Du génome à l'interactome à l'aube des NGS de 3ème génération

Jérôme Azé

LRI, CNRS UMR 8623, Équipe Bioinfo
Équipe Projet INRIA AMIB

Outline

Contexte

Pourquoi séquence t-on l'ADN ?

- Extraire les informations encodées dans l'ADN (séquence de nucléotides A, C, G et T)
- Recenser de manière exhaustive les séquences codantes d'ADN (gènes)
- Comprendre les interactions entre ces gènes codant pour une ou plusieurs protéines

Contexte

Intérêts multiples

- Thérapeutique : mieux comprendre les interactions entres protéines $\longrightarrow$ pouvoir activer ou inhiber certaines interactions (drug design)
- Agroalimentaire : meilleur rendement de certaines cultures (reproduction contrôlée, maladie spécifique éradiquée)
- Environnemental : protéines dégradant la bio-masse, production de carburant vert, . . .

Comment séquence t-on l'ADN ?

Première génération : clonage (1977 –)

- Méthode Sanger : séquençage par synthèse

Deuxième génération : amplification PCR (2005 –)

- Méthode 454 : séquençage par synthèse
- Méthode Solexa : séquençage par synthèse
- Méthode SOLID : séquençage par ligation

Troisième génération : molécule unique (fin 2012 ? –)

- Méthode Pacific Bioscience : séquençage par synthèse : molécule unique
- Méthode ion torrent : séquençage par synthèse
- Méthode nanopore : séquençage molécule unique

Comment séquence t-on l'ADN ?

Première génération : clonage (1977 –)

- Méthode Sanger : séquençage par synthèse

Deuxième génération : amplification PCR (2005 –)

- Méthode 454 : séquençage par synthèse
- Méthode Solexa : séquençage par synthèse
- Méthode SOLID : séquençage par ligation

Troisième génération : molécule unique (fin 2012 ? –)

- Méthode Pacific Bioscience : séquençage par synthèse : molécule unique
- Méthode ion torrent : séquençage par synthèse
- Méthode nanopore : séquençage molécule unique

Comment séquence t-on l'ADN ?

Première génération : clonage (1977 –)

- Méthode Sanger : séquençage par synthèse

Deuxième génération : amplification PCR (2005 –)

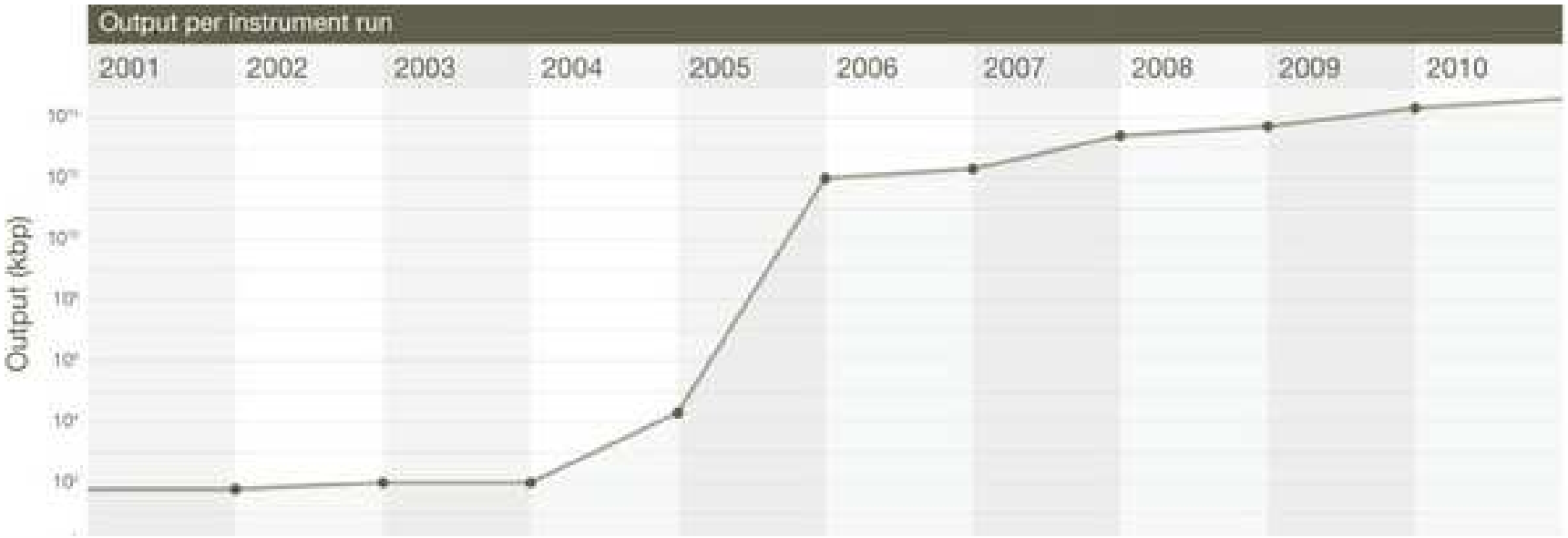
- Méthode 454 : séquençage par synthèse
- Méthode Solexa : séquençage par synthèse
- Méthode SOLID : séquençage par ligation

Troisième génération : molécule unique (fin 2012 ? –)

- Méthode Pacific Bioscience : séquençage par synthèse : molécule unique
- Méthode ion torrent : séquençage par synthèse
- Méthode nanopore : séquençage molécule unique

Quelques chiffres

L'évolution des performances des séquençeurs de 1ère et 2ème génération



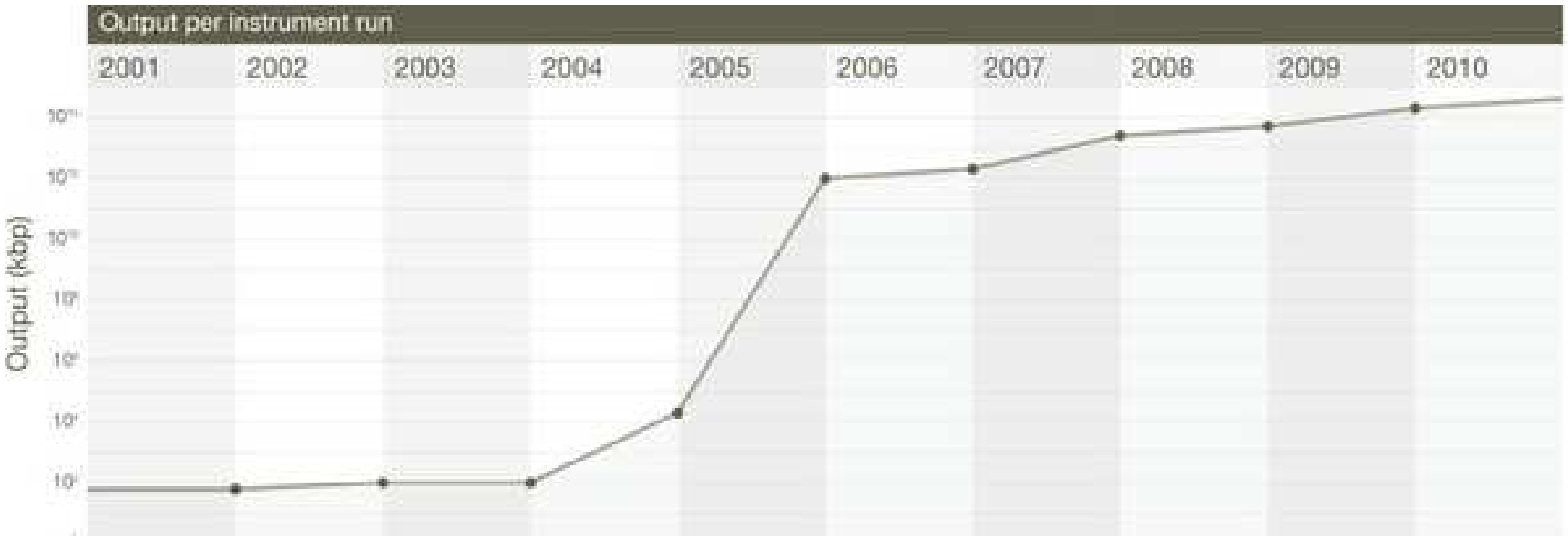
Mardis ER, A decade's perspective on DNA sequencing technology, Nature, 2011

Pour une même somme

- La quantité de séquençage double tous les 5 mois
- La quantité de RAM double tous les 14 mois.

Quelques chiffres

L'évolution des performances des séquençeurs de 1ère et 2ème génération



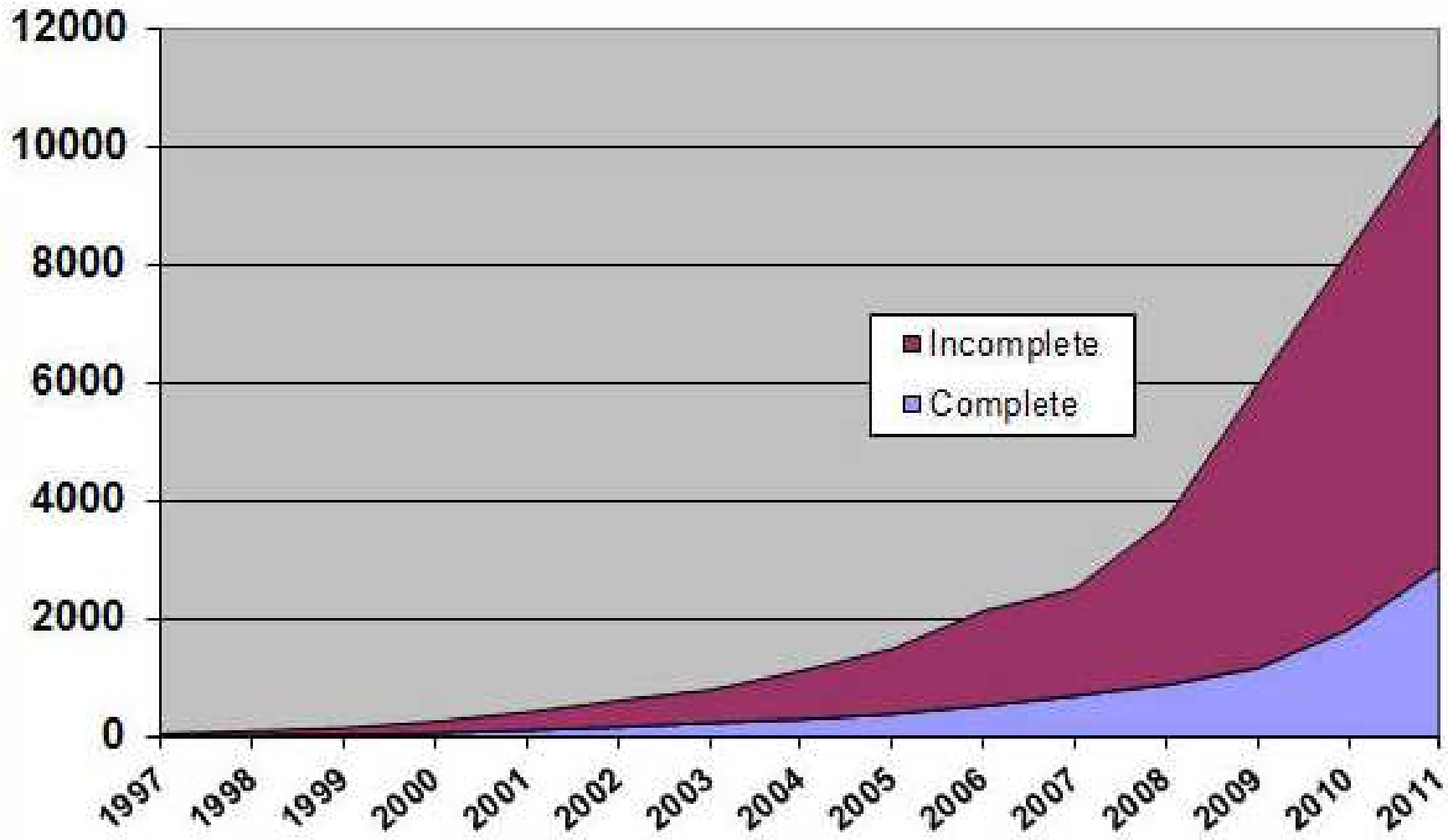
Mardis ER, A decade's perspective on DNA sequencing technology, Nature, 2011

Pour une même somme

- La quantité de séquençage double tous les 5 mois
- La quantité de RAM double tous les 14 mois.

Quelques chiffres

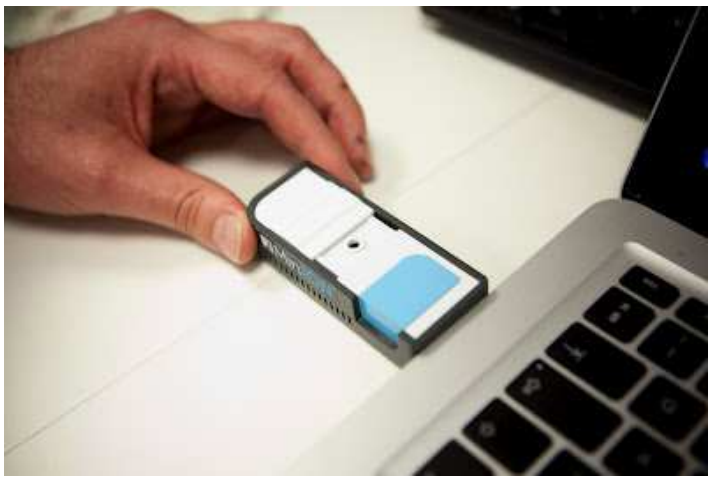
Génomes recensés sur GOLD (octobre 2011)



Troisième génération : “single molecule”

Oxford Nanopore : MinION, GridION, fin 2012 - début 2013

- Annoncé lors de l’**A**dvances in **G**enome **B**iology and **T**echnology (15 au 18 février 2012, USA)
- MinION : séquençeur jetable sur port USB, $\sim 1\,Gb$, coût $< 1000\$$.



