# Prédiction grande échelle des interactions protéine/protéine

Yann Ponty

Analytical Genomics
A. Carbone's lab
INSERM U511/Paris 6
Paris, France

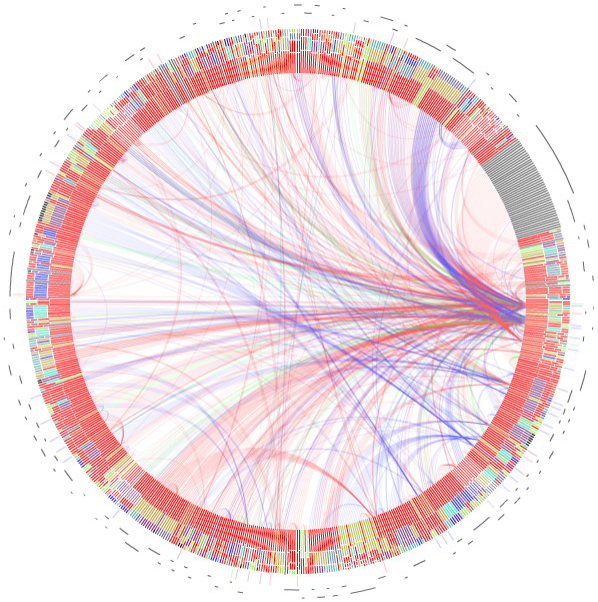
30 Avril 2009

## Travail en cours

Lewis Hine, 1920

Encore pas mal de boulons à serrer ...

# Prédiction d'interaction protéines/protéines



Réseau d'interaction de E. coli

PDB : $\sim$15000 modèles de protéines humaines

Lesquelles de ces protéines interagissent dans la cellule ?

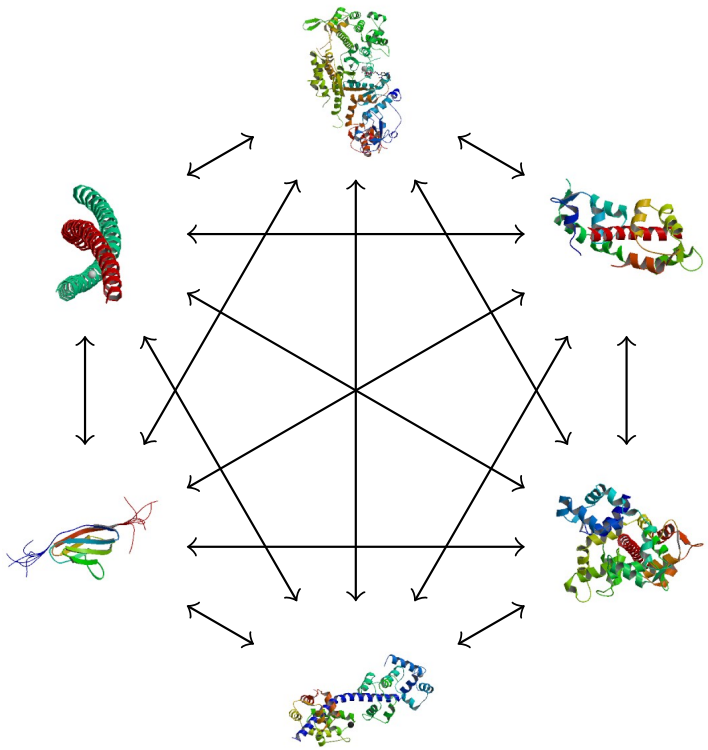
But : Établir des interactions/réseaux d'interaction inconnus
$\Rightarrow$ Portée théorique et thérapeutique

# Approche combinatoire

Docking croisé : Tester chaque couple de protéines, utiliser les résultats pour prédire les couples en interaction.

Données et approches disponibles :

- Biochimiques :
  Interaction fondée au niveau moléculaire
  $\Rightarrow$ Simulation de l'assemblage (Docking)
- Génomiques :
  Pression sélective pesant sur le réseau
  $\Rightarrow$ Conservation, mutations compensatoires
- Fonctionnelles : Protéines PDB connues
  $\Rightarrow$ Annotation (presque) systématique

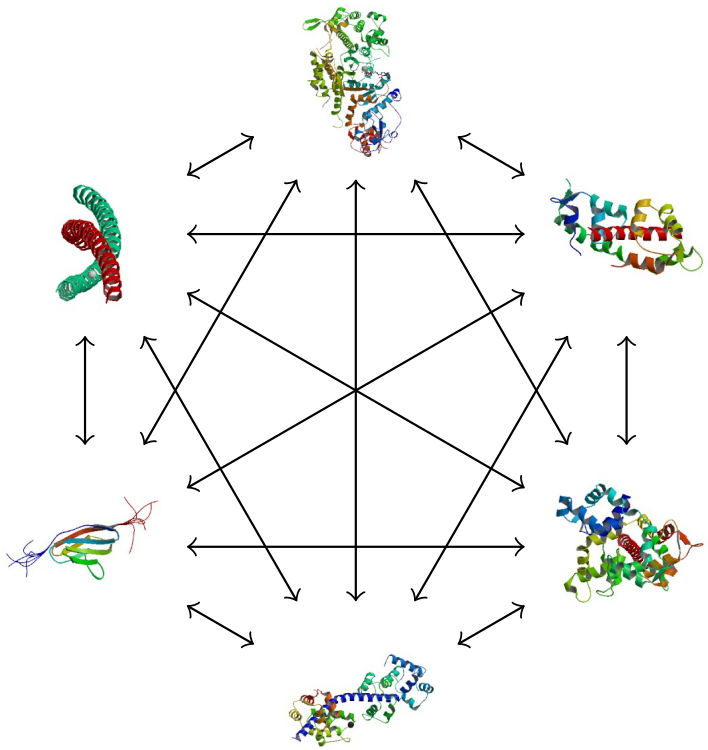


# Approche combinatoire

Docking croisé : Tester chaque couple de protéines, utiliser les résultats pour prédire les couples en interaction.

Données et approches disponibles :

- Biochimiques :
  Interaction fondée au niveau moléculaire
  $\Rightarrow$ Simulation de l'assemblage (Docking)
- Génomiques :
  Pression sélective pesant sur le réseau
  $\Rightarrow$ Conservation, mutations compensatoires
- Fonctionnelles : Protéines PDB connues
  $\Rightarrow$ Annotation (presque) systématique

