Inférence de réseaux d'interaction protéine-protéine par apprentissage statistique

Application au réseau d'interaction autour de la protéine CFTR

Céline Brouard

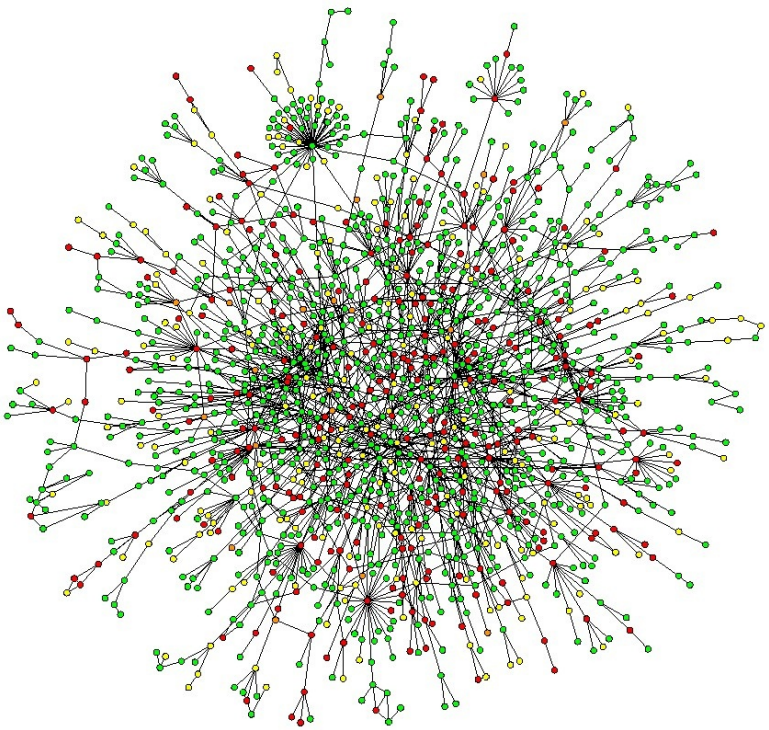
Direction : Florence d'Alché-Buc et Aleksander Edelman

14 février 2013

Interactions protéine-protéine

- La plupart des protéines réalisent leurs fonctions en interagissant avec d'autres protéines

Réseau d'interaction protéine-protéine :

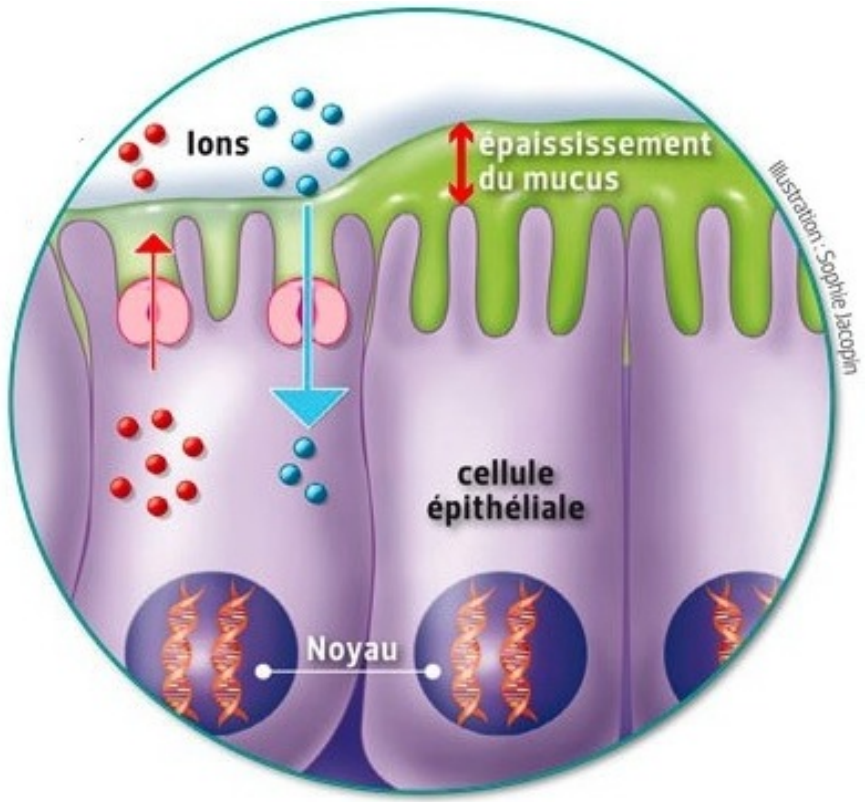


Réseau PPI de la levure

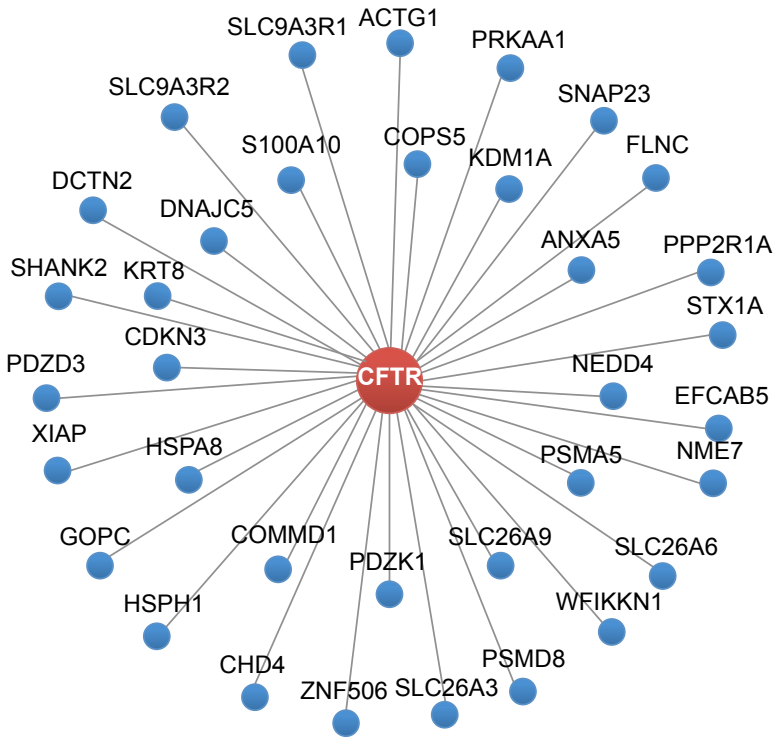
- Nœuds $\leftrightarrow$ protéines
- Une arête entre deux nœuds signifie l'existence d'une interaction entre les protéines correspondantes
- La dynamique et la localisation des interactions dans la cellule ne sont pas prises en compte

La protéine CFTR et son implication dans la mucoviscidose

- **Mucoviscidose** :
  - maladie létale d'origine génétique
  - liée à des mutations du gène *CFTR*, entraînant une altération de la protéine codée par ce gène
- **Protéine CFTR**
  - fonction principale : régulation du transport des ions chlorure à travers la membrane cellulaire



La protéine CFTR et son implication dans la mucoviscidose



- Capacité de CFTR à interagir avec un grand nombre de protéines
  $\Rightarrow$ Impact sur la stabilité, la localisation et la fonction de CFTR
- Importance de l'identification de ces interactions pour mieux comprendre le fonctionnement et la régulation de CFTR

Motivation (4)

- Limitations des méthodes de détection expérimentale existantes

Objectifs de la thèse

- développer des méthodes de prédiction *in silico* de PPI qui puissent être appliquées chez l'homme
- proposer un cadre général pour résoudre ce problème

Apprentissage supervisé (1)

- Ensemble d'observations $x_i \in \mathcal{X}$ associées à des sorties $y_i \in \mathcal{Y}$, appelées étiquettes :

$$\left.\begin{array}{cc} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{array}\right\} \text{Ensemble d'apprentissage}$$

$$\left.\begin{array}{ccc} x & \rightarrow & ? \end{array}\right\} \text{Nouvelle prédiction}$$

Objectif :

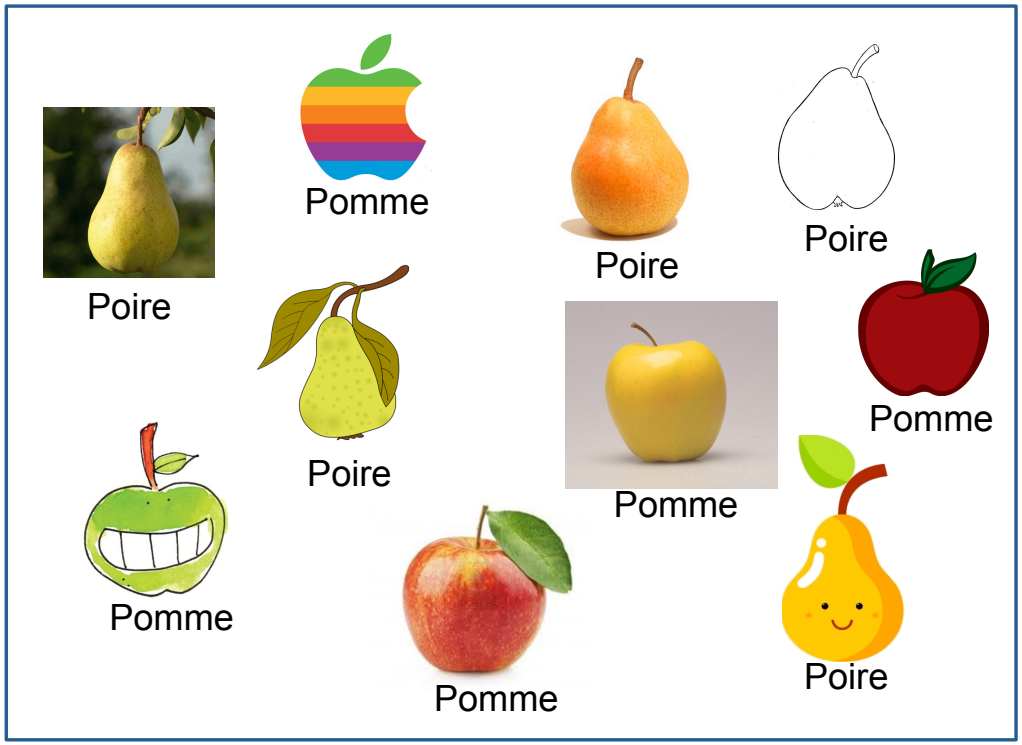
Apprendre une fonction $f : \mathcal{X} \rightarrow \mathcal{Y}$ afin de prédire l'étiquette associée à une nouvelle observation

- Classification / Régression

Apprentissage supervisé (2)

Exemple : classification d'images

Ensemble d'apprentissage



Pomme/Poire ?