Troisième génération : “single molecule”

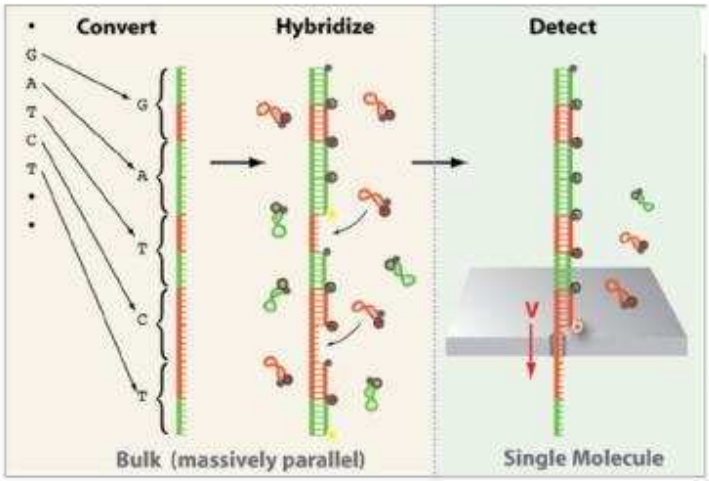
Oxford Nanopore : MinION, GridION, fin 2012 - début 2013

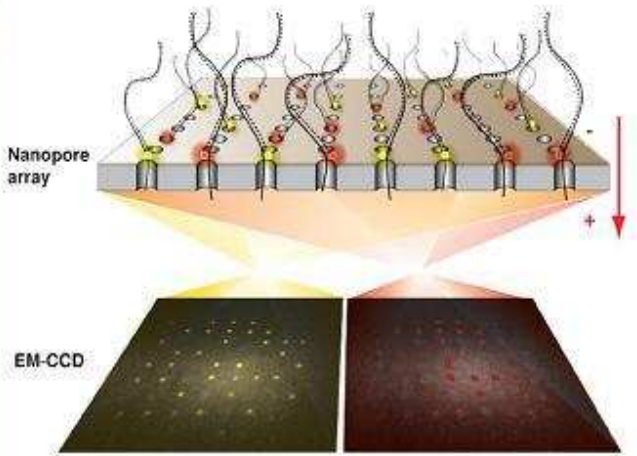
- GridION :
  - Plus de $10 Gb/jour/module$ (a priori, $\sim 25\,Gb$)
  - Environ $30\,k\$/module$;
  - 20 modules GridION $\rightarrow$ un génome humain / 15 minutes.

Troisième génération : “single molecule”

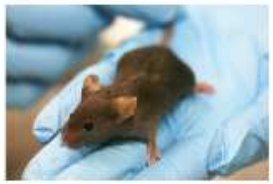
Noblegen Biosciences, 2014 ?

- Technologie “optipore” (optical detection + nanopore)
- Taille des fragments $<$ 200 bases
- Capacité de séquençage annoncée : 500 Gb/heure

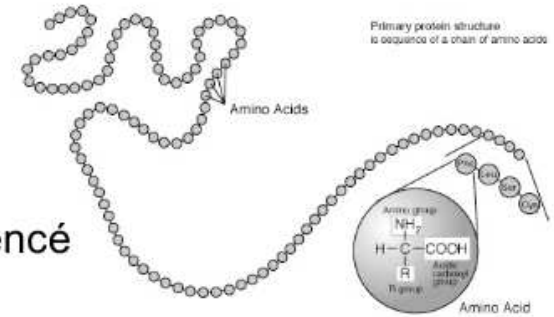




Quelques finalités du séquençage



Génome séquencé

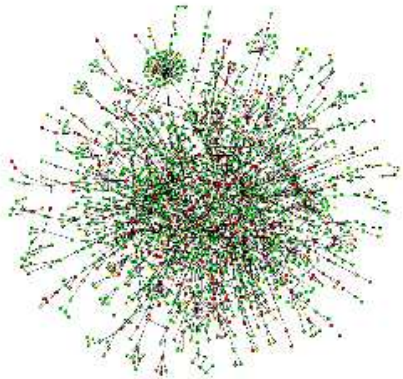


Annotation des protéines

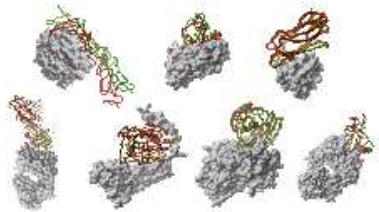
EMBL-EBI

SIB
Swiss Institute of Bioinformatics

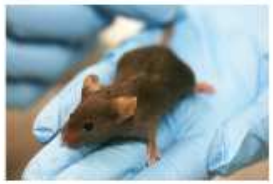
Interactions protéine-protéine



Structure 3D des protéines

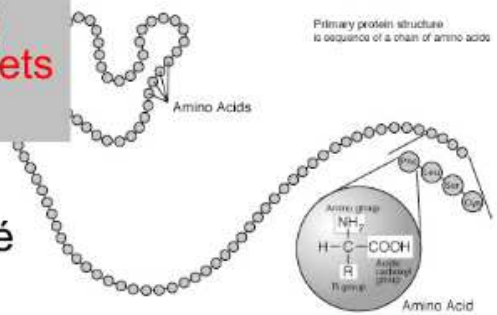


Quelques finalités du séquençage



3 173 génomes complets
10 479 génomes incomplets
1 918 projets en cours

Génome séquencé

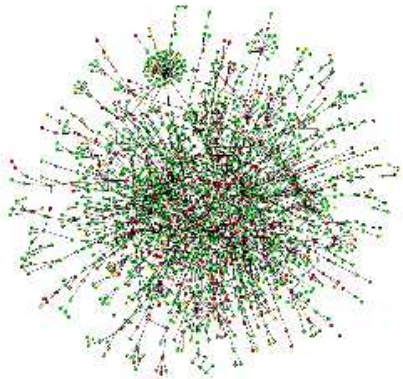


Annotation des protéines

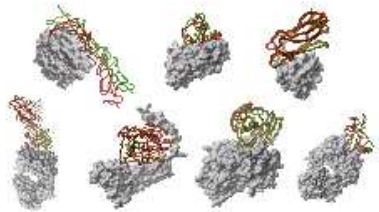
EMBL-EBI

SIB
Swiss Institute of Bioinformatics

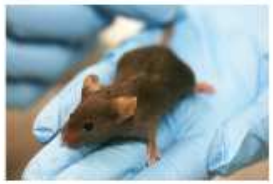
Interactions protéine-protéine



Structure 3D des protéines

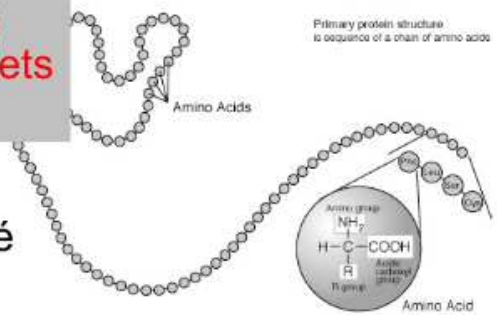


Quelques finalités du séquençage



3 173 génomes complets
10 479 génomes incomplets
1 918 projets en cours

Génome séquencé



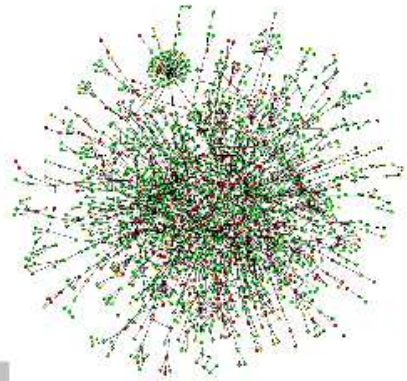
Annotation des protéines

EMBL-EBI

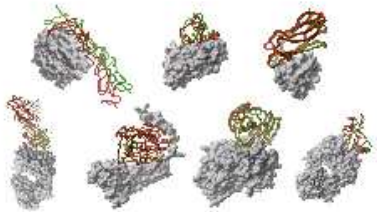
SIB
Swiss Institute of Bioinformatics

535 248 protéines annotées

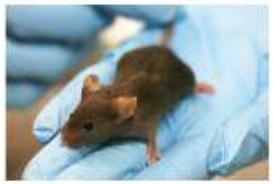
Interactions protéine-protéine



Structure 3D des protéines

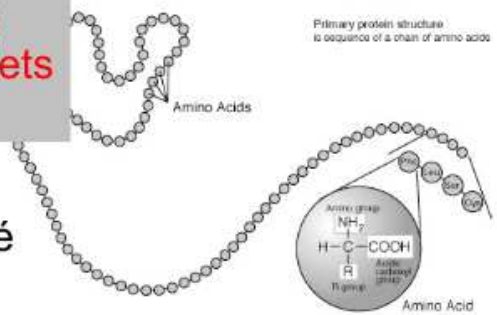


Quelques finalités du séquençage



3 173 génomes complets
10 479 génomes incomplets
1 918 projets en cours

Génome séquencé



Annotation des protéines

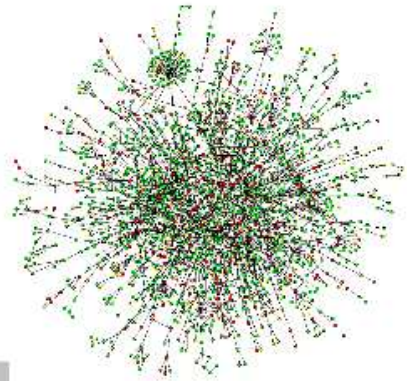
EMBL-EBI

20 639 311 protéines

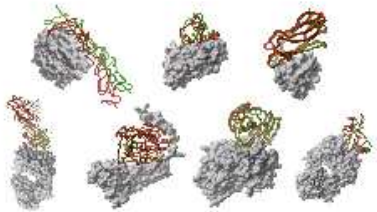
SIB
Swiss Institute of Bioinformatics

535 248 protéines annotées

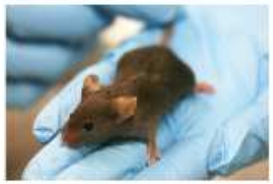
Interactions protéine-protéine



Structure 3D des protéines

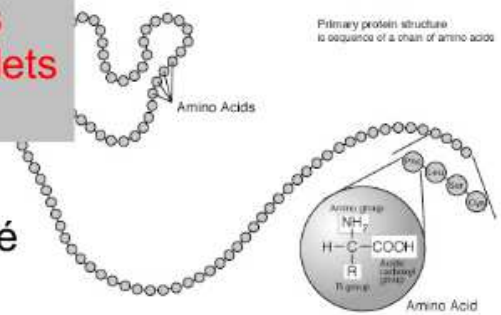


Quelques finalités du séquençage



3 173 génomes complets
10 479 génomes incomplets
1 918 projets en cours

Génome séquencé



Annotation des protéines

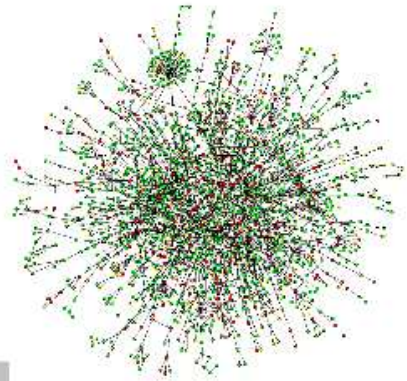
EMBL-EBI

20 639 311 protéines

SIB
Swiss Institute of Bioinformatics

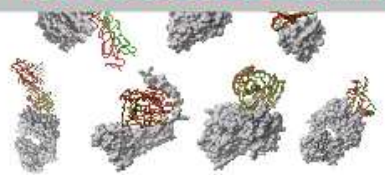
535 248 protéines annotées

Interactions protéine-protéine

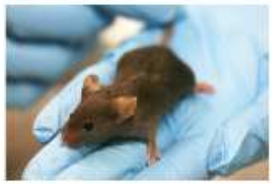


Structure 3D des protéines

74 460 structures connues

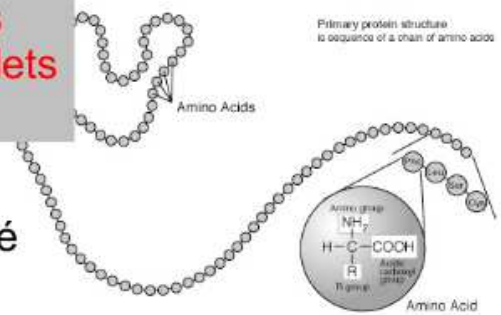


Quelques finalités du séquençage



3 173 génomes complets
10 479 génomes incomplets
1 918 projets en cours

Génome séquencé



Annotation des protéines

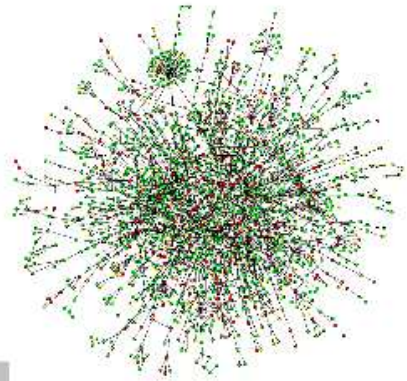
EMBL-EBI

20 639 311 protéines

SIB
Swiss Institute of Bioinformatics

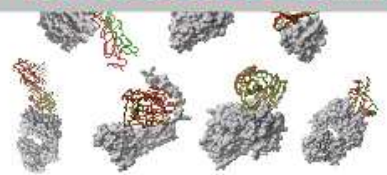
535 248 protéines annotées

Interactions protéine-protéine



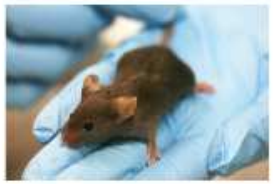
Structure 3D des protéines

74 460 structures connues



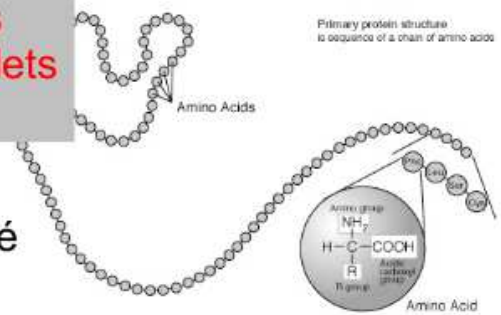
Dont 3 569 complexes

Quelques finalités du séquençage



3 173 génomes complets
10 479 génomes incomplets
1 918 projets en cours

Génome séquencé



Annotation des protéines

EMBL-EBI

20 639 311 protéines

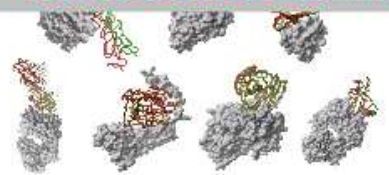
SIB
Swiss Institute of Bioinformatics

535 248 protéines annotées

Interactions protéine-protéine

Structure 3D des protéines

74 460 structures connues



Dont 3 569 complexes

Outline

Prédiction d'interactions protéine-protéine

Principaux objectifs

Pouvoir, à partir d'un génome nouvellement séquençé ou à partir d'un génome déjà connu

1. Prédire les interactions protéine-protéine
2. Élaguer l'immense ensemble de paires de protéines possiblement en interaction sans perdre de vraies interactions
3. Découvrir de nouvelles interactions protéine-protéine

Prédiction d'interactions protéine-protéine

Approche proposée (D. de Vienne et J. Azé)

- Hypothèse forte : les protéines en interaction co-évoluent dans différentes espèces
- Utilisation de descripteurs "simples" et rapides à calculer pour chaque paire de protéines
- Utiliser des approches de "Machine Learning" pour apprendre à ordonner les paires de protéines selon une "probabilité" décroissante d'être en interaction

Étude de cas

- Prédiction de l'interactome d'*Escherichia coli*
- Seules les paires de protéines présentes dans des complexes avérés ont été considérées comme étant en interaction

Prédiction d'interactions protéine-protéine

Approche proposée (D. de Vienne et J. Azé)

- Hypothèse forte : les protéines en interaction co-évoluent dans différentes espèces
- Utilisation de descripteurs "simples" et rapides à calculer pour chaque paire de protéines
- Utiliser des approches de "Machine Learning" pour apprendre à ordonner les paires de protéines selon une "probabilité" décroissante d'être en interaction

Étude de cas

- Prédiction de l'interactome d'*Escherichia coli*
- Seules les paires de protéines présentes dans des complexes avérés ont été considérées comme étant en interaction

Vue globale de l'approche

Données

- Un ensemble de protéines
- Un ensemble de génomes completement séquencés et l'**arbre phylogénétique** associé
- Un groupe de séquences de protéines orthologues

Pour l'apprentissage

- Un ensemble d'interactions protéine-protéine **positives**
- Un ensemble d'interactions protéine-protéine **négatives**

Pour la prédiction

- L'ensemble des interactions protéine-protéine potentielles

Vue globale de l'approche

Données

- Un ensemble de protéines
- Un ensemble de génomes completement séquencés et l'**arbre phylogénétique** associé
- Un groupe de séquences de protéines orthologues

Pour l'apprentissage

- Un ensemble d'interactions protéine-protéine **positives**
- Un ensemble d'interactions protéine-protéine **négatives**

Pour la prédiction

- L'ensemble des interactions protéine-protéine potentielles

Vue globale de l'approche

Données

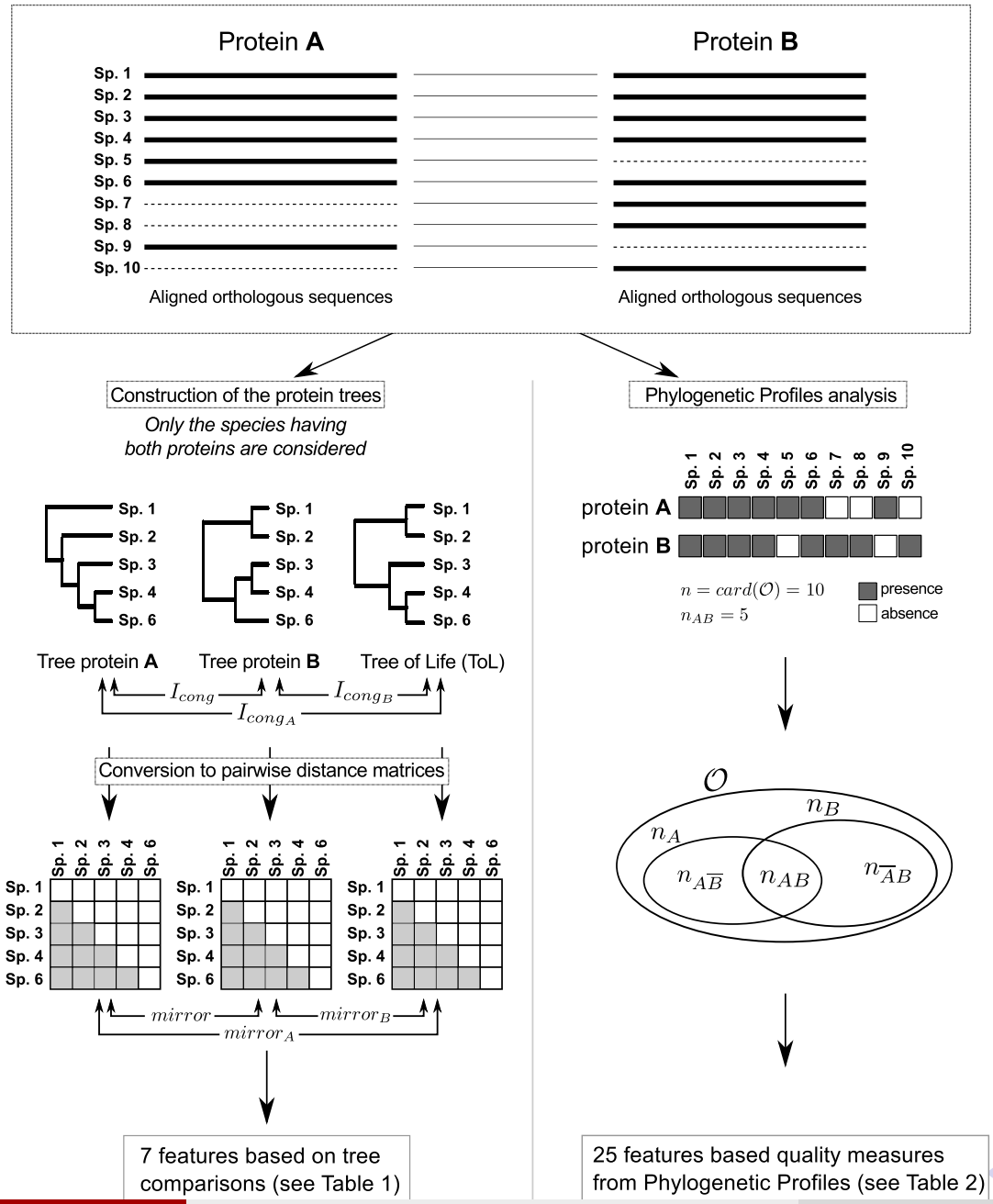
- Un ensemble de protéines
- Un ensemble de génomes completement séquencés et l'**arbre phylogénétique** associé
- Un groupe de séquences de protéines orthologues

Pour l'apprentissage

- Un ensemble d'interactions protéine-protéine **positives**
- Un ensemble d'interactions protéine-protéine **négatives**

Pour la prédiction

- L'ensemble des interactions protéine-protéine potentielles


Protein A
Protein B
Sp. 1
Sp. 2
Sp. 3
Sp. 4
Sp. 5
Sp. 6
Sp. 7
Sp. 8
Sp. 9
Sp. 10
Aligned orthologous sequences
Aligned orthologous sequences
Construction of the protein trees
Only the species having both proteins are considered
Phylogenetic Profiles analysis
Tree protein A
Tree protein B
Tree of Life (ToL)
I_{cong}
I_{cong_B}
I_{cong_A}
Conversion to pairwise distance matrices
$mirror$
$mirror_B$
$mirror_A$
protein A
protein B
$n = card(\mathcal{O}) = 10$
$n_{AB} = 5$
presence
absence
$\mathcal{O}$
n_A
n_B
$n_{A\overline{B}}$
n_{AB}
$n_{\overline{A}B}$
7 features based on tree comparisons (see Table 1)
25 features based quality measures from Phylogenetic Profiles (see Table 2)

Plan

Données étudiées

Données fournies par *Juan et al. (PNAS, 2008)*

- 2 177 séquences de protéines issues du génome de *E. coli*
- Un ensemble de 115 génomes prokariotes entièrement séquencés
- Pour chaque protéine, l'ensemble des séquences orthologues dans les autres génomes
- Pour des considérations techniques, seules les protéines ayant au moins 7 séquences orthologues ont été conservées
- De manière similaire, seules les paires de protéines partageant au moins 7 séquences orthologues ont été conservées

Description des données

Données fournies par *Juan et al. (PNAS, 2008)*

- 2 177 séquences de protéines issues du génome de *E. coli*
- 628 interactions protéine-protéine sont considérées comme des interactions **positives** (protéines présentes dans des complexes)
- 76 202 interactions protéine-protéine considérées comme **négatives** (non présentes dans des complexes)

Descripteurs utilisés pour décrire et prédire les interactions protéine-protéine

Tous les descripteurs sont liés à la notion de co-évolution.

- Comparaison des arbres phylogénétiques (7 descripteurs)
- Comparaison des profils phylogénétiques (28 descripteurs)

Plan

Méthodes

Approches classiques

- Ordonner les paires de protéines selon un ou plusieurs critères
- Utiliser un seuil prédéfini pour prédire les paires en interaction
- Meilleure approche publiée : **context-mirror** by *Juan et al. (PNAS 2008)*

Notre approche

- Apprentissage de plusieurs classifieurs via une boîte à outils (Weka)
- Combinaison des différents rangs induits par ces modèles pour obtenir un score unique
- Tri des paires de protéines selon ce score

Méthodes

Approches classiques

- Ordonner les paires de protéines selon un ou plusieurs critères
- Utiliser un seuil prédéfini pour prédire les paires en interaction
- Meilleure approche publiée : **context-mirror** by *Juan et al. (PNAS 2008)*

Notre approche

- Apprentissage de plusieurs classifieurs via une boîte à outils (Weka)
- Combinaison des différents rangs induits par ces modèles pour obtenir un score unique
- Tri des paires de protéines selon ce score

Combinaison des prédictions

Détails du calcul

$$S(pair) = S_{pos}(pair)/S_{neg}(pair)$$

$$S_{pos}(pair) = \begin{cases} P_{pos} \times e^{n_{pos}} & \text{if } n_{pos} > 0 \\ 1 & \text{otherwise} \end{cases}$$

$$S_{neg}(pair) = \begin{cases} P_{neg} \times e^{n_{neg}} & \text{if } n_{neg} > 0 \\ 1 & \text{otherwise} \end{cases}$$

$$\text{with} \begin{cases} n_{pos} = \sum_{c \in Classifiers} \mathbb{1}_{P_c(pair) \geq 0.5} \\ P_{pos} = \prod_{c \in Classifiers} P_c(pair) \times \mathbb{1}_{P_c(pair) \geq 0.5} \end{cases}$$

$$\text{and} \begin{cases} n_{neg} = \sum_{c \in Classifiers} \mathbb{1}_{P_c(pair) < 0.5} \\ P_{neg} = \prod_{c \in Classifiers} (1 - P_c(pair)) \times \mathbb{1}_{P_c(pair) < 0.5} \end{cases}$$

Cadre d'évaluation

Cadre d'évaluation

- 3 folds Cross-Validation
- 30 itérations
- Les évaluations ont été réalisées selon les critères suivants :
  - Courbe ROC et aire sous la courbe ROC (ROC-AUC)
  - Courbes de précision et de rappel

Plan

Critère de qualité : Aire sous la courbe ROC

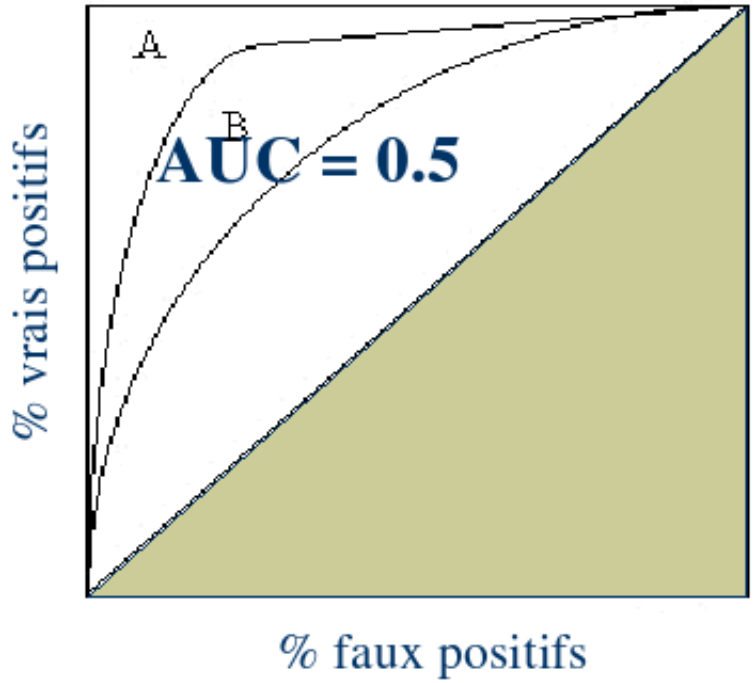
Évaluation d'un test : compromis entre

- % de vrais positifs
- % de faux positifs (1 - % de vrais négatifs)

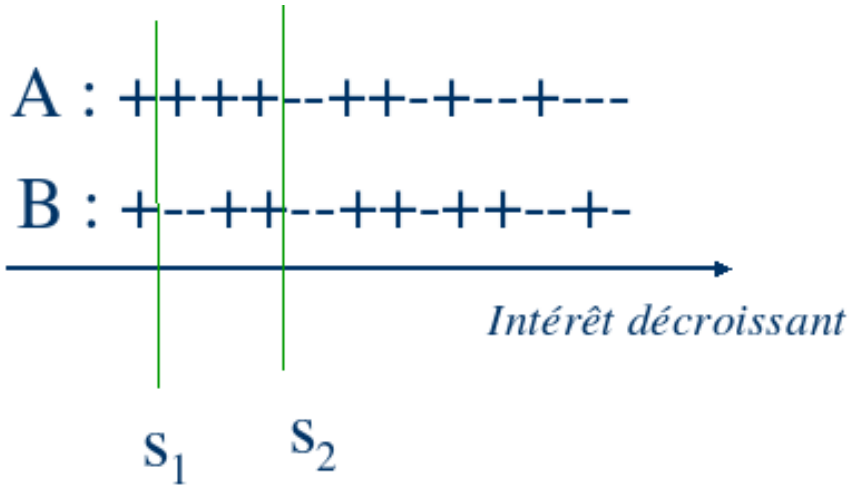
Critère plus fiable que la précision

Ling, Huang, Zhang, AI'03

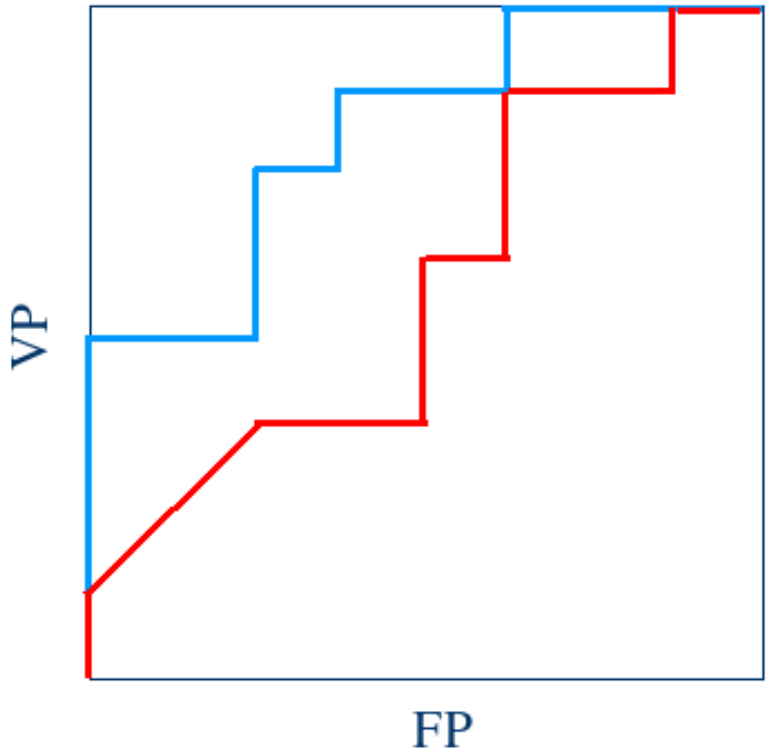
- Insensible à la distribution
- Indépendant des coûts d'erreur



Comparaison de deux ordres



	S_1	S_2
A	VP = 1/8 FP = 0	VP = 4/8 FP = 0
B	VP = 1/8 FP = 0	VP = 3/8 FP = 2/8



Comparaison de deux ordres

Critère réellement optimisé

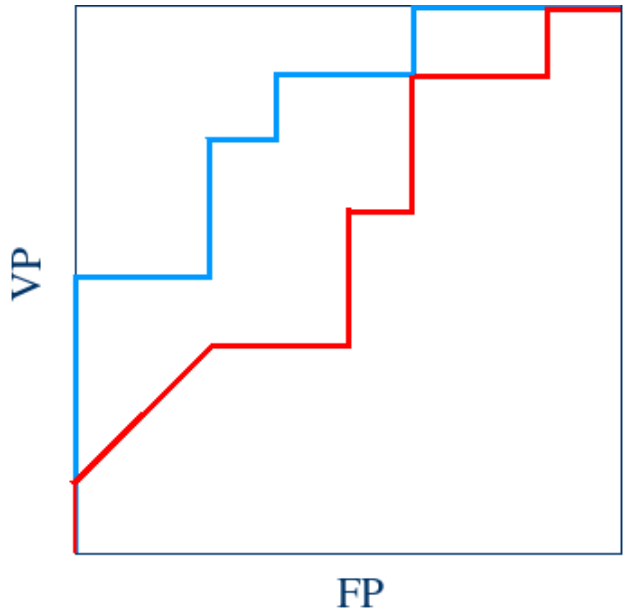
Maximiser l'aire sous la courbe ROC $\Longleftrightarrow$ minimiser la somme des rangs des ex. positifs

Rangs (+, A) : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Rangs (+, B) : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

	$\sum$ Rangs (+,.)
A : ++++--++-+--+---	48
B : +--++--++-++--+-	65

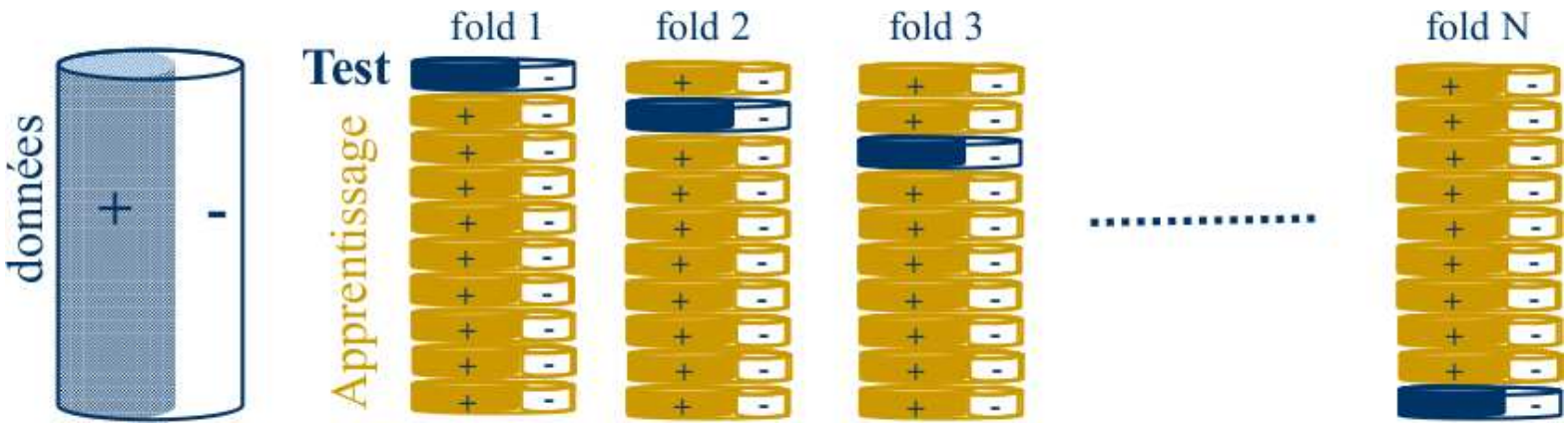
Intérêt décroissant



Principe de la validation croisée

N-Validation Croisée

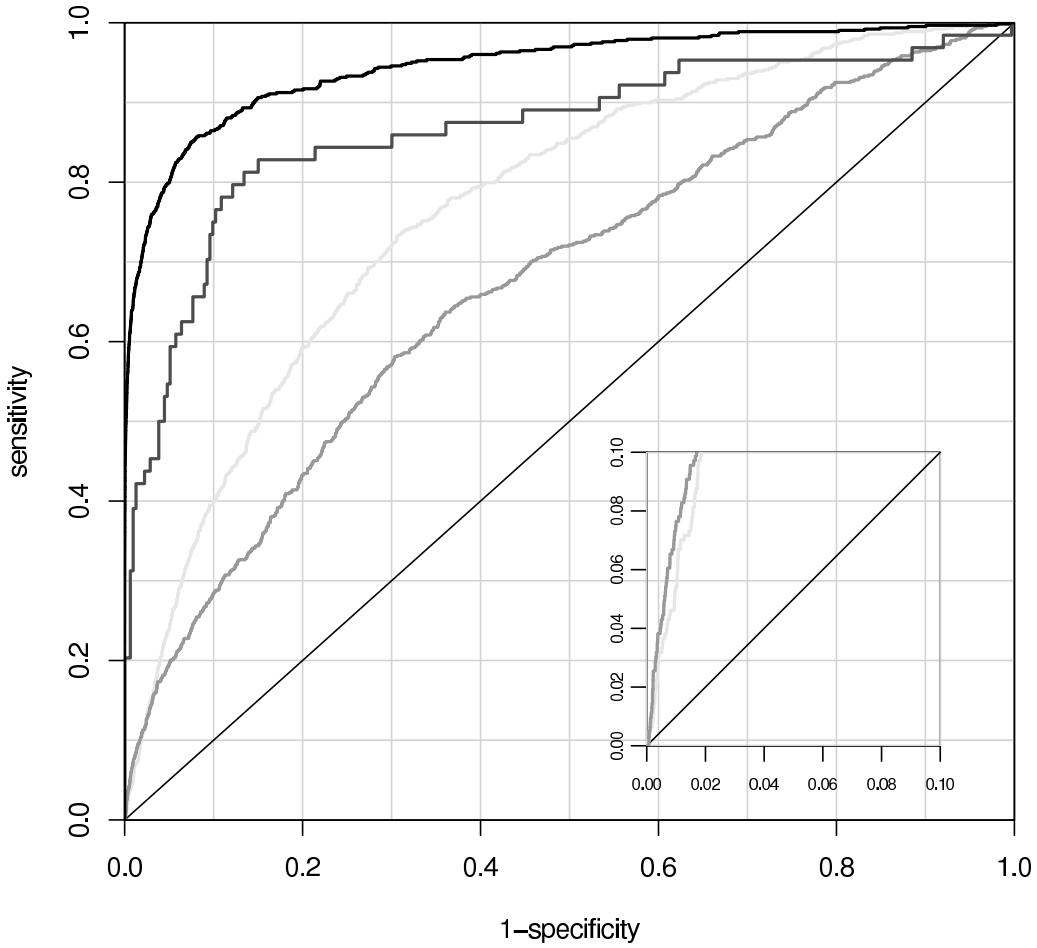
- Séparer les données en N ensembles disjoints (partition)
- Garantir que la distribution des classes dans les N ensembles est comparable à la distribution initiale (stratification)
- Utiliser $N - 1$ ensembles pour apprendre et le complémentaire pour évaluer
- Répéter le processus “apprentissage / test” N fois.



Plan

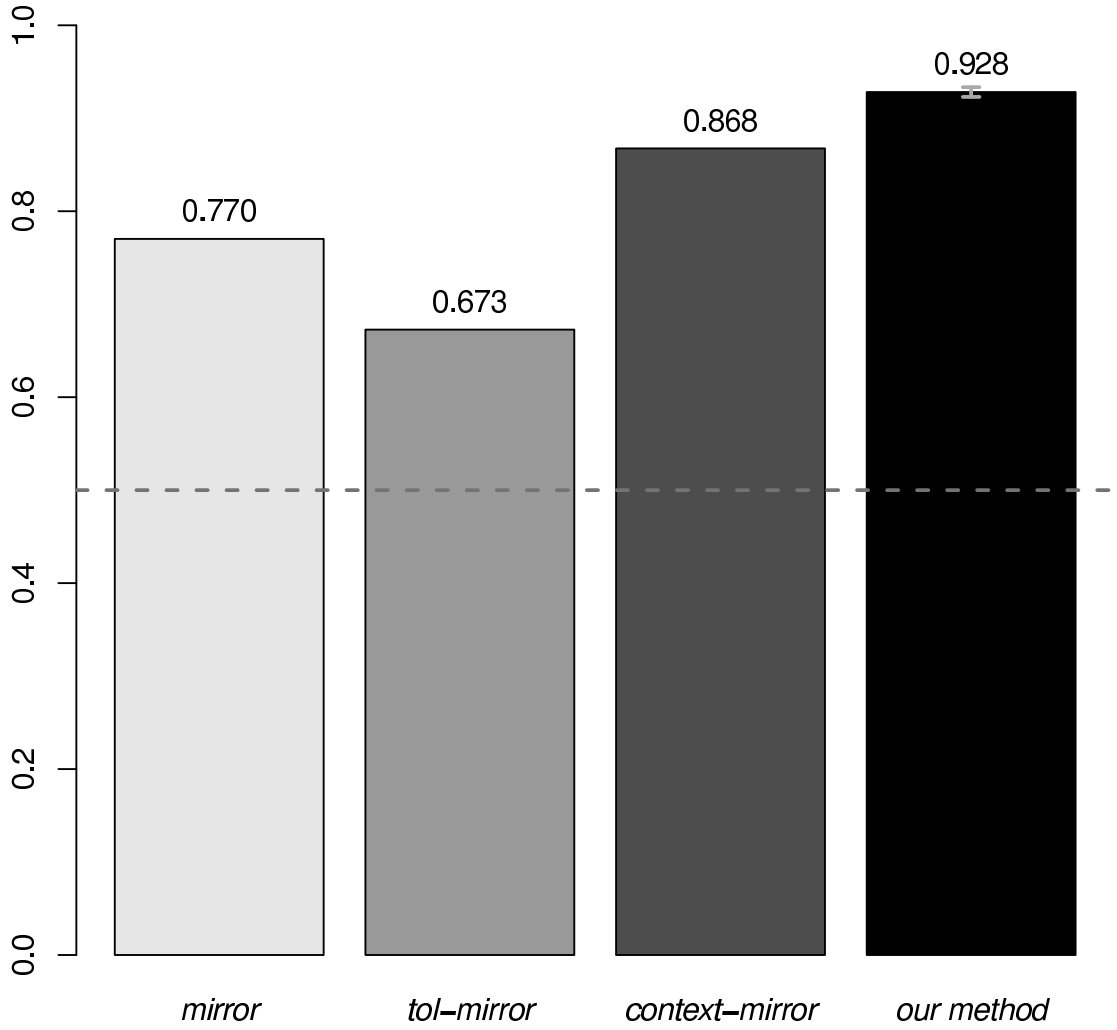
Comparaison avec les approches de référence

Comparaison des courbes ROC



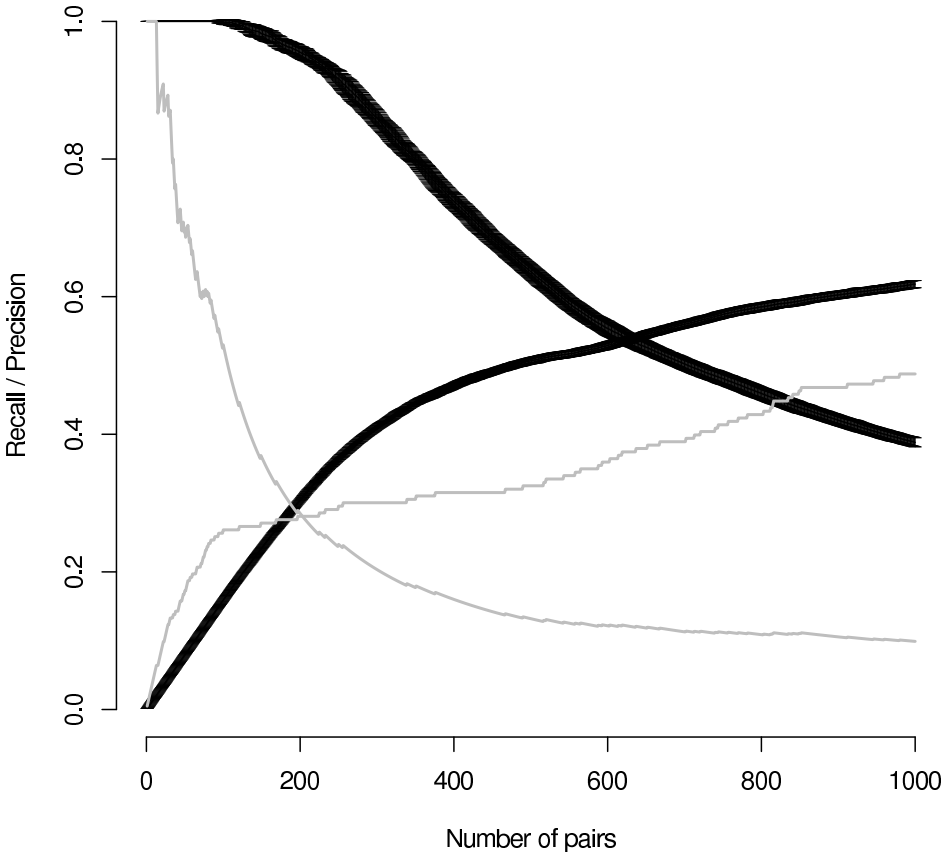
Comparaison avec les approches de référence

Comparaison des ROC-AUC



Comparaison avec les approches de référence

Comparaison des courbes Précision et Rappel avec l'approche "context-mirror"



Précision = 100%

	nb paires	Rappel
context-mirror	13	2%
notre approche	90	14.6%

Rappel = 50%

	nb paires	Précision
context-mirror	$> 1\,000$	$\sim 10\%$
notre approche	~ 450	$\sim 62\%$

Impact des différents descripteurs utilisés pour apprendre les modèles prédéctifs

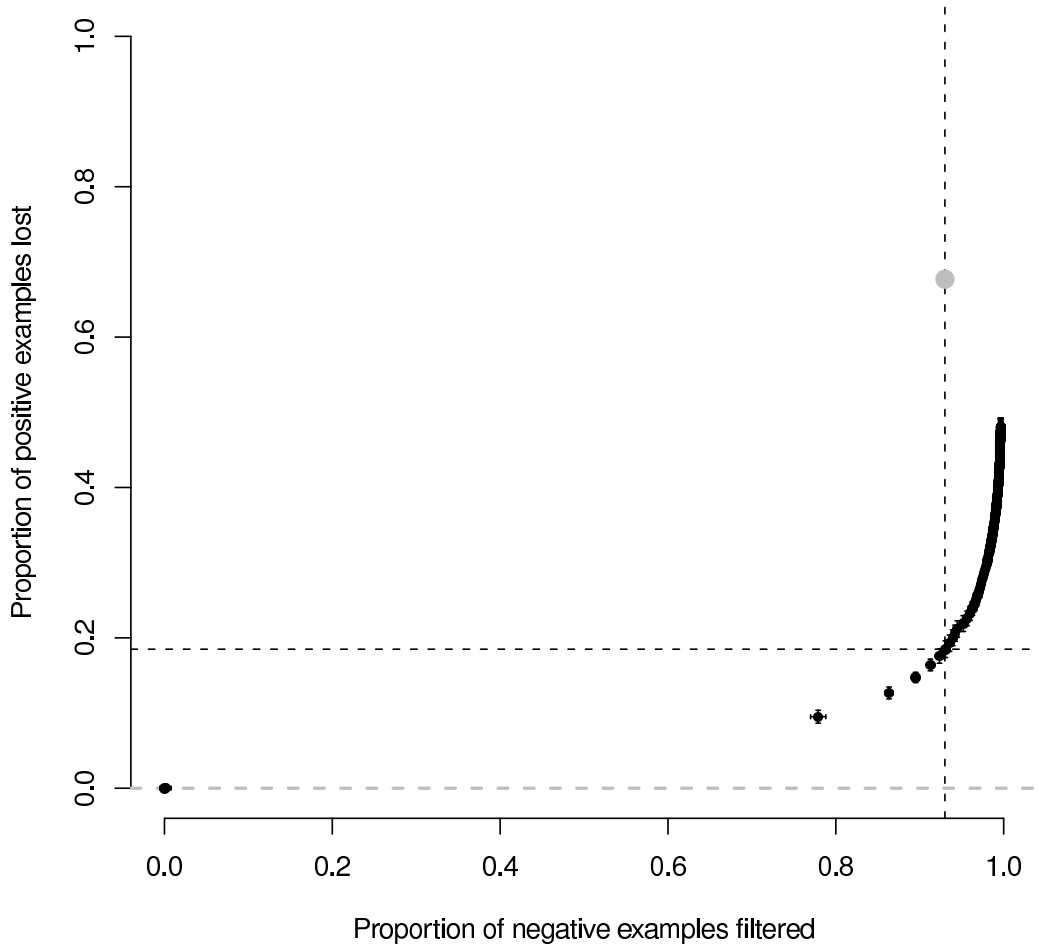
Efficacité de la prédiction pour les différents ensemble de descripteurs

	Ensemble de descripteurs				
	topology[1]	*matrix*[2]	*tree*[3]	*PP*[4]	***ALL***[5]
AUC (mean)	0.78	0.80	0.84	0.92	0.93
AUC (sd)	0.0054	0.0045	0.0039	0.0036	0.0028

[1] Features included : I_{cong}, $I_{cong}A$ and $I_{cong}B$
[2] Features included : *mirror*, $mirror_A$, $mirror_B$ and $tol - mirror$
[3] Features included : I_{cong}, $I_{cong}A$, $I_{cong}B$, *mirror*, $mirror_A$, $mirror_B$ and $tol - mirror$
[4] Features included : all Phylogenetic Profile features
[5] Features included : all features

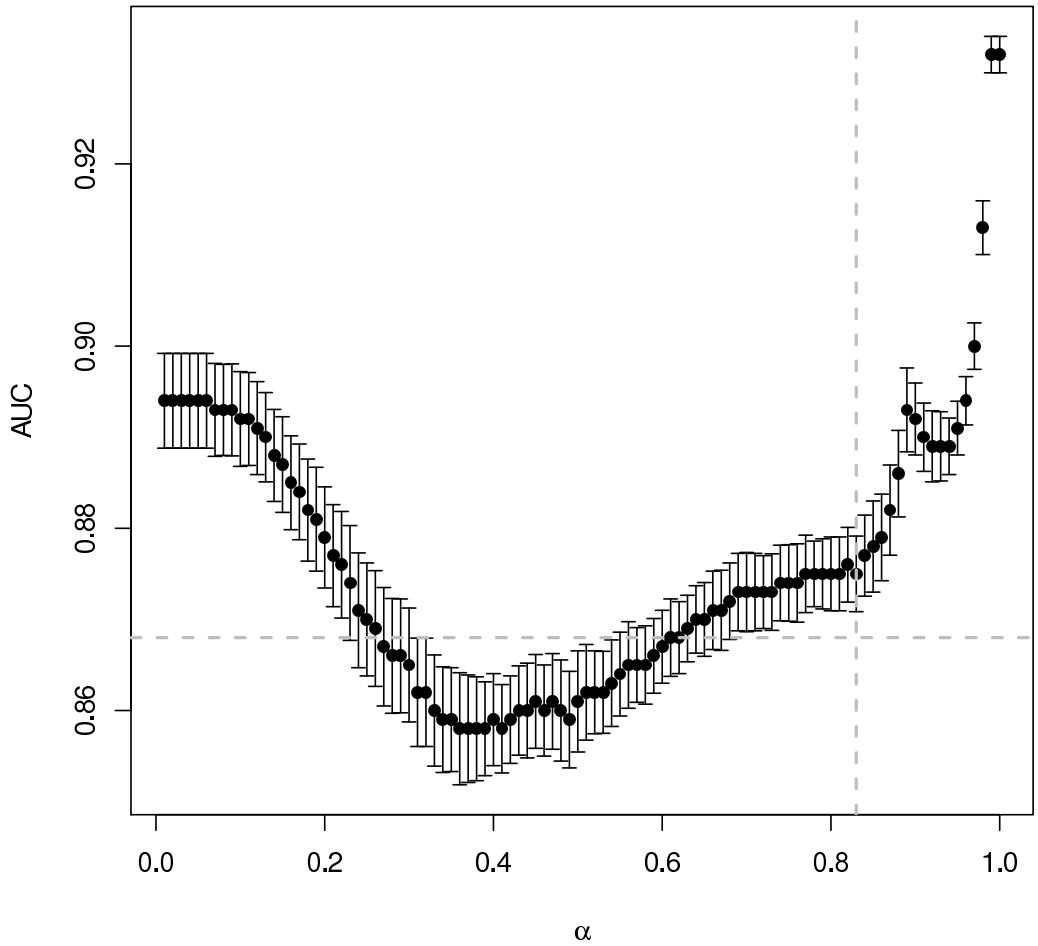
Impact sur l'élagage

Effet de l'élagage des paires prédites négatives sur la proportion de paires positives perdues



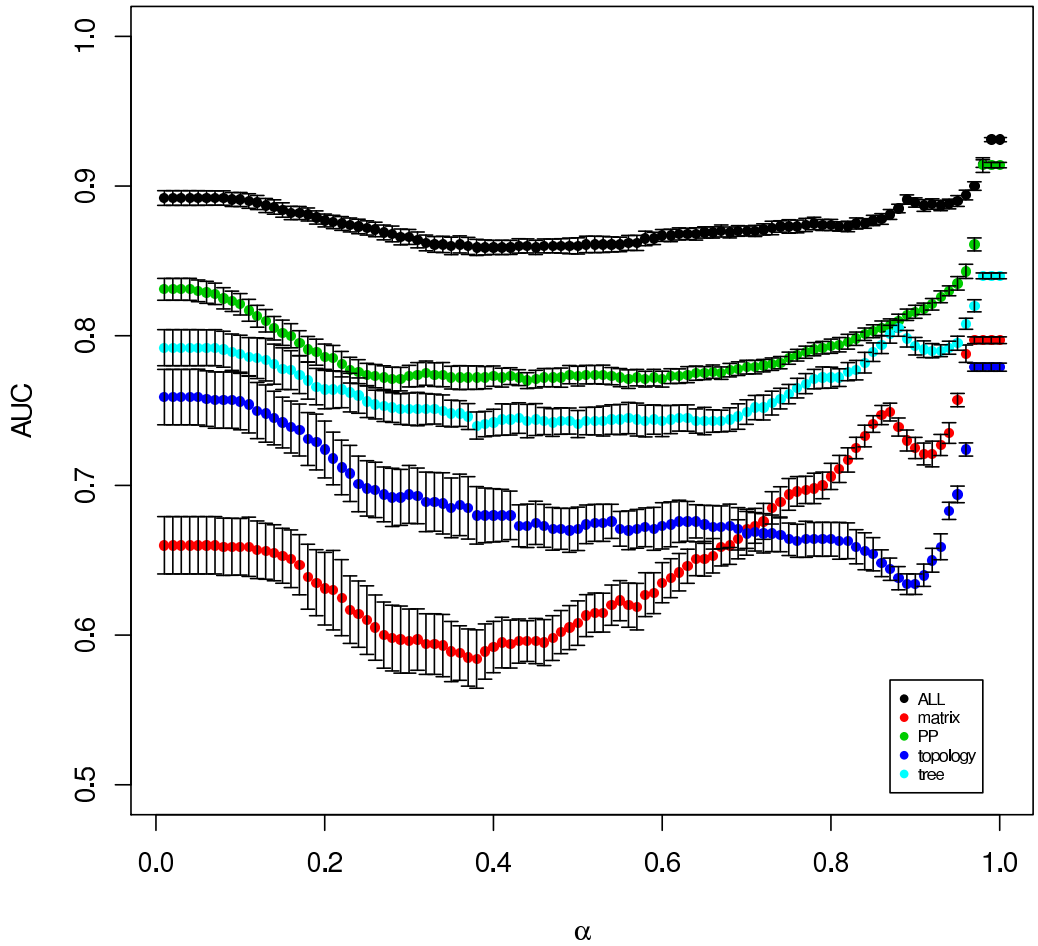
Impact sur l'élagage

Effet de l'élagage des paires prédites négatives sur la proportion de paires positives perdues



Impact sur l'élagage

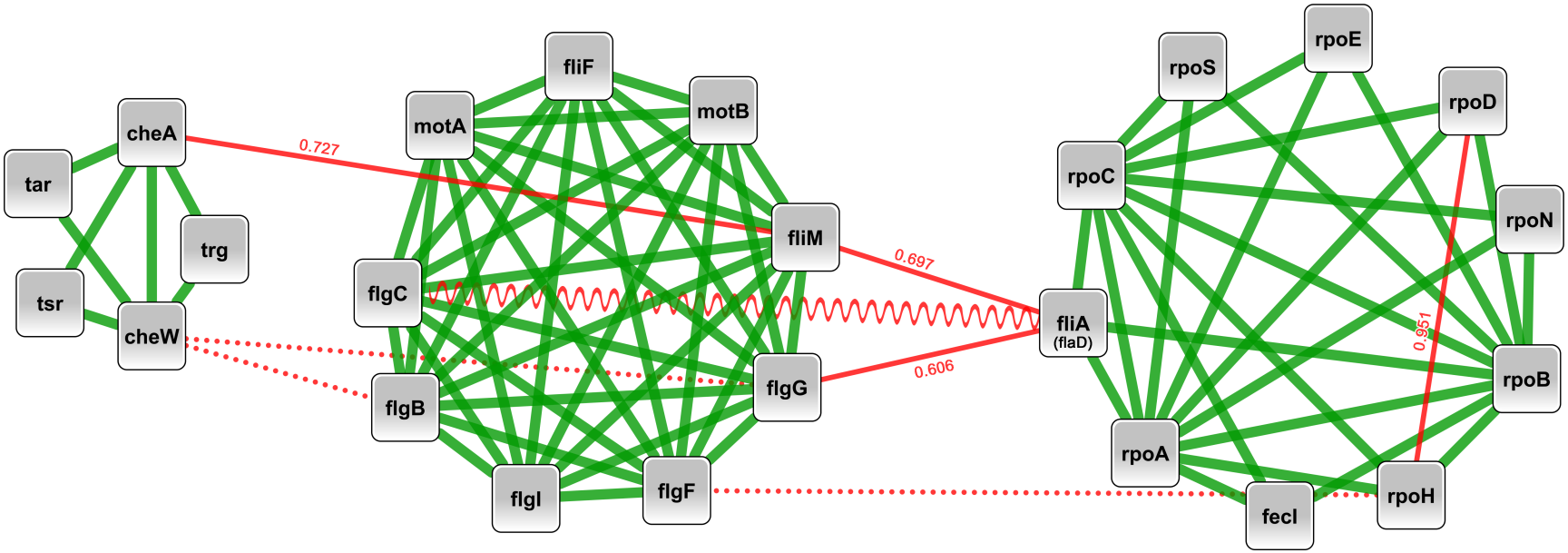
Effet de l'élagage des paires prédites négatives sur la proportion de paires positives perdues



Découverte de nouvelles interactions protéine-protéine

Focus sur les 50 premières paires négatives

- Pour 23 paires parmi les 50, la base de données STRING contient des preuves d'interactions possibles
- Focus sur un sous-ensemble de protéines pour y découvrir de nouvelles interactions



Performances et perspectives

Temps de calcul

- Calcul des descripteurs : 3*h*(*PP*) et 20*h* (*tree*)
- Apprentissage des modèles : 3*h*
- Prédiction, tri, élagage : quelques minutes

Validation à plus large échelle

- Validation expérimentale (non triviale)
- Étude d'autres organismes : 48 champignons (eukaryotes)
- Validation via la prédiction de complexes protéiques connus

Performances et perspectives

Temps de calcul

- Calcul des descripteurs : 3*h*(*PP*) et 20*h* (*tree*)
- Apprentissage des modèles : 3*h*
- Prédiction, tri, élagage : quelques minutes

Validation à plus large échelle

- Validation expérimentale (non triviale)
- Étude d'autres organismes : 48 champignons (eukaroytes)
- Validation via la prédiction de complexes protéine-protéine

Performances et perspectives

Temps de calcul

- Calcul des descripteurs : 3h(PP) et 20h ($tree$)
- Apprentissage des modèles : 3h
- Prédiction, tri, élagage : quelques minutes

Validation à plus large échelle

- Validation expérimentale (non triviale)
- Étude d'autres organismes : 48 champignons (eukaroytes)
- Validation via la prédiction de complexes protéine-protéine

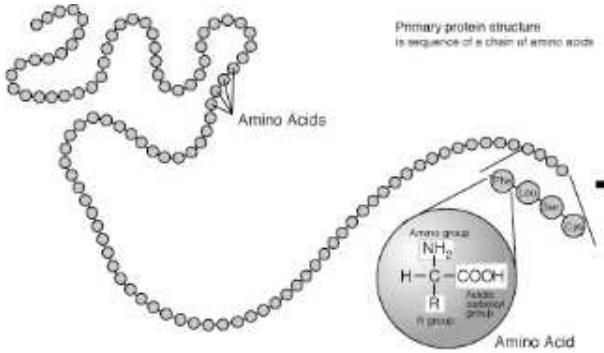
Performances et perspectives

Temps de calcul

- Calcul des descripteurs : 3h(PP) et 20h ($tree$)
- Apprentissage des modèles : 3h
- Prédiction, tri, élagage : quelques minutes

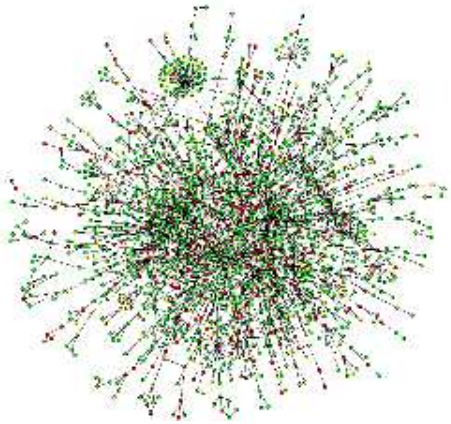
Validation à plus large échelle

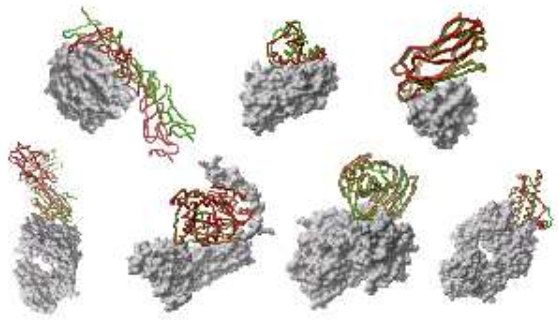
- Validation expérimentale (non triviale)
- Étude d'autres organismes : 48 champignons (eukaroytes)
- Validation via la prédiction de complexes protéine-protéine



Interactions protéine-protéine

Structure 3D des protéines





Outline

Complexes protéine - protéine

Pourquoi modéliser les complexes protéine - protéine ?

- La fonction d'un très grand nombre de protéines implique la formation d'un complexe avec une autre protéine (ex : anticorps-antigène, hormone-récepteur).
- La connaissance de la structure du complexe permet de comprendre le mécanisme : intérêt thérapeutique (ex : construction rationnelle d'inhibiteurs).
- La résolution expérimentale de la structure d'un complexe est problématique, coûteuse, voire impossible.
- Dans un organisme "simple" comme la Levure, on sait qu'il existe plus de 12 000 complexes différents.

Complexes protéine - protéine

Le problème

- Structures de deux protéines A et B connues
- Quelle est la meilleure conformation possible pour l'association de A et B ?
- Est-ce que cette association est suffisamment bonne pour que le complexe existe in vivo ?

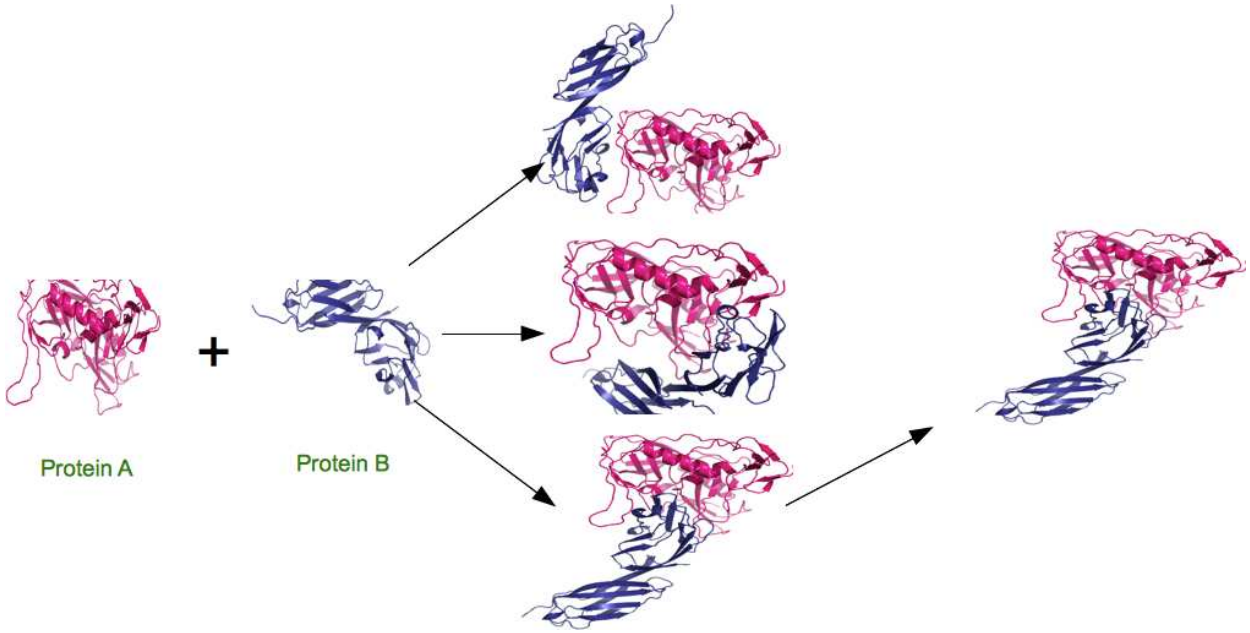
Existant

Dans les méthodes existantes, le problème est traité en deux étapes :

1. Génération d'un grand nombre de conformations
2. Utilisation d'une fonction de score pour classer les conformations

Principes communs à de nombreuses approches

Génération de conformations



Évaluation

- Utilisation de fonction énergétique
- Mesure de l'emboîtement
- Utilisation de connaissances expertes

Plan

Modélisation des protéines

Niveau atomique

- Précis mais très coûteux en temps de calcul ($\sim$ 10 000 atomes/protéine)
- Trop de degrés de liberté
- Trop sensible à la flexibilité

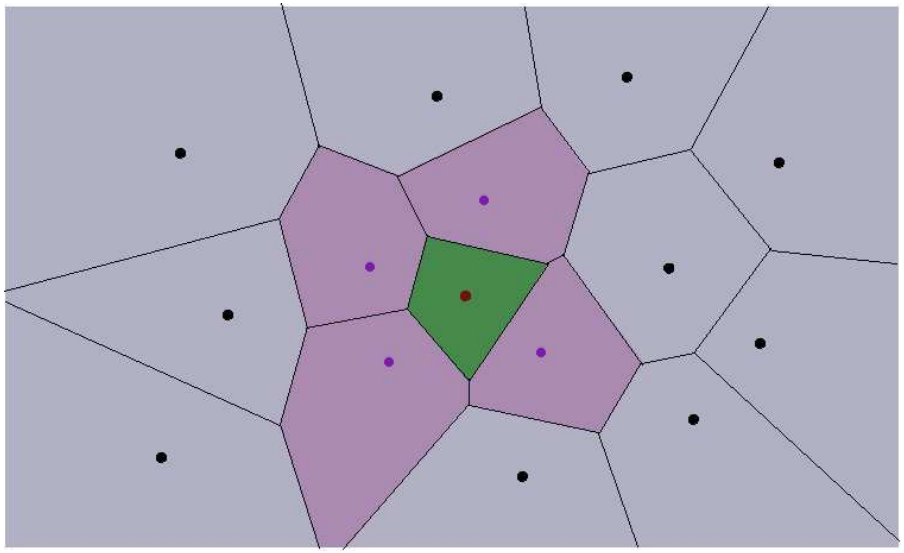
Niveau acides aminés (résidus)

- Moins précis que le modèle atomique mais suffisamment porteur d'informations
- Plus facile et plus rapide à évaluer ($\sim$ 400 résidus/protéine)
- Moins sensible à la flexibilité due à l'interaction

Modélisation d'une protéine par un polyèdre

Diagrammes de Voronoï

- Si on considère un ensemble de sites E : $V(p_i)$ est l'ensemble des points plus proche du site p_i que de tous les autres sites.
- $E = \{p_1, \ldots, p_n\}$
- $V(p_i) = \{x \in R^d : ||x - p_i|| \leq ||x - p_j||, \forall j \leq n\}$

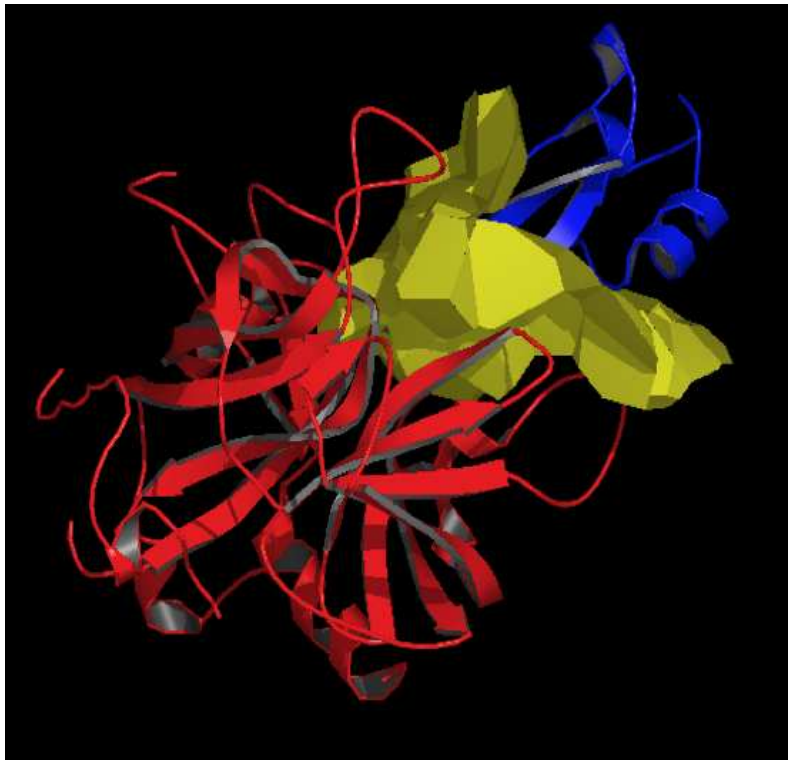


Description des complexes pour l'apprentissage

Extraction de mesures à partir des Voronoï

96 paramètres d'apprentissage

- Surface de l'interface (1)
- Nombre de résidus dans le cœur de l'interface (1)
- Volume de voronoï de chaque type de résidus (20)
- Fréquence d'apparition de chaque type de résidus (20)
- Fréquence des paires de résidus en contact (21 après regroupement)
- Distance de paires entre résidus (21 après regroupement)
- Fraction des résidus du cœur de l'interface de chaque catégorie) (6)
- Volume moyen de chaque catégorie (6)

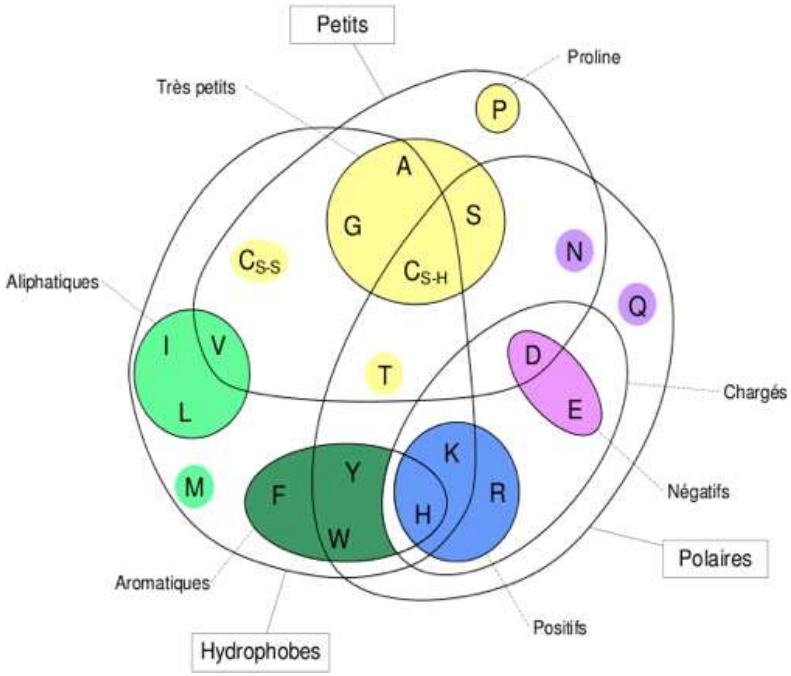


Description des complexes pour l'apprentissage

Extraction de mesures à partir des Voronoï

96 paramètres d'apprentissage

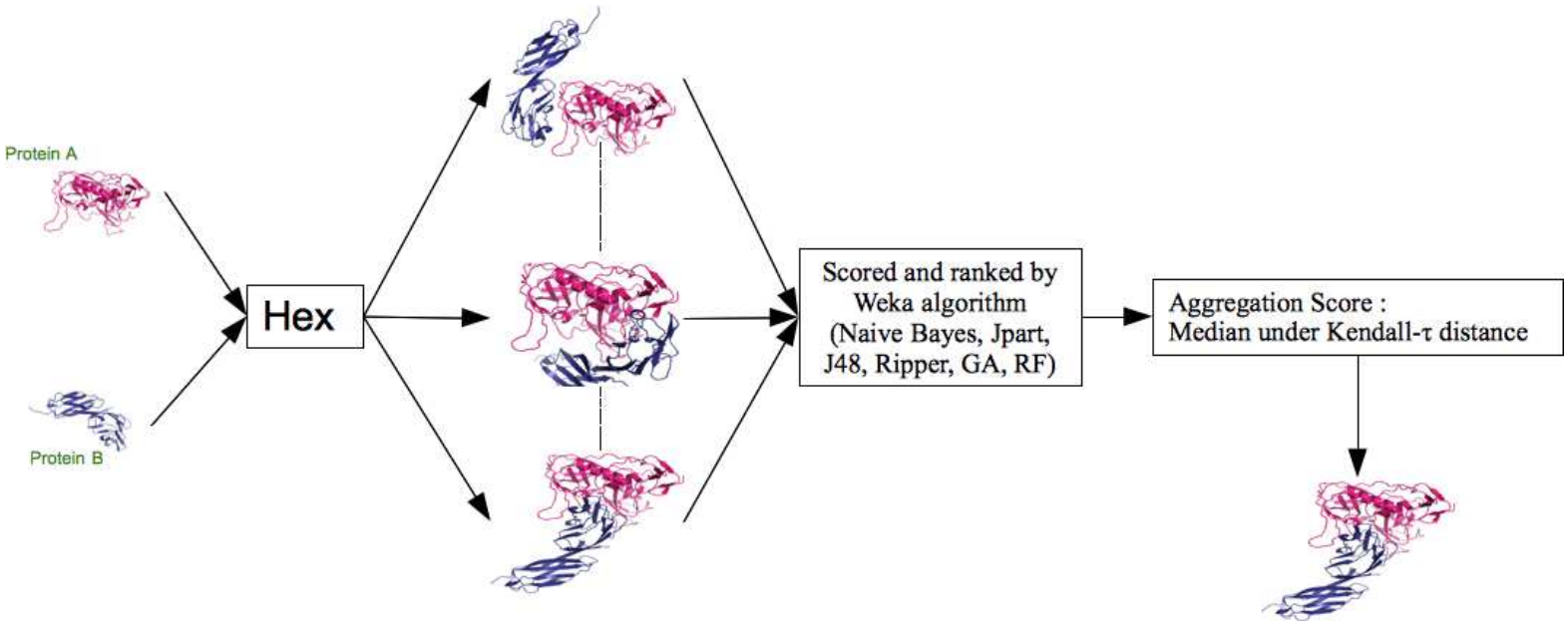
- Surface de l'interface (1)
- Nombre de résidus dans le cœur de l'interface (1)
- Volume de voronoï de chaque type de résidus (20)
- Fréquence d'apparition de chaque type de résidus (20)
- Fréquence des paires de résidus en contact (21 après regroupement)
- Distance de paires entre résidus (21 après regroupement)
- Fraction des résidus du cœur de l'interface de chaque catégorie) (6)
- Volume moyen de chaque catégorie (6)



Apprentissage et prédiction

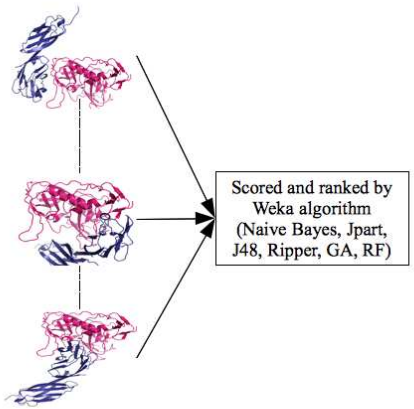
Combinaison de classifieurs multiples

- Utilisation de Hex pour générer rapidement un large ensemble de conformations plausibles
- Utilisation de plusieurs classifieurs pour associer un score à chaque conformation (Weka)
- Calcul d'un score consensuel



Apprentissage et prédiction

Étape de combinaison des classifieurs



$$\begin{cases} RF & = & [1\,2\,4\,5\,9\,10\,12\,\ldots\,50] \\ GA & = & [4\,9\,1\,5\,3\,10\,12\,\ldots\,50] \\ Ripper & = & [1\,4\,6\,2\,9\,10\,12\,\ldots\,36] \\ NaivesBayes & = & [4\,3\,1\,5\,9\,10\,12\,\ldots\,23] \\ \ldots & & \end{cases}$$

Comment aggréger efficacement l'ensemble des rangs prédits par les différents classifieurs ?

Apprentissage et prédiction

La distance Kendall-τ

- La distance de Kendall-τ est une mesure de dissimilarité permettant de traiter des tris de même taille et sans égalité
- Elle considère le nombre de paires de rangs en désaccord entre les différentes tris

Considérons deux permutations π^1 et π^2,

$$d_{KT}(\pi^1,\pi^2) = \#\{(i,j) : i < j \text{ et } \quad ((\pi^1[i] < \pi^1[j] \text{ et } \pi^2[i] > \pi^2[j]) \text{ ou } (\pi^1[i] > \pi^1[j] \text{ et } \pi^2[i] < \pi^2[j]))\}$$

Plan

Résultats

Protein-Protein Benchmark 4.0 (51 complexes)

top *N*	Hex	Kendall-τ	sumRank	medRank	b-RF	b-PART	b-JRip	b-J48	b-NaiveBayes
5	13(15)	13(20)	13(18)	9(16)	11(20)	12(19)	11(14)	10(12)	12(14)
10	18(30)	19(37)	20(35)	18(36)	19(36)	17(31)	20(29)	18(27)	13(20)
15	22(39)	23(48)	21(46)	23(51)	23(50)	20(39)	22(41)	23(40)	20(37)
25	26(58)	27(71)	25(67)	26(72)	26(69)	27(60)	27(62)	25(61)	25(64)
35	28(77)	29(90)	28(86)	28(85)	28(86)	29(85)	30(82)	28(78)	28(87)
45	30(96)	31(105)	30(105)	30(105)	31(106)	30(102)	30(102)	29(99)	30(106)
average	12,29	11,39	12,32	12,42	12,16	12,65	12,13	13,97	13,81

- La stratégie Kendall-τ présente des résultats comparables à l'ensemble des autres méthodes pour toute valeur de *N*
- En moyenne, le rang de la première conformation acceptable est meilleur pour Kendall-τ
- Ce "léger" mieux représente t-il un vrai gain sur la qualité des conformations prédites ?

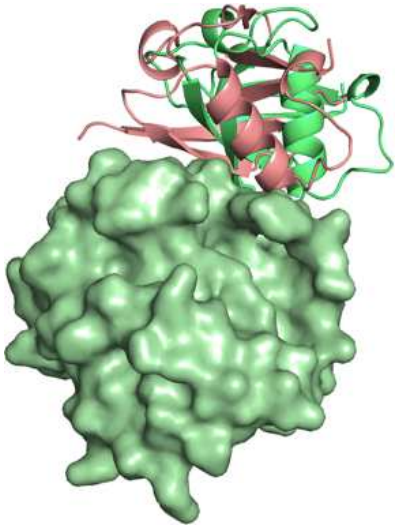
Résultats

Protein-Protein Benchmark 4.0 (51 complexes)

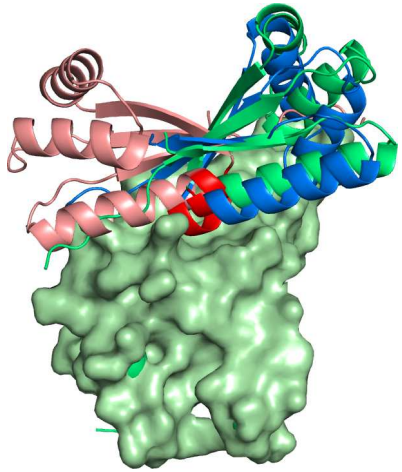
top *N*	Hex	Kendall-τ	sumRank	medRank	b-RF	b-PART	b-JRip	b-J48	b-NaiveBayes
5	13(15)	13(20)	13(18)	9(16)	11(20)	12(19)	11(14)	10(12)	12(14)
10	18(30)	19(37)	20(35)	18(36)	19(36)	17(31)	20(29)	18(27)	13(20)
15	22(39)	23(48)	21(46)	23(51)	23(50)	20(39)	22(41)	23(40)	20(37)
25	26(58)	27(71)	25(67)	26(72)	26(69)	27(60)	27(62)	25(61)	25(64)
35	28(77)	29(90)	28(86)	28(85)	28(86)	29(85)	30(82)	28(78)	28(87)
45	30(96)	31(105)	30(105)	30(105)	31(106)	30(102)	30(102)	29(99)	30(106)
average	12,29	11,39	12,32	12,42	12,16	12,65	12,13	13,97	13,81

- La stratégie Kendall-τ présente des résultats comparables à l'ensemble des autres méthodes pour toute valeur de *N*
- En moyenne, le rang de la première conformation acceptable est meilleur pour Kendall-τ
- Ce "léger" mieux représente t-il un vrai gain sur la qualité des conformations prédites ?

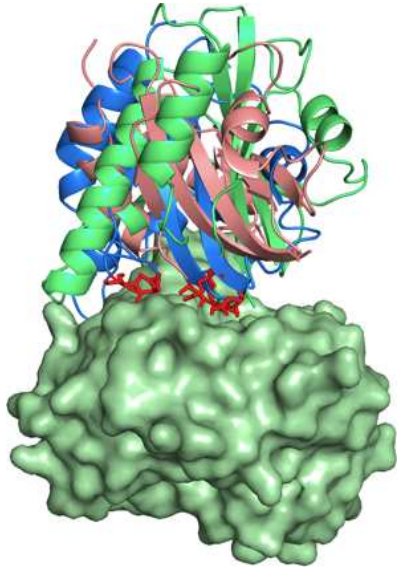
Pertinence biologique réelle . . .



1FLD
Ras-RalGds complex
rang(Kendall-τ) = 7
RMSD = 10.79 Å

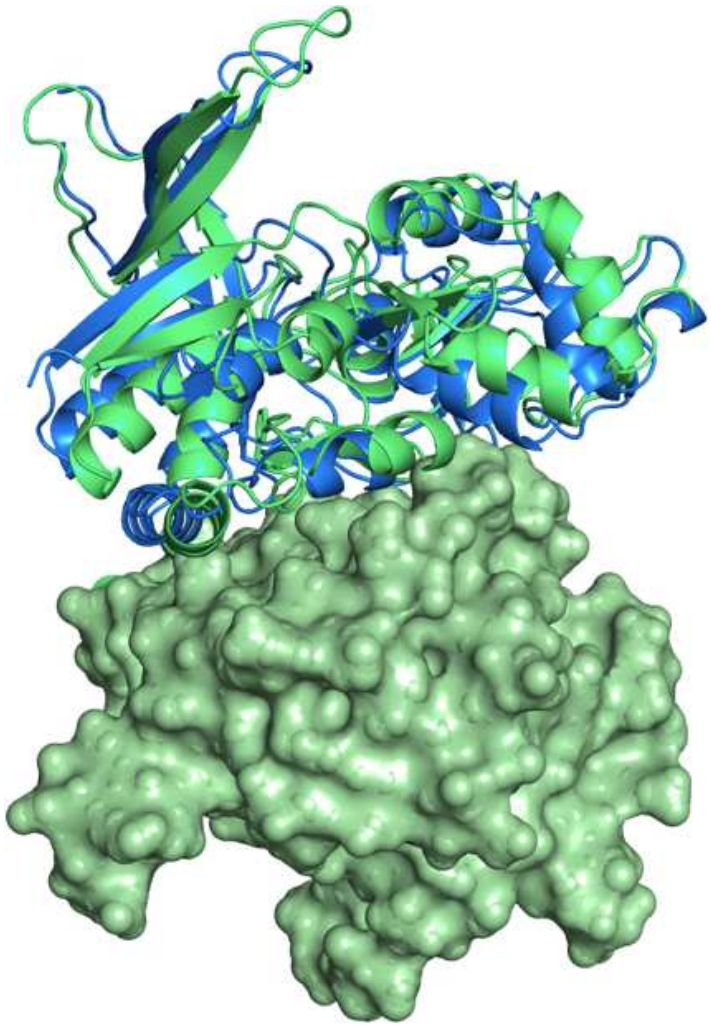


1R6Q
ClpNS/ClpA chaperone
rang(Kendall-τ) = 1
RMSD = 30.06 Å



2O3B
NucA/NuiA complex
rang(Kendall-τ) = 5
RMSD = 16.96 Å

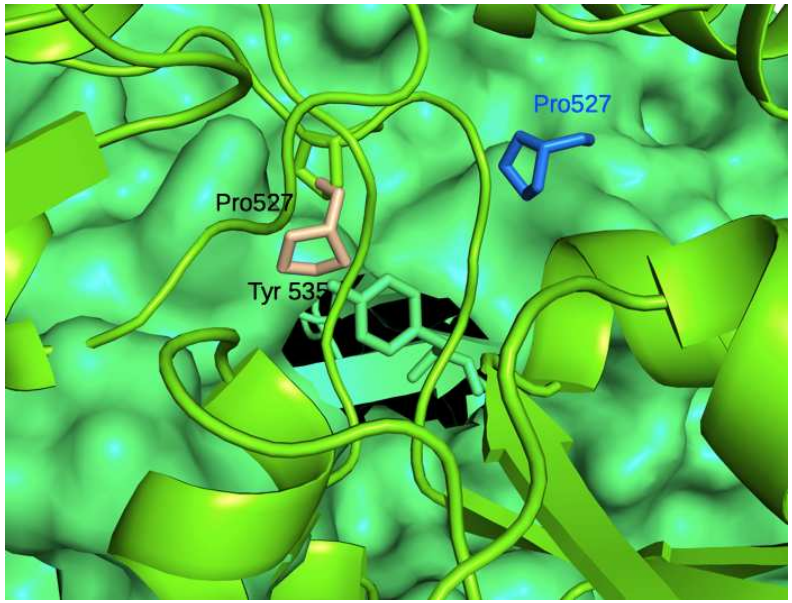
The NR1/NR2A ligand-binding cores complex



- Role crucial dans le système nerveux central des mammifères
- Hex génère des conformations poches de la conformation biologique
- Meilleure conformation de HEX (rang 17 avec RMSD = 7.69 Å)
- Même conformation en rang 45 avec Kendall-τ

The NR1/NR2A ligand-binding cores complex

- L'activité du canal ionique est régulée par des contacts VdW entre Y535 et P527
- L'analyse du cœur de l'interface montre que cette conformation essentielle proposée par Hex est perdue
- Rang(Kendall-τ) = 9 (RMSD = 17.95 Å)
- Rang pour Hex : 48

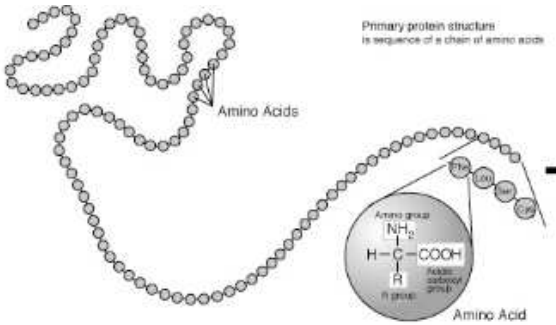


Performances

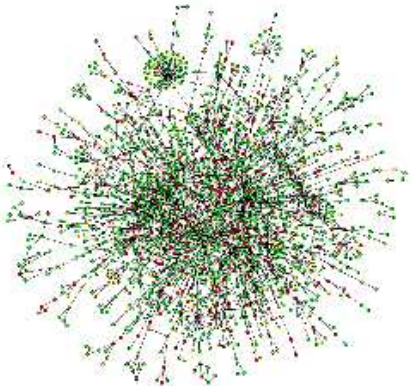
Temps de calcul

- Calculs des descripteurs pour un complexe protéine-protéine
  - Hex : $\sim$ 2 minutes pour générer 500 conformations
  - Voronoï (8 cœurs) : $\sim$ 4 minutes (partie coûteuse = solvatation et désolvatation)
- Espace disque : $\sim$ 1 Go par complexe (données temporaires pour les calculs de Voronoï)
- Apprentissage des modèles : $\sim 1h$
- Prédiction : quelques minutes
- Kendall-τ : $<$ 1h

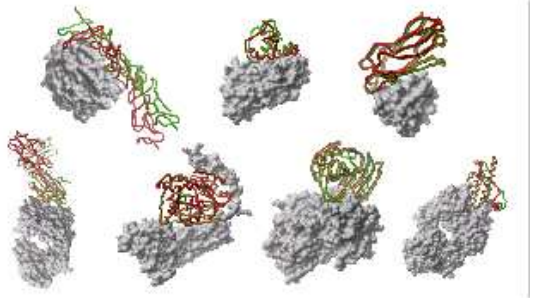
Conclusion et perspectives



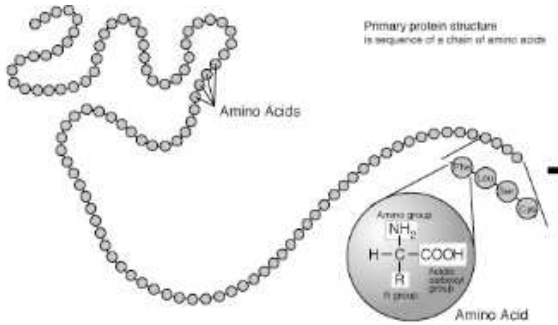
Interactions protéine-protéine



Structure 3D des protéines



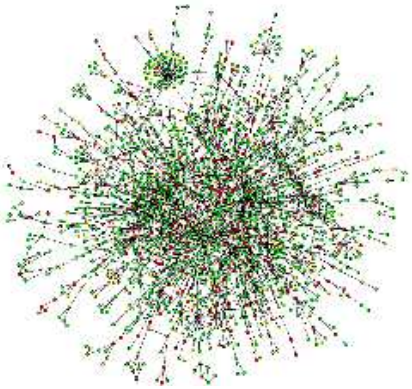
Conclusion et perspectives

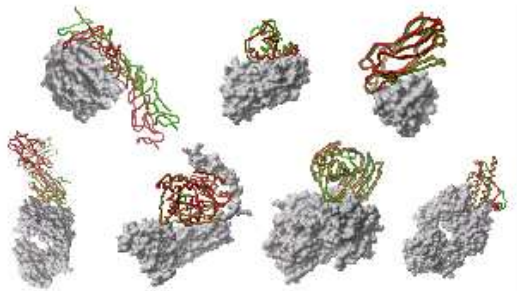


Pondérer les interactions dans la matrice de cross-docking

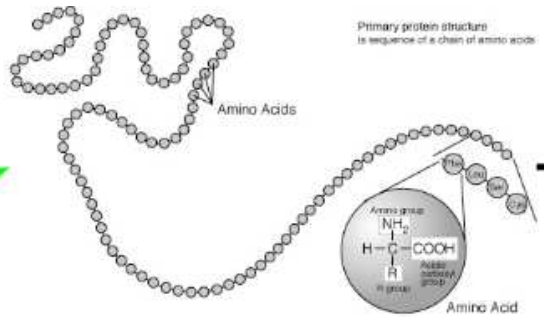
Interactions protéine-protéine

Structure 3D des protéines





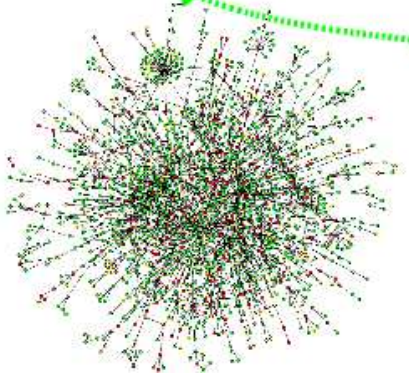
Conclusion et perspectives

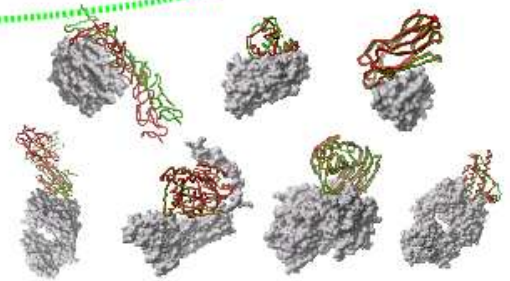


Pondérer les interactions dans la matrice de cross-docking

Interactions protéine-protéine

Structure 3D des protéines





Remerciements

NGS

- Jean-François Gibrat (INRA, MIG, Jouy-en-Josas)

PPI

- Damien de Vienne (CRG, Barcelone)
- Bernard Labedan (LRI, France)

Docking

- Thomas Bourquard (INRIA, Nancy)
- Anne Poupon (INRA, Tours)
- Dave Ritchie (INRIA, Nancy)
- Sylvie Hamel (Université de Montréal)
- Julie Bernauer (LIX, France)