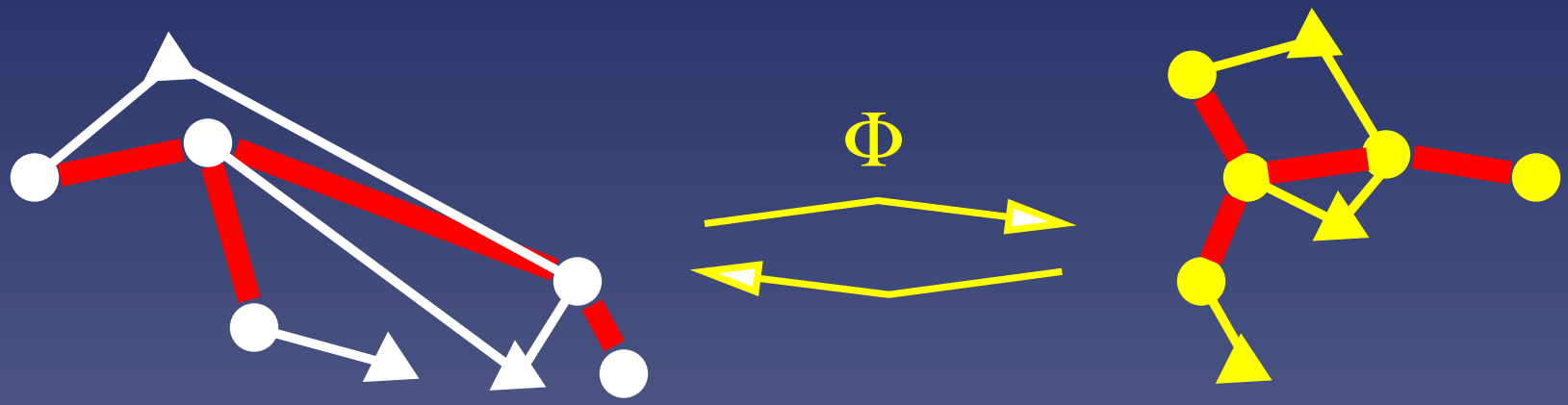
Inférence de graphe par apprentissage de métrique



Inférence de graphe par apprentissage de métrique



Inférence de graphe par apprentissage de métrique



Une méthode avec noyaux (*NIPS 04*)

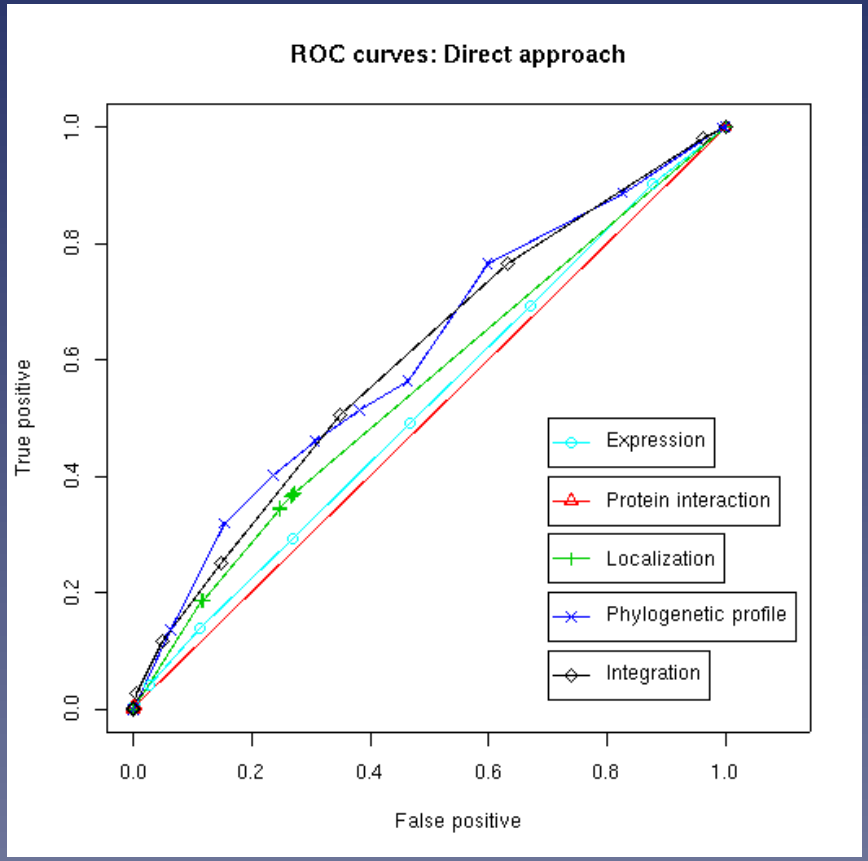
- Apprend une mesure de similarité qui "colle" avec la partie du graphe connue: $\Phi = (f_1, \ldots, f_d)$ avec

$$f_i = \underset{f \perp \{f_1,\ldots,f_{i-1}\}, var(f)=1}{\arg\min} \left\{ \sum_{i \sim j} (f(x_i) - f(x_j))^2 + \lambda ||f||_k^2 \right\}.$$

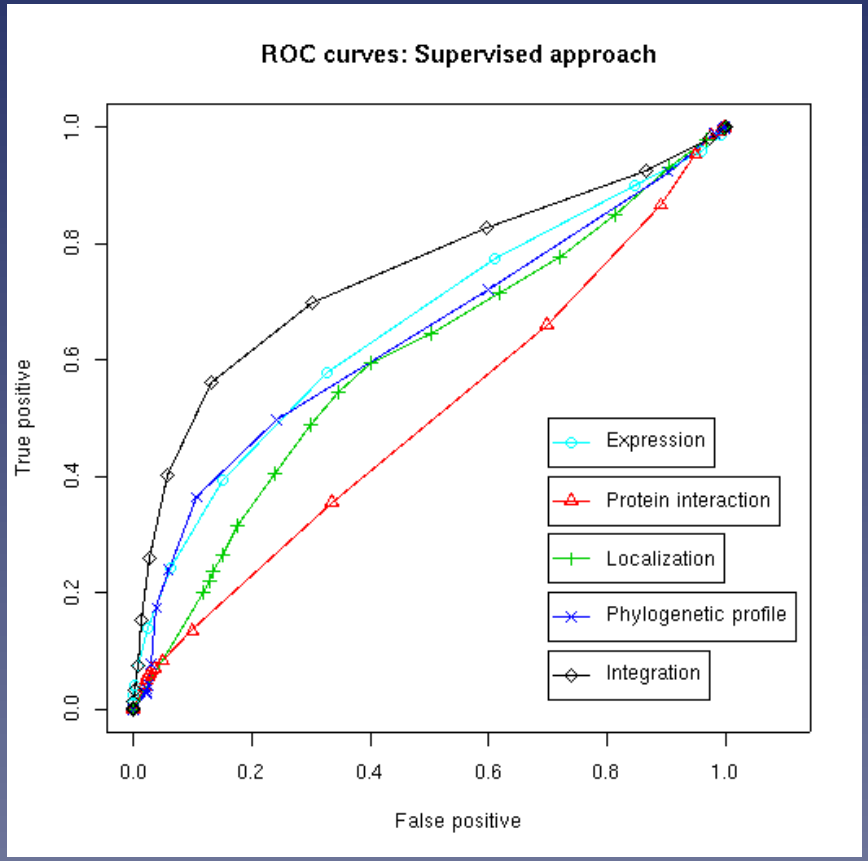
- Prédit les arêtes par similarité décroissante
- Equivalent à résoudre:

$$(LK_V + \lambda I)\alpha = \mu K_V \alpha.$$

Apprentissage de graphe (non supervisé)



Apprentissage de graphe (supervisé)



Conclusion

Conclusion

- Les noyaux sont utiles pour la représentation et l'intégration de données biologiques; algorithmes puissants
- Beaucoup de travail récent sur la construction de noyaux; leur optimisation et apprentissage joueront un rôle croissant
- Utile pour l'apprentissage statistique, mais comment augmenter la sémantique de cette représentation ?
- Validation de ces approches par des résultats utilisables en biologie et médecine

Remerciements

- Ecole des Mines : M. Cuturi, P. Mahé, M. Hue, J. Vermorel, F. Rapaport, C. Lajaunie, V. Stoven, JP Chiles, M. Schmitt, B. Legait...
- Université de Kyoto: Y. Yamanishi, M. Kanehisa, T. Akutsu, H. Saigo, N. Ueda, ...
- Autre: K. Tsuda, B. Schölkopf, W. Noble, N. Cristianini, M. Jordan, T. Evgeniou, J. Abernethy, R. Vert, A. Tsybakov, N. Vayatis, E. Barillot, P. David, JY. Coppée, I. Kondor, ...