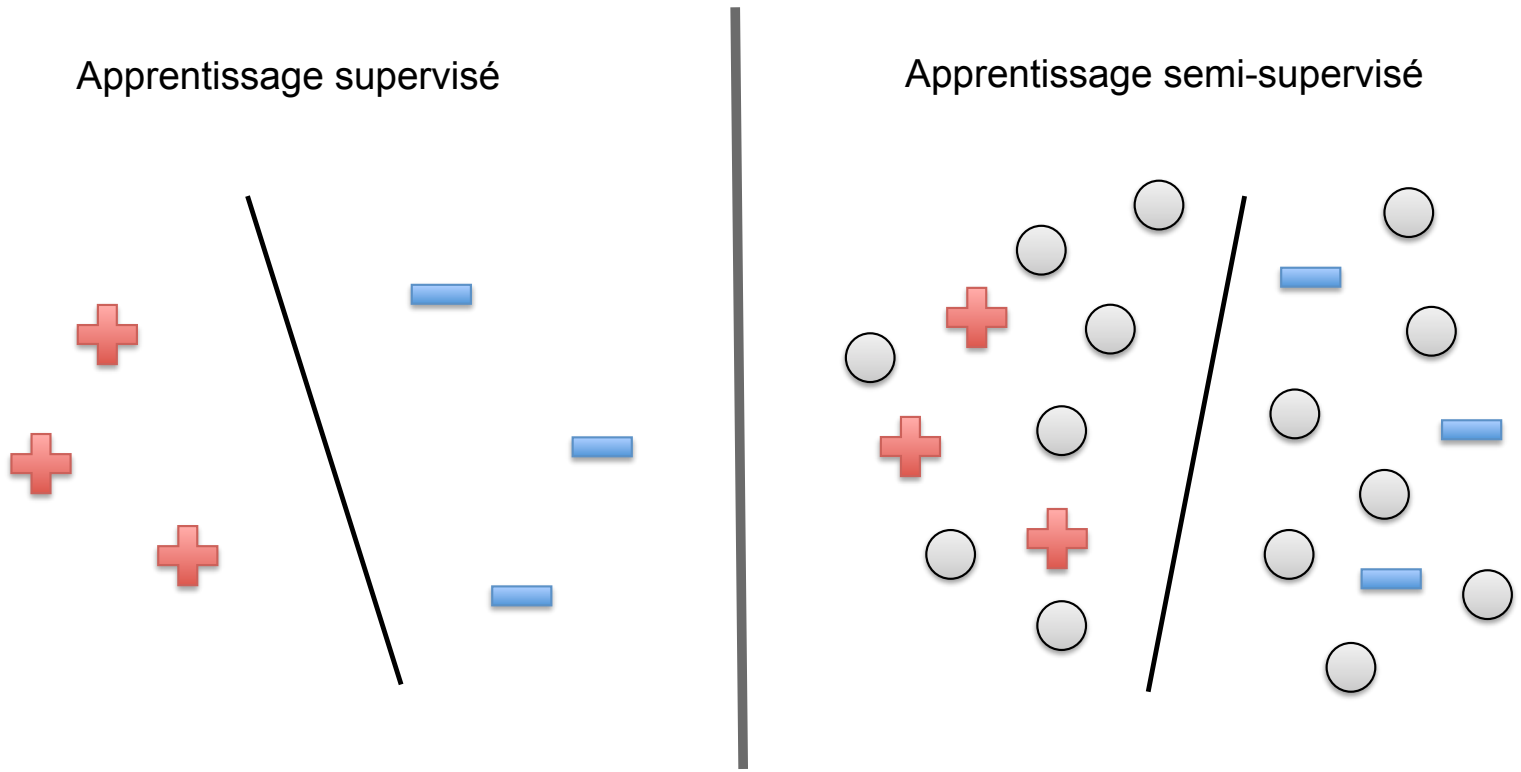
Apprentissage supervisé (3)

Exemples d'applications en bioinformatique :

- Diagnostic / prognostic (cancer)
- Prédiction de fonction
- Prédiction de structures
- Prédiction de localisations
- Inférence de réseaux de régulation
- ...

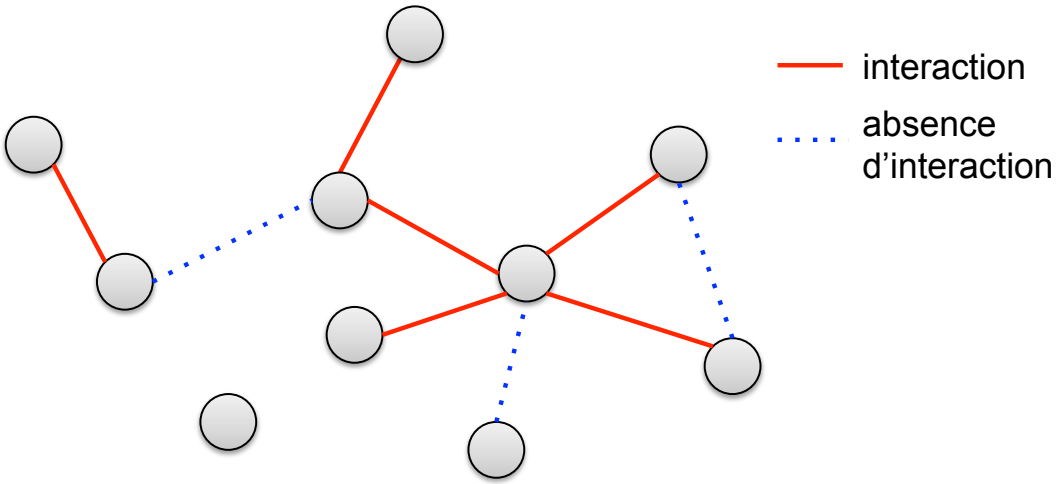
Apprentissage semi-supervisé

- Utilise à la fois des données étiquetées et des données non étiquetées lors de l'apprentissage



- Cas particulier : apprentissage transductif

Apprentissage supervisé pour la prédiction de liens



x_i : paire de protéines (u,u')
y_i : étiquette liée à la présence ou l'absence d'arête

Objectif : apprendre une fonction de prédiction

$$f : (u, u') \longrightarrow \begin{cases} 1 & \text{s'il existe une interaction entre les protéines } u \text{ et } u' \\ 0 & \text{sinon} \end{cases}$$

à partir

- d'un ensemble d'interactions et d'absences d'interactions connues
- d'informations sur les protéines

Descriptions des protéines

- Séquence d'acides aminés
- Structure tridimensionnelle
- Localisation cellulaire
- Annotations fonctionnelles
- Expression du gène qui code pour la protéine
- Signature phylogénétique
- Domaines
- Interologues
- ...

Difficultés inhérentes au problème

- Peu ou pas d'absences d'interaction confirmées, et donc présence possible de faux négatifs

- Peu de données étiquetées (interactions connues)

  $\Rightarrow$ résolution du problème dans le cadre de l'apprentissage semi-supervisé

Approches existantes pour la prédiction de liens

Approches supervisées

- SVM avec un noyau entre paires de nœuds [Ben-Hur & Noble, 2005] [Martin et al., 2005] [Vert et al., 2007] [Hue et al., 2010]
- Apprentissage d'un noyau ou d'une similarité
  - kCCA [Yamanishi et al., 2004]
  - Apprentissage de métrique [Vert & Yamanishi, 2005]
  - Régression à noyau de sortie [Geurts et al., 2006, 2007]
- Classifieurs locaux [Bleakley et al., 2007]
- Approches d'ensemble [Qi et al., 2007] [De Vienne & Azé, 2012]

Approches transductives

- Expansion de l'ensemble d'apprentissage [Yip & Gerstein, 2009]
- Complétion de matrice [Kato et al., 2005] [Yamanishi & Vert, 2007]
- Propagation de liens [Kashima et al., 2009]

Méthodes à noyaux (1)

Fonction noyau : $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

- fonction symétrique et semi-définie positive :

$$\forall (x, x') \in \mathcal{X} \times \mathcal{X},\ k(x, x') = k(x', x),$$

$$\forall n \in \mathbb{N}, \forall c_1, \ldots, c_n \in \mathbb{R}, \forall x_1, \ldots, x_n \in \mathcal{X},\ \sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

Théorème [Aronszajn, 1950]

k est un noyau semi-défini positif sur $\mathcal{X}$ si et seulement si il existe un espace de Hilbert $\mathcal{F}_x$ et une fonction $\phi : \mathcal{X} \to \mathcal{F}_x$ tels que

$$\forall x, x' \in \mathcal{X},\ k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}_x}$$

- Exemples :
  - Noyau linéaire : $k(x, x') = \langle x, x' \rangle$
  - Noyau polynomial : $k(x, x') = (\langle x, x' \rangle + c)^d$
  - Noyau gaussien : $k(x, x') = exp(-\|x - x'\|^2 / 2\sigma^2)$

Méthodes à noyaux (2)

noyau $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$\downarrow$

Espace de Hilbert à noyau reproduisant (RKHS) $\mathcal{H}$

$\downarrow$

$\arg\min_{h \in \mathcal{H}} \sum_{i=1}^{\ell} V(h(x_i), y_i) + \lambda_1 \|h\|_{\mathcal{H}}^2 \to$ ex : $V(h(x_i), y_i) = (h(x_i) - y_i)^2$

$\downarrow$

Théorème de représentation :

$\hat{h}(x) = \sum_{i=1}^{\ell} \alpha_i k(x, x_i)$

$\downarrow$

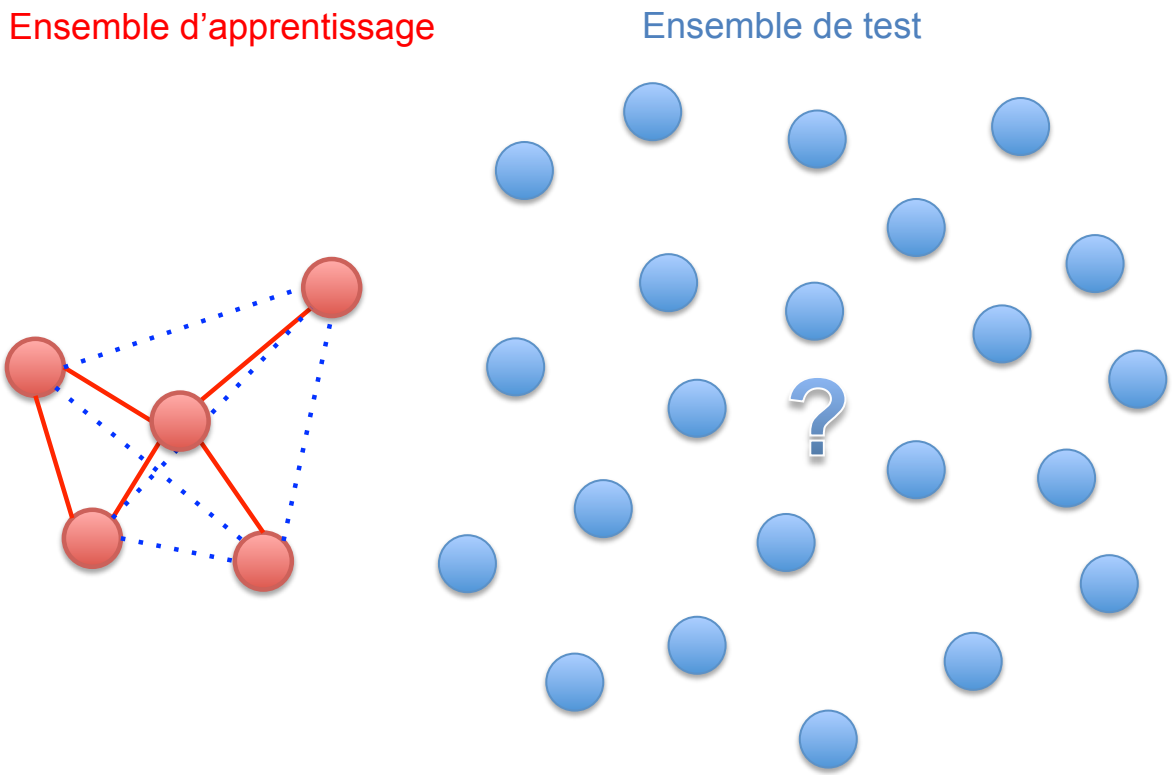
Incorporation du modèle dans le problème d'optimisation

$\downarrow$

Résolution du problème d'optimisation

Sommaire

Cadre pour la prédiction de liens



- $\mathcal{U}$: ensemble des nœuds
- $\mathcal{U}_\ell$: ensemble de nœuds pour lesquels les présences et absences de liens sont supposées connues
- A_ℓ : matrice d'adjacence du sous-graphe connu

Inférence supervisée de graphe par apprentissage d'un noyau

Conversion du problème de classification binaire à partir de paires d'objets en un **problème d'apprentissage d'un noyau** :

- κ_y : noyau de sortie mesurant la similarité entre protéines en tant que nœuds dans le graphe
- Construction d'une fonction de classification à partir de $\hat{\kappa_y}$:

$$\forall (u, u') \in \mathcal{U} \times \mathcal{U},\ f_\theta(u, u') = \text{sgn}(\hat{\kappa_y}(u, u') - \theta)$$

Régression à noyau de sortie

κ_y étant un noyau semi-défini positif, il existe un espace de Hilbert $\mathcal{F}_y$ et une fonction caractéristique $y : \mathcal{U} \to \mathcal{F}_y$ tels que :

$$\forall (u, u') \in \mathcal{U} \times \mathcal{U},\ \kappa_y(u, u') = \langle y(u), y(u') \rangle_{\mathcal{F}_y}$$

- Approximation de y par une fonction $h : \mathcal{U} \to \mathcal{F}_y \in \mathcal{H}$

$$\forall (u, u') \in \mathcal{U} \times \mathcal{U},\ \hat{\kappa_y}(u, u') = \langle h(u), h(u') \rangle_{\mathcal{F}_y}$$

- Apprentissage de h : **régression à noyau de sortie** (OKR)

  $\Rightarrow$ au lieu d'apprendre un classifieur à partir de paires d'objets, on apprend une fonction d'une seule variable à valeurs dans un espace de Hilbert.

- Travaux précédents : *Output Kernel Regression Trees* [Geurts et al., 2006, 2007]

Régression à noyau de sortie

κ_y étant un noyau semi-défini positif, il existe un espace de Hilbert $\mathcal{F}_y$ et une fonction caractéristique $y : \mathcal{U} \to \mathcal{F}_y$ tels que :

$$\forall (u, u') \in \mathcal{U} \times \mathcal{U},\ \kappa_y(u, u') = \langle y(u), y(u') \rangle_{\mathcal{F}_y}$$

- Approximation de y par une fonction $h : \mathcal{U} \to \mathcal{F}_y \in \mathcal{H}$

$$\forall (u, u') \in \mathcal{U} \times \mathcal{U},\ \hat{\kappa_y}(u, u') = \langle h(u), h(u') \rangle_{\mathcal{F}_y}$$

- Apprentissage de h : **régression à noyau de sortie** (OKR)

  $\Rightarrow$ au lieu d'apprendre un classifieur à partir de paires d'objets, on apprend une fonction d'une seule variable à valeurs dans un espace de Hilbert.

- Travaux précédents : *Output Kernel Regression Trees* [Geurts et al., 2006, 2007]

Régression à noyau de sortie dans le cadre semi-supervisé

Extension de la régression à noyau de sortie au cadre de l'apprentissage semi-supervisé :

- Les méthodes à base d'arbres ne sont pas appropriées
- La théorie des RKHS fournit un cadre rigoureux pour la régression semi-supervisée [Zhou et al., 2004 ; Belkin et al., 2006] dans le cas de fonctions à valeurs scalaires.

$\Rightarrow$ Dans le cas d'OKR : se tourner vers la théorie des RKHS dédiée aux fonctions à valeurs dans un espace de Hilbert.

Sommaire

Théorie des RKHS pour les fonctions à valeurs dans un espace de Hilbert

Théorie des RKHS avec des **noyaux à valeur opérateur** [Senkene & Tempel'man, 1973 ; Michelli & Pontil, 2005 ; Caponetto et al., 2008]

- Applications existantes :
  - Apprentissage multitâches [Michelli & Pontil, 2005 ; Argyriou & Pontil, 2008]
  - Prédiction de données fonctionnelles [Kadri et al., 2010]
  - Classification structurée [Dinuzzo et al., 2011]

Définition d'un noyau à valeur opérateur

- $\mathcal{F}_y$: espace de Hilbert
- $\mathcal{L}(\mathcal{F}_y)$: ensemble des opérateurs linéaires bornés de $\mathcal{F}_y$ dans lui-même.

Noyau à valeur opérateur :
(Senkene & Tempel'man, 1973 ; Caponnetto et al., 2008)

$\mathcal{K}_x : \mathcal{U} \times \mathcal{U} \to \mathcal{L}(\mathcal{F}_y)$ est un noyau à valeur opérateur si :

- $\forall (u, u') \in \mathcal{U} \times \mathcal{U},\ \mathcal{K}_x(u, u') = \mathcal{K}_x(u', u)^*$
- $\forall m \in \mathbb{N},\ \forall \{(u_i, \mathbf{y}_i)\}_{i=1}^m \subseteq \mathcal{U} \times \mathcal{F}_y,$

$$\sum_{i,j=1}^{m} \langle \mathbf{y}_i, \mathcal{K}_x(u_i, u_j)\mathbf{y}_j \rangle_{\mathcal{F}_y} \geq 0 \ .$$

Construire un RKHS à partir d'un noyau à valeur opérateur

Théorème (Senkene & Tempel'man, 1973 ; Micchelli & Pontil, 2005)

Si $\mathcal{K}_x : \mathcal{U} \times \mathcal{U} \to \mathcal{L}(\mathcal{F}_y)$ est un noyau à valeur opérateur, alors il existe un unique RKHS $\mathcal{H}$ admettant $\mathcal{K}_x$ comme noyau reproduisant, c'est à dire qui vérifie :

$$\forall u \in \mathcal{U}, \forall \mathbf{y} \in \mathcal{F}_y,\ \langle h, \mathcal{K}_x(\cdot, u)\mathbf{y} \rangle_{\mathcal{H}} = \langle h(u), \mathbf{y} \rangle_{\mathcal{F}_y}$$

Théorème de représentation dans le cas supervisé

- $\{(u_i, \mathbf{y}_i)\}_{i=1}^{\ell} \subseteq \mathcal{U} \times \mathcal{F}_y$: exemples d'apprentissage

Problème d'optimisation

$$\arg\min_{h \in \mathcal{H}} \sum_{i=1}^{\ell} \|h(u_i) - \mathbf{y}_i\|_{\mathcal{F}_y}^2 + \lambda_1 \|h\|_{\mathcal{H}}^2 \ ,$$

avec $\lambda_1 > 0$

[Micchelli & Pontil 2005] ont montré que

$$\hat{h}(\cdot) = \sum_{j=1}^{\ell} \mathcal{K}_x(\cdot, u_j)\mathbf{c}_j, \quad \mathbf{c}_j \in \mathcal{F}_y, j = 1, \cdots, \ell.$$

Les vecteurs $\mathbf{c}_j$ sont solutions de : $\sum_{i=1}^{\ell}(\lambda_1 \delta_{ij} + \mathcal{K}_x(u_j, u_i))\mathbf{c}_i = \mathbf{y}_j$.

Théorème dans le cas semi-supervisé (1)

Ajout d'une contrainte de continuité pour bénéficier des données non étiquetées (Zhou et al. 2004, Belkin et al. 2006).

- $\{u_i\}_{i=\ell+1}^{\ell+n} \subseteq \mathcal{U}$: ensemble additionnel d'exemples non étiquetés,
- W : matrice mesurant la similarité entre les objets dans l'espace d'entrée.

Problème d'optimisation

$$\arg\min_{h\in\mathcal{H}} \sum_{i=1}^{\ell} \|h(u_i) - \mathbf{y}_i\|_{\mathcal{F}_y}^2 + \lambda_1 \|h\|_{\mathcal{H}}^2 + \lambda_2 \sum_{i,j=1}^{\ell+n} W_{ij} \|h(u_i) - h(u_j)\|_{\mathcal{F}_y}^2 ,$$

avec λ_1 et $\lambda_2 > 0$,

Théorème dans le cas semi-supervisé (2)

$$\arg\min_{h\in\mathcal{H}} \sum_{i=1}^{\ell} \|h(u_i) - \mathbf{y}_i\|^2_{\mathcal{F}_y} + \lambda_1 \|h\|^2_{\mathcal{H}} + \lambda_2 \sum_{i,j=1}^{\ell+n} W_{ij} \|h(u_i) - h(u_j)\|^2_{\mathcal{F}_y}$$

$$\downarrow$$

Réécriture du problème d'optimisation :

$$\arg\min_{h\in\mathcal{H}} \sum_{i=1}^{\ell} \|h(u_i) - \mathbf{y}_i\|^2_{\mathcal{F}_y} + \lambda_1 \|h\|^2_{\mathcal{H}} + 2\lambda_2 \sum_{i,j=1}^{\ell+n} L_{ij} \langle h(u_i), h(u_j)\rangle_{\mathcal{F}_y},$$

où L est le Laplacien associé à W

Théorème dans le cas semi-supervisé (3)

Théorème [Brouard, d'Alché-Buc and Szafranski, 2011]

La fonction $\hat{h}$ minimisant ce problème d'optimisation admet la forme suivante :

$$\hat{h}(\cdot) = \sum_{j=1}^{\ell+n} \mathcal{K}_x(\cdot, u_j)\mathbf{c}_j, \quad \mathbf{c}_j \in \mathcal{F}_y.$$

Les vecteur $\mathbf{c}_j$ sont solutions du système d'équations suivant :

$$V_j\mathbf{y}_j = V_j \sum_{i=1}^{\ell+n} \mathcal{K}_x(u_j, u_i)\mathbf{c}_i + \lambda_1\mathbf{c}_j + 2\lambda_2 \sum_{i=1}^{\ell+n} L_{ij} \sum_{m=1}^{\ell+n} \mathcal{K}_x(u_i, u_m)\mathbf{c}_m,$$

où $V_j = \begin{cases} I_{dim(\mathcal{F}_y)} & \text{lorsque } j \leq \ell \\ 0_{dim(\mathcal{F}_y)} & \text{lorsque } \ell \leq j \leq (\ell+n) \end{cases}$

Sommaire

Régression à noyaux d'entrée et de sortie pour la prédiction de liens (1)

- IOKR (Input Output Kernel Regression)
- Choix d'une paire $(\kappa_y, \mathcal{K}_x)$

Noyau de sortie utilisé : noyau de diffusion [Kondor & Lafferty, 2002]

- Définition d'une matrice de Gram K_{Y_ℓ} à partir de A_ℓ :

$$K_{Y_\ell} = \exp(-\beta L),$$

où $L = D_\ell - A_\ell$ avec D_ℓ la matrice diagonale des degrés.

- $\kappa_{y,\ell} : \mathcal{U}_\ell \times \mathcal{U}_\ell \to \mathbb{R}$, noyau associé à K_{Y_ℓ} :

$$\forall (u, u') \in \mathcal{U}_\ell \times \mathcal{U}_\ell,\ K_{Y_\ell}(u, u') = \kappa_{y,\ell}(u, u').$$

Régression à noyaux d'entrée et de sortie pour la prédiction de liens (2)

- Résultats théoriques sur les noyaux à valeur opérateur décomposables : $\mathcal{K}_x(u,u') = \kappa_x(u,u') \times A$, ($A$ de taille $dim(\mathcal{F}_y) \times dim(\mathcal{F}_y)$.)

Noyau à valeur opérateur en entrée

On définit $\mathcal{K}_x$ de la façon suivante :

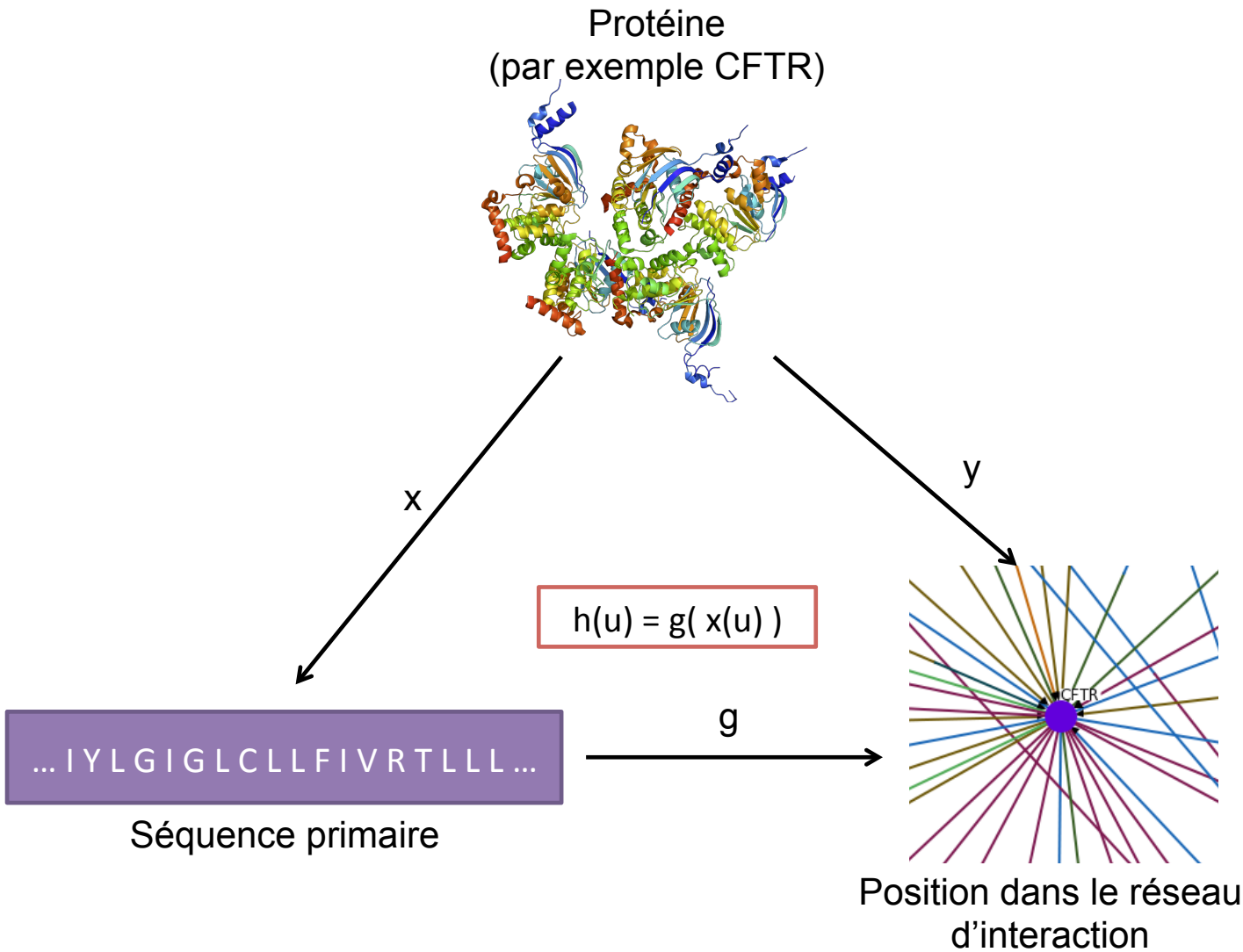
$$\begin{aligned} \mathcal{K}_x : \quad \mathcal{U} \times \mathcal{U} &\to \mathcal{L}(\mathcal{F}_y) \\ (u,u') &\mapsto \kappa_x(u,u') \times I_{dim(\mathcal{F}_y)} \end{aligned}$$

- $\kappa_x : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$: noyau scalaire en entrée
- Il existe donc un espace de Hilbert $\mathcal{F}_x$ et une fonction caractéristique $x : \mathcal{U} \to \mathcal{F}_x$ tels que :

$$\forall (u,u') \in \mathcal{U} \times \mathcal{U},\ \kappa_x(u,u') = \langle x(u), x(u') \rangle_{\mathcal{F}_x}$$

IOKR pour la prédiction de liens

Cas du noyau décomposable $\mathcal{K}_x(u, u') = \kappa_x(u, u') \times I_{dim(\mathcal{F}_y)}$



Modèles

Notations :

- $Y_\ell = [y(u_1), \ldots, y(u_\ell)]$
- $X_\ell = [x(u_1), \ldots, x(u_\ell)], X_{\ell+n} = [x(u_1), \ldots, x(u_{\ell+n})]$
- $K_{X_\ell} = X_\ell^T X_\ell, K_{X_{\ell+n}} = X_{\ell+n}^T X_{\ell+n}$
- $U = [I_\ell, 0_{\ell \times (\ell+n)}]$

Cas supervisé

$$\hat{h}(u) = Y_\ell \left(\lambda_1 I_\ell + K_{X_\ell}\right)^{-1} X_\ell^T x(u)$$

- modèle linéaire proposé par Cortes et al. (2005) dans le cadre de la reformulation de KDE (Kernel Dependency Estimation).

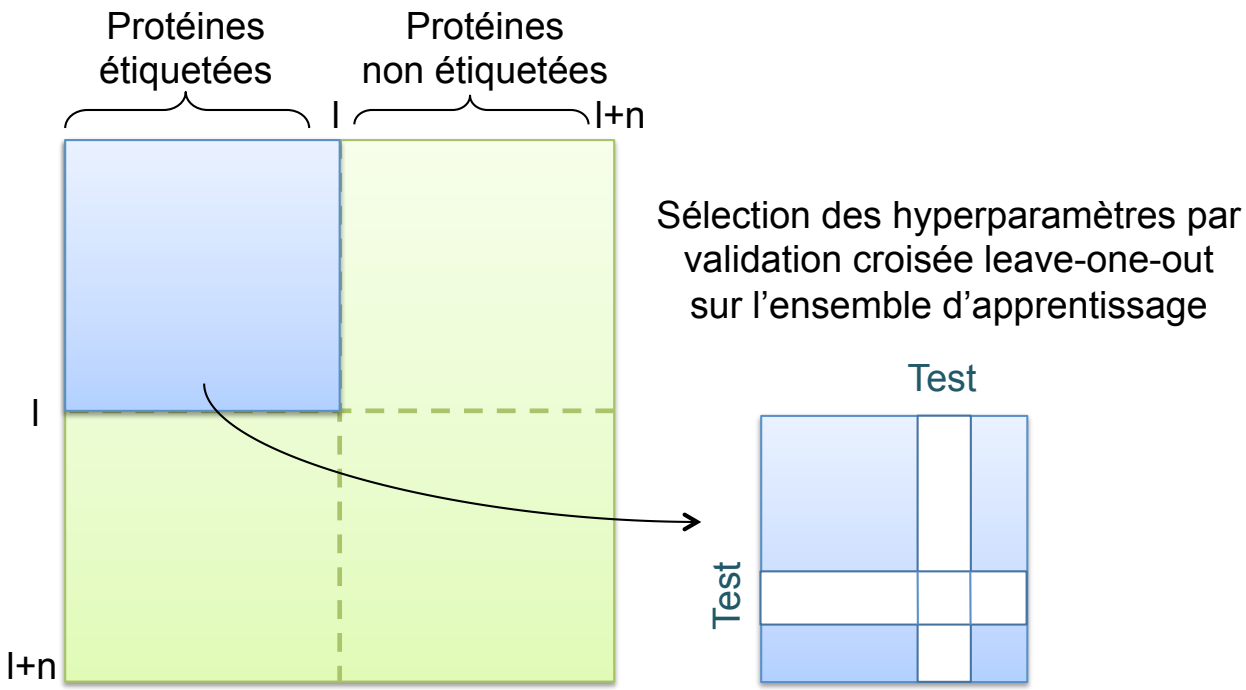
Cas semi-supervisé

$$\hat{h}(u) = Y_\ell \left(\lambda_1 I_{\ell+n} + K_{X_{\ell+n}}(U^T U + 2\lambda_2 L)\right)^{-1} X_{\ell+n}^T x(u)$$

Sommaire

Protocole (1)

- Cadre transductif
- Pour différentes valeurs de ℓ, nous avons échantillonné uniformément un sous-ensemble de nœuds pour l'ensemble d'apprentissage et utilisé les autres comme exemples de test. (répété 10 fois)



- Remarque : 10% de nœuds étiquetés $\rightarrow$ 1% seulement d'interactions étiquetées

Protocole (2)

Classe prédite \ Classe réelle	1	0
1	Vrais positifs (TP)	Faux positifs (FP)
0	Faux négatifs (FN)	Vrais négatifs (TN)
	P = TP + FN	N = FP + TN

TPR = TP/P
FPR = FP/N

Précision = TP / (TP + FP)
Rappel = TP / P = TPR

- **Courbe ROC** : comportement du taux de vrais positifs en fonction du taux de faux positifs
- **Courbe Précision-Rappel (PR)** : comportement de la précision en fonction du rappel.

Mesures d'évaluation :
aires sous les courbes ROC et PR (AUC-ROC et AUC-PR)

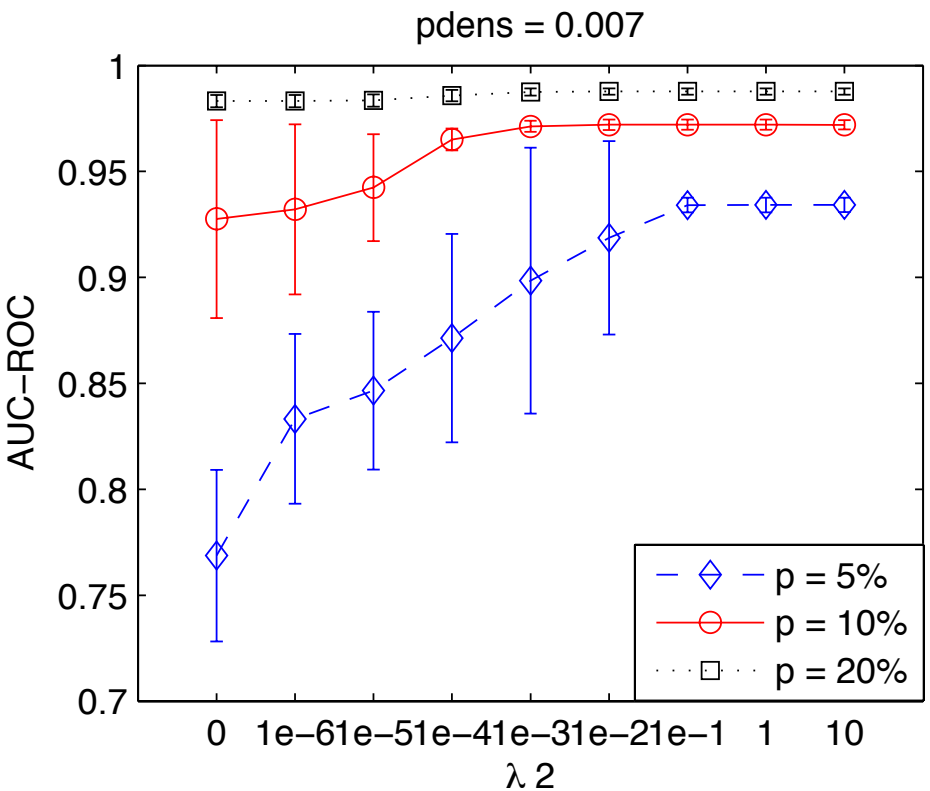
Réseaux synthétiques (1)

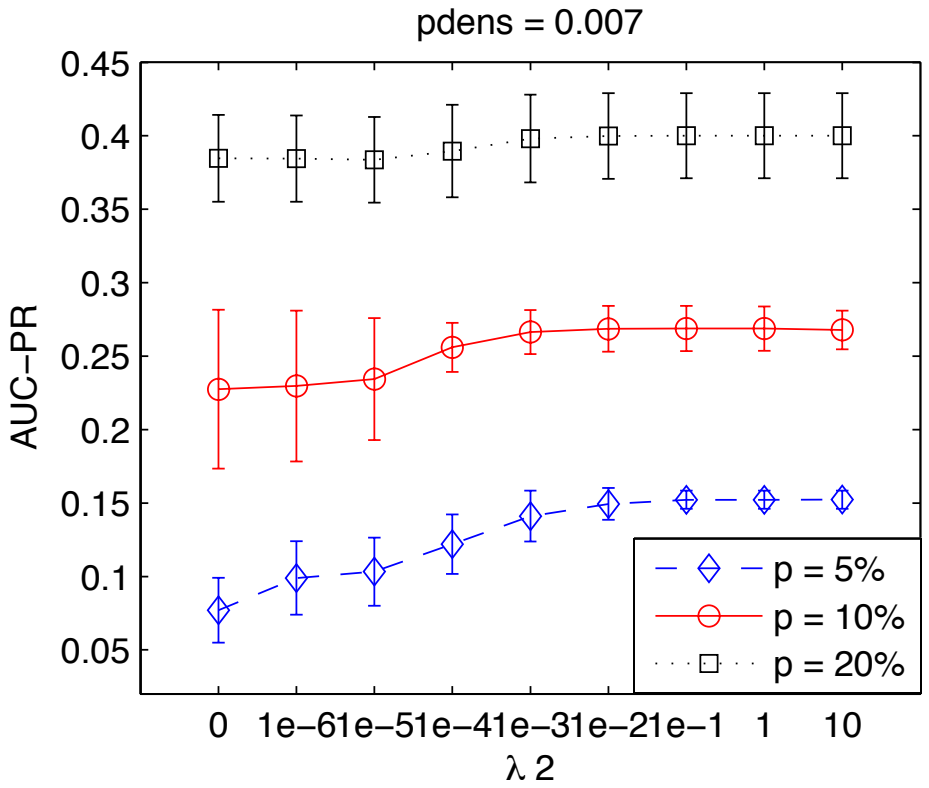
Test dans le cas idéal : l'information donnée en entrée correspond quasiment à l'information de sortie

- Graphes aléatoires (700 nœuds) échantillonnés à partir d'une loi de Erdős-Renyi
- Plusieurs densités : 0.007, 0.01 et 0.02
- Obtention des vecteurs caractéristiques en entrée :
  - Application de l'ACP à noyaux sur le noyau de diffusion associé au graphe
  - Utilisation des composantes permettant de capturer 95% de la variance

Réseaux synthétiques (2)

Réseaux de densité 0.007





Application au réseau PPI de la levure (1)

- Problème test
- Réseau : 984 protéines, 2438 interactions [von Mering et al., 2002 ; Kato et al., 2005]
- Entrées :
  - Expressions de gènes $\rightarrow k_{exp}$
  - Profils phylogéniques $\rightarrow k_{phy}$
  - Localisations $\rightarrow k_{loc}$
  - Données de double hybride de la levure $\rightarrow k_{y2h}$
- Combinaison des différents noyaux :

$$k_{int} = (k_{exp} + k_{loc} + k_{phy} + k_{y2h})/4$$

- Cadre supervisé : comparaison avec les résultats de [Bleakley et al., 2007] en utilisant la même procédure de 5-CV

Application au réseau PPI de la levure (2)

Comparaison réalisée par [Bleakley et al., 2007] complétée

a) AUC-ROC :

Méthodes	exp	loc	phy	y2h	int
kCCA	81.4 ± 1.1	49.1 ± 14.6	67.8 ± 2.2	48.1 ± 2.4	87.8 ± 0.9
kML	82.9 ± 1.2	76.3 ± 1.1	71.7 ± 1.8	64.4 ± 1.9	88.1 ± 1.2
EM	80.6 ± 1.1	76.7 ± 3.8	71.0 ± 1.3	57.2 ± 2.7	89.3 ± 1.1
Local	78.1 ± 1.1	77.1 ± 2.9	$\mathbf{75.5 \pm 2.4}$	77.8 ± 1.2	87.6 ± 1.8
OK3+ET	82.1 ± 1.1	$\mathbf{81.0 \pm 1.3}$	75.1 ± 1.9	$\mathbf{80.6 \pm 1.6}$	89.2 ± 1.2
IOKR-ridge	$\mathbf{83.3 \pm 2.1}$	74.7 ± 3.6	69.6 ± 1.5	60.8 ± 3.5	$\mathbf{91.0 \pm 0.4}$

b) AUC-PR :

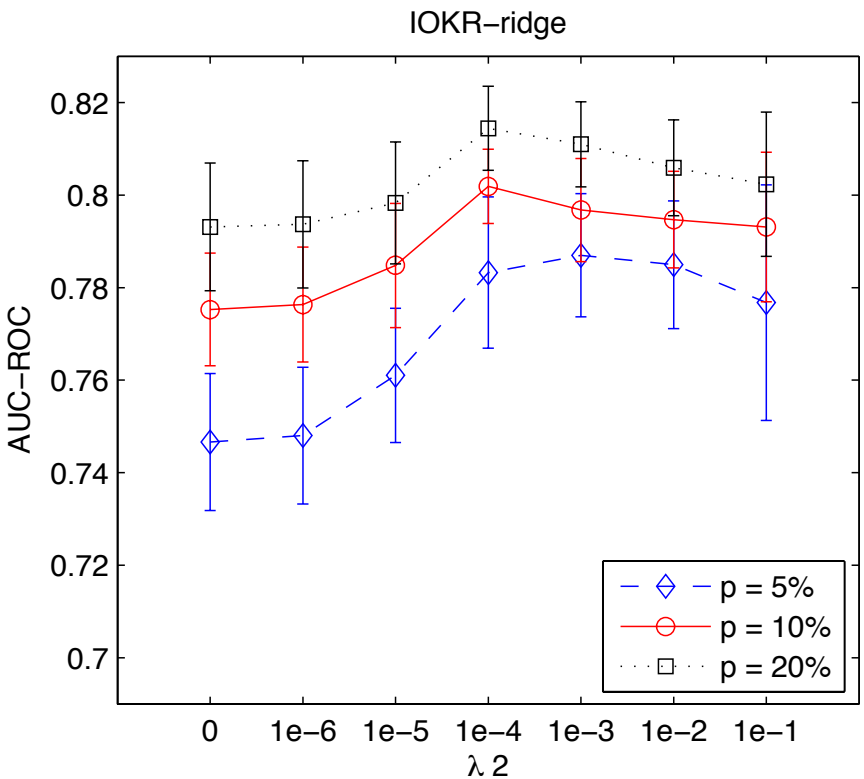
Méthodes	exp	loc	phy	y2h	int
kCCA	8.0 ± 1.9	2.1 ± 3.8	1.9 ± 0.4	2.1 ± 0.6	11.4 ± 1.5
kML	7.0 ± 2.4	1.2 ± 0.1	1.5 ± 0.2	0.8 ± 0.1	6.5 ± 2.2
EM	6.3 ± 1.2	5.5 ± 1.1	3.2 ± 0.5	10.4 ± 1.0	19.1 ± 1.3
Local	2.6 ± 0.4	3.7 ± 0.9	2.1 ± 0.3	7.6 ± 1.6	25.5 ± 3.4
OK3+ET	10.3 ± 3.7	$\mathbf{8.0 \pm 1.5}$	$\mathbf{3.7 \pm 0.9}$	5.7 ± 1.2	14.3 ± 3.8
IOKR-ridge	$\mathbf{13.7 \pm 4.4}$	7.0 ± 1.3	2.6 ± 0.4	$\mathbf{12.9 \pm 2.9}$	$\mathbf{27.2 \pm 6.5}$

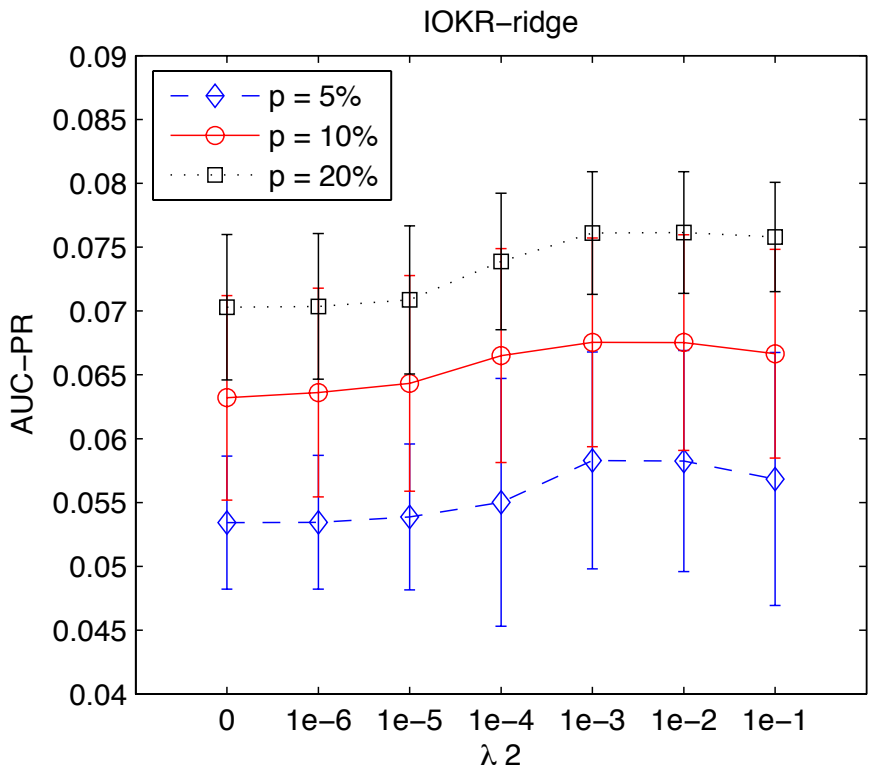
kCCA [Yamanishi et al., 2004] ; kML [Vert & Yamanishi, 2005] ; EM [Kato et al., 2005] ; Local [Bleakley et al., 2007] ; OK3+ET [Geurts et al., 2007]

Application au réseau PPI de la levure (3)

Cadre transductif

Noyau d'entrée : k_{exp}





Elaboration d'un nouveau jeu de données sur la levure

- Problèmes soulevés par le réseau précédent :
  - redondances entre les informations d'entrée et de sortie
  - interactions issues de complexes protéiques et de données indirectes (contexte génomique, expressions de gènes, interactions génétiques)

$\Rightarrow$ Construction d'un réseau d'interaction protéine-protéine chez la levure à partir de la base de données DIP (Database of Interacting Proteins)

- Considération des protéines annotées pour chacun des noyaux d'entrée et impliquées dans au moins une interaction

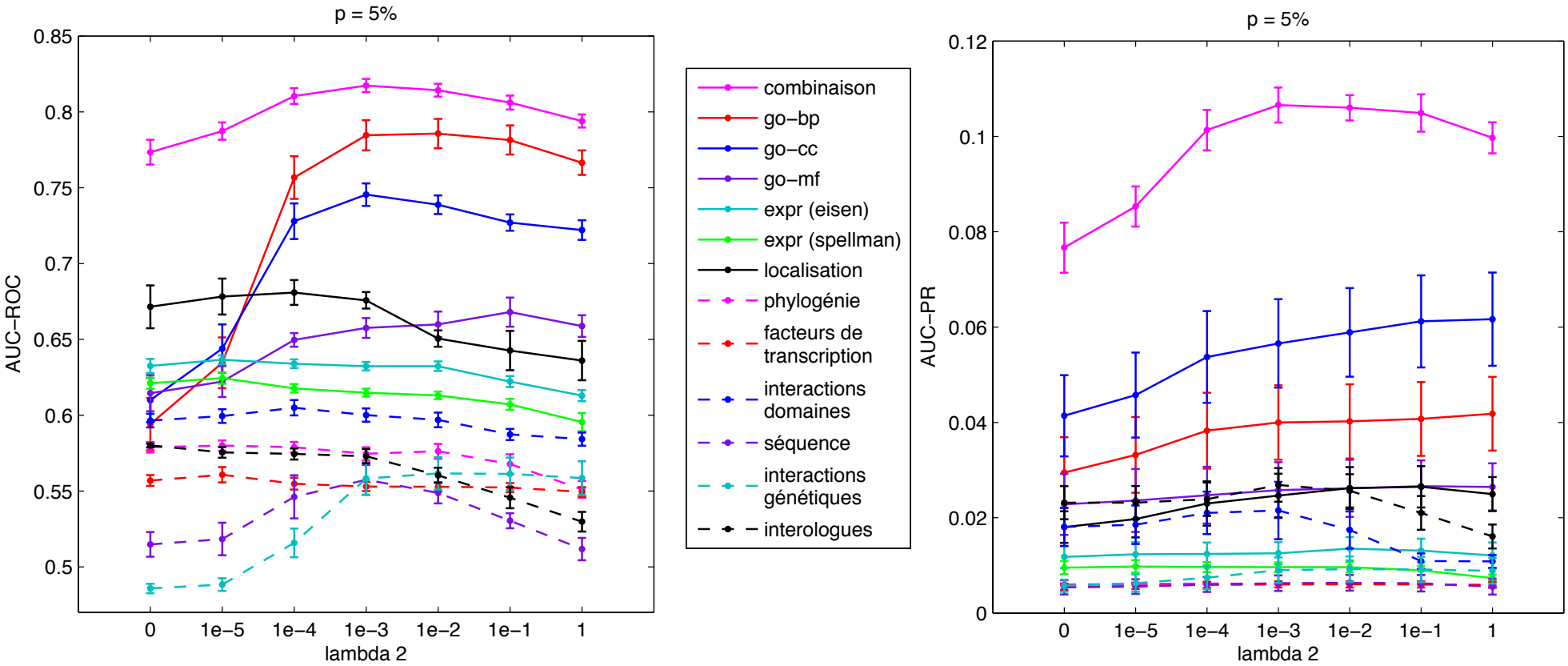
  $\Rightarrow$ obtention d'un réseau de 815 nœuds avec une densité de liens de 0.0054

Application au réseau PPI de la levure

Informations	**Type de noyau**
Expressions de gènes [Eisen et al., 1998]	gaussien
Expressions de gènes [Spellman et al., 1998]	gaussien
Localisations cellulaires [MIPS]	gaussien
Interactions génétiques [BioGRID]	gaussien
Séquence primaire [NCBI Protein]	k-spectrum
Interactions domaine-domaine [Pfam, DOMINE]	diffusion
Facteurs de transcription [YEASTRACT]	gaussien
Processus biologiques [Gene Ontology]	gaussien
Fonctions moléculaires [Gene Ontology]	gaussien
Composants cellulaires [Gene Ontology]	gaussien
Interologues [Inparanoid, DIP, MINT, BioGRID]	diffusion
Profils phylogénétiques [Phylopro]	gaussien

Apport de l'apprentissage semi-supervisé

Cas où 5% des données sont étiquetées

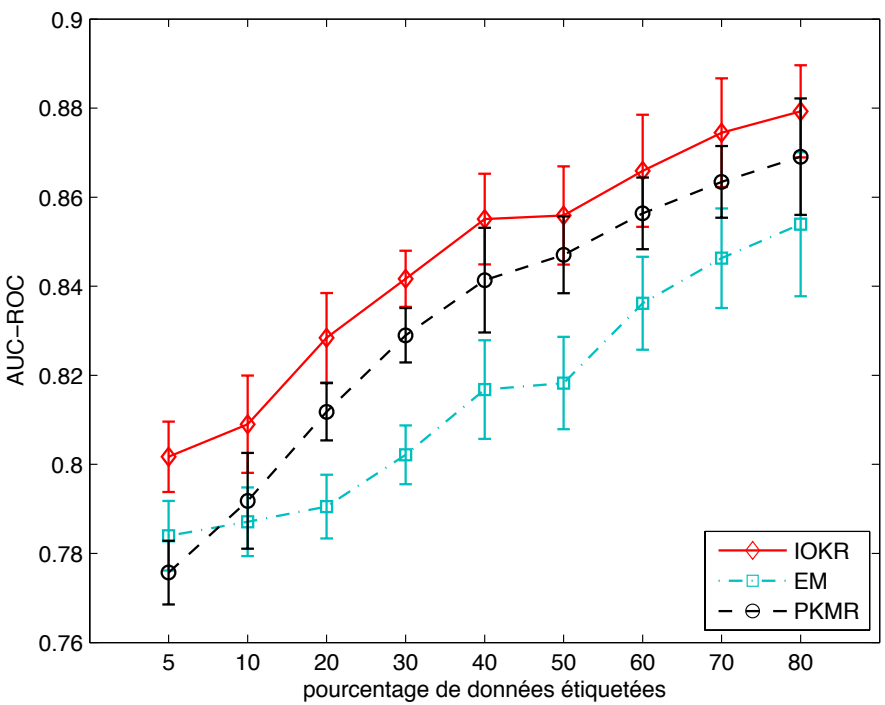


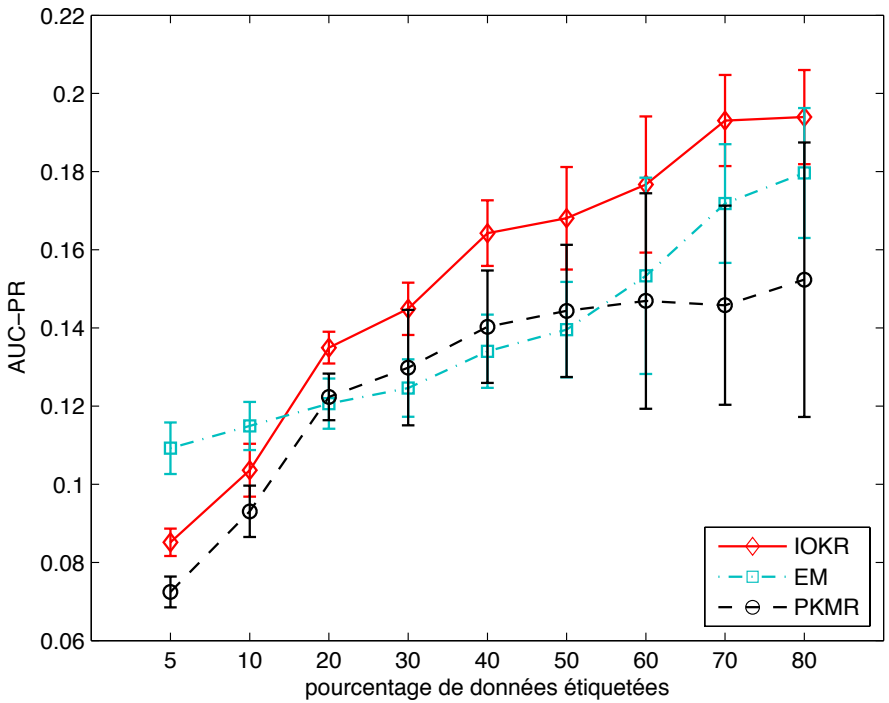
Combinaison : $\forall u, u' \in \mathcal{U}$, $\hat{\kappa_y}(u, u') = \frac{1}{p} \sum_{j=1}^{p} \hat{\kappa_y}^{(j)}(u, u')$,
où $\hat{\kappa_y}^{(j)}$ correspond à l'approximation du noyau de sortie obtenue lorsque le j-ème noyau est utilisé en entrée, et p au nombre de noyaux considérés.

Comparaison avec des approches transductives

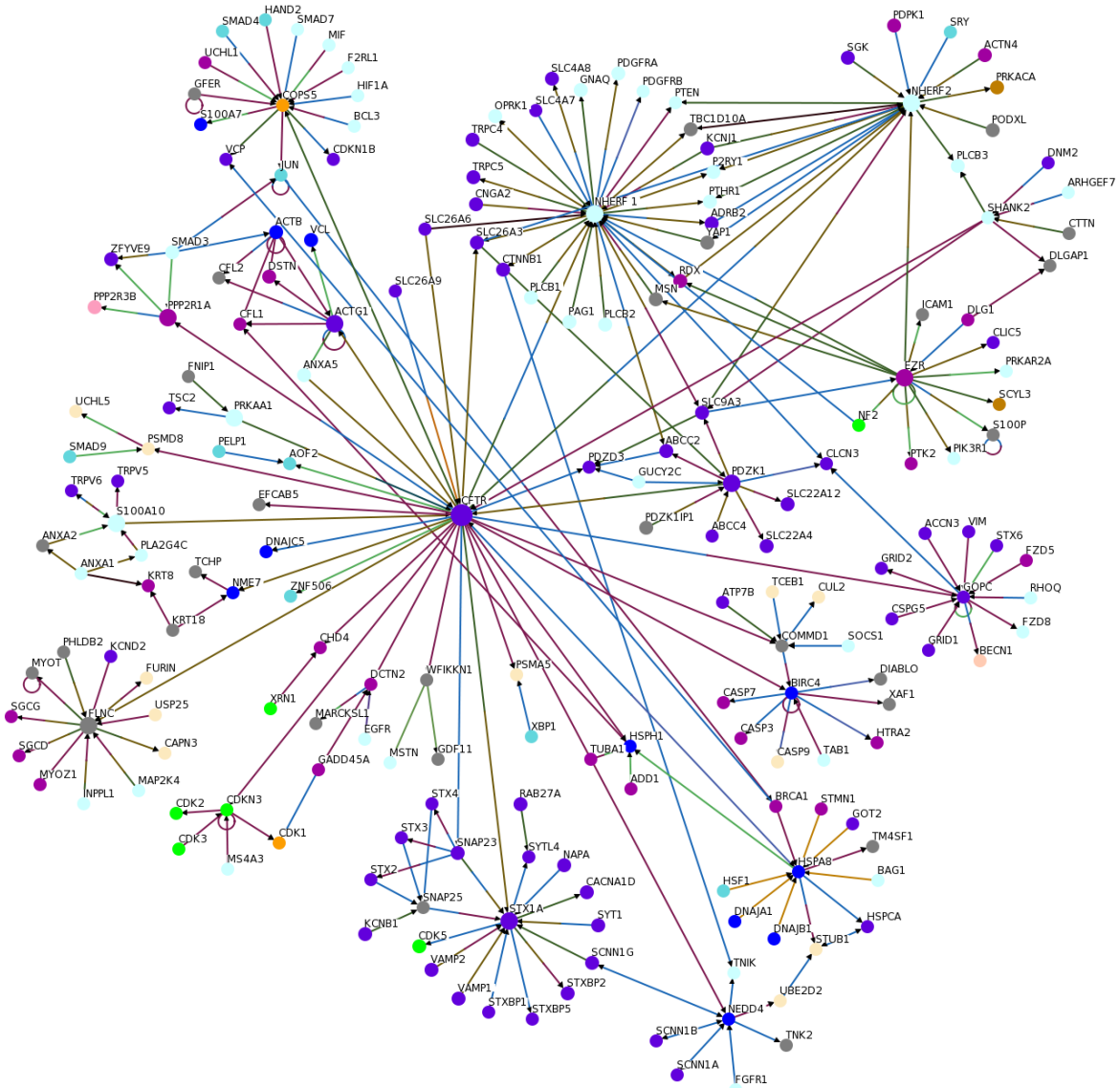
- EM [Kato et al., 2005]
- PKMR (Penalized Kernel Matrix Regression) [Yamanishi & Vert, 2007]

Résultats correspondant à
la combinaison des prédictions obtenues pour chacun des noyaux d'entrée :





Inférence du réseau d'interaction autour de CFTR (1)



- Réseau de 198 protéines
- Vérification manuelle des interactions dans la littérature (BioGRID, DIP, MINT, Intact, NextProt)

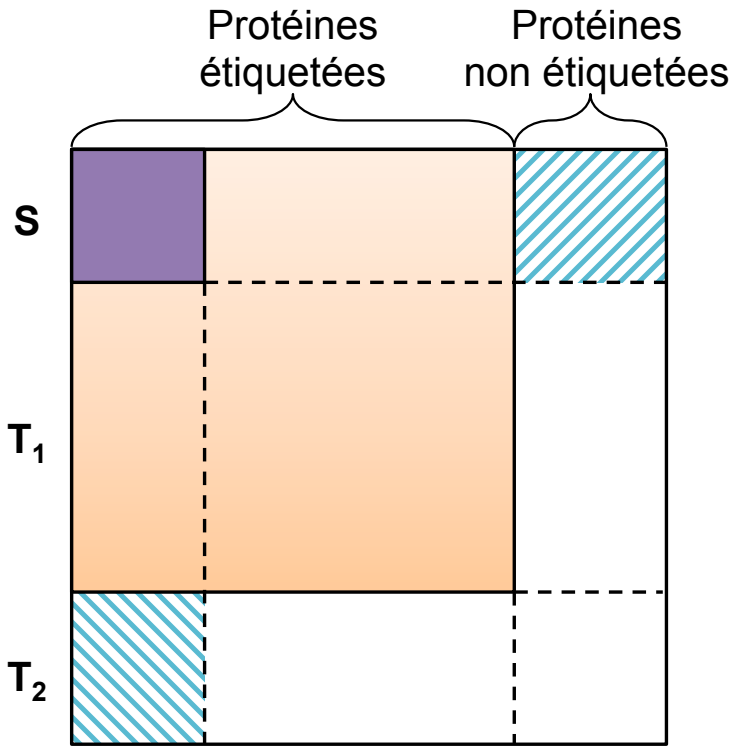
Inférence du réseau d'interaction autour de CFTR (2)

Informations	**Type de noyau**
Expressions des gènes [Su et al., 2004]	gaussien
Expressions des protéines [The Human Protein Atlas]	gaussien
Localisations cellulaires [The Human Protein Atlas]	gaussien
Séquence primaire [NCBI Protein]	k-spectrum
Interactions domaine-domaine [Pfam, DOMINE]	diffusion
Processus biologiques [Gene Ontology]	gaussien
Fonctions moléculaires [Gene Ontology]	gaussien
Composants cellulaires [Gene Ontology]	gaussien
Interologues [Inparanoid, DIP, MINT, BioGRID, Intact]	diffusion
Profil phylogénétique [BLASTP]	gaussien

Protocole

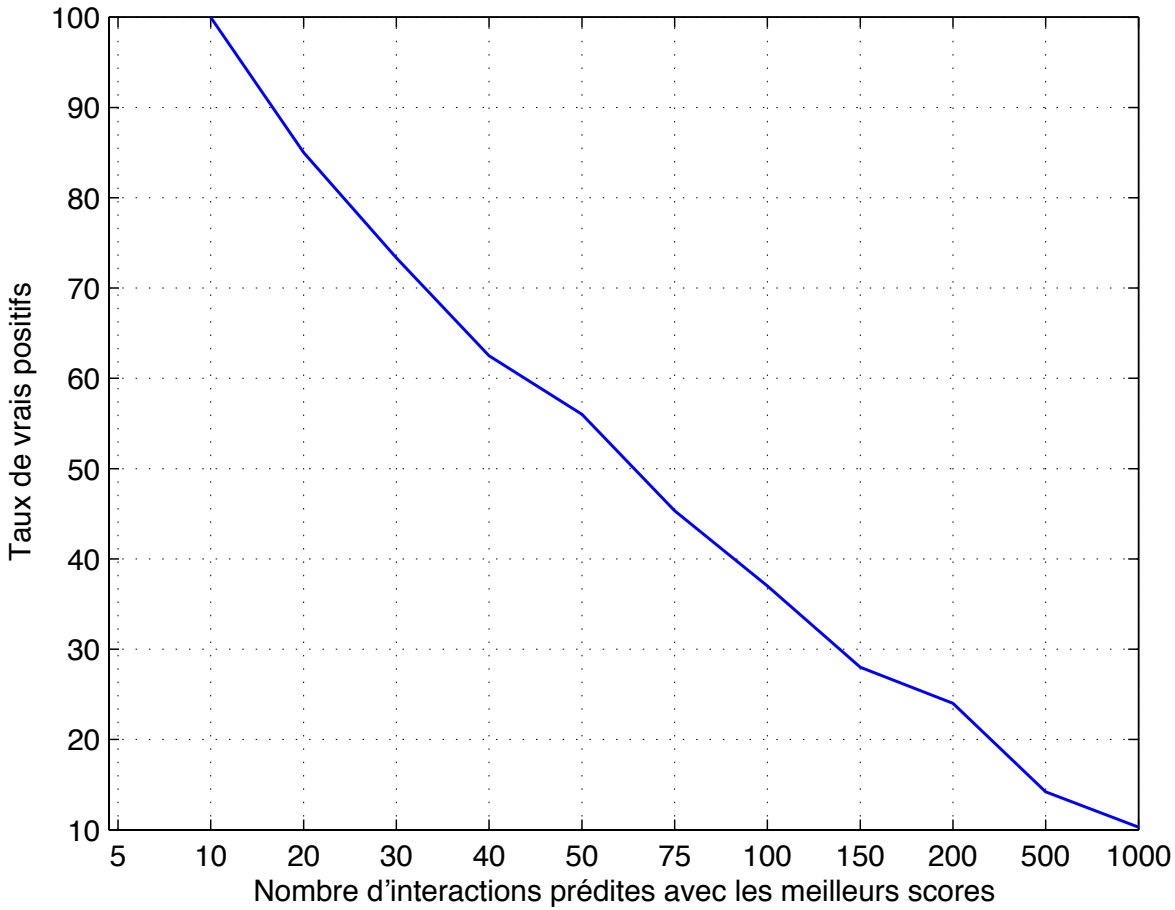
Prise en compte des annotations manquantes

- S : CFTR + protéines interagissant directement avec CFTR (34)
- T : ensemble des protéines interagissant avec les protéines de S (163)
- Prédiction des interactions entre les protéines de S et celles de T
- Plusieurs itérations :
  - A la i-ème itération, T est divisé uniformément en deux sous-ensembles $T_{1,i}$ et $T_{2,i}$
  - A la fin, combinaison des prédictions



Prédiction d'interactions connues

- Interactions prédites entre deux protéines u et u' ordonnées selon la valeur prise par $\hat{\kappa_y}(u, u')$
- Taux de vrais positifs obtenu pour les n premières prédictions :



Prédiction de nouvelles interactions

- Liste des interactions établie par une étude de la littérature pour les 100 premières prédictions obtenues

Prot 1	Prot 2	Méthode	Référence
XIAP	PTEN	étude enzymatique	Van Themsche C (2009)
NEDD4	PTEN	pull down	Wang X (2008)
NEDD4	PTEN	étude enzymatique	Wang X (2008)
SNAP23	VAMP2	pull down	Kawanishi M (2000)
SNAP23	VAMP1	double hybride	Ravichandran V (1996)
SNAP23	STX6	pull down	Martin-Martin B (2000)
SNAP23	STXBP2	in vivo	Schraw TD (2003)
DNAJC5	STUB1	affinity capture western	Schmidt BZ (2009)
ANXA5	ANXA1	co-localisation	Arur S (2003)

Sommaire

Conclusion

- Contribution à la problématique de la prédiction de PPI
- Introduction d'un nouveau cadre pour la régression à sorties structurées
- Extension de OKR au cas de l'apprentissage semi-supervisé
- Résultats théoriques pour les noyaux à valeur opérateur décomposables
- Définition de deux modèles : IOKR-ridge et IOKR-margin
- Illustration de l'approche par des résultats numériques

Perspectives

- **Validation expérimentale** : collaboration avec le LAMBE
- **Amélioration des performances** :
  - Critère de maximisation des AUC
  - Choix du noyau de sortie
  - Choix du noyau d'entrée : descriptions plus riches, choix d'un autre noyau à valeur opérateur
  - Utilisation de méthodes d'ensemble
- **Prédiction de fonction**
- **Apprentissage multitâches et apprentissage par transfert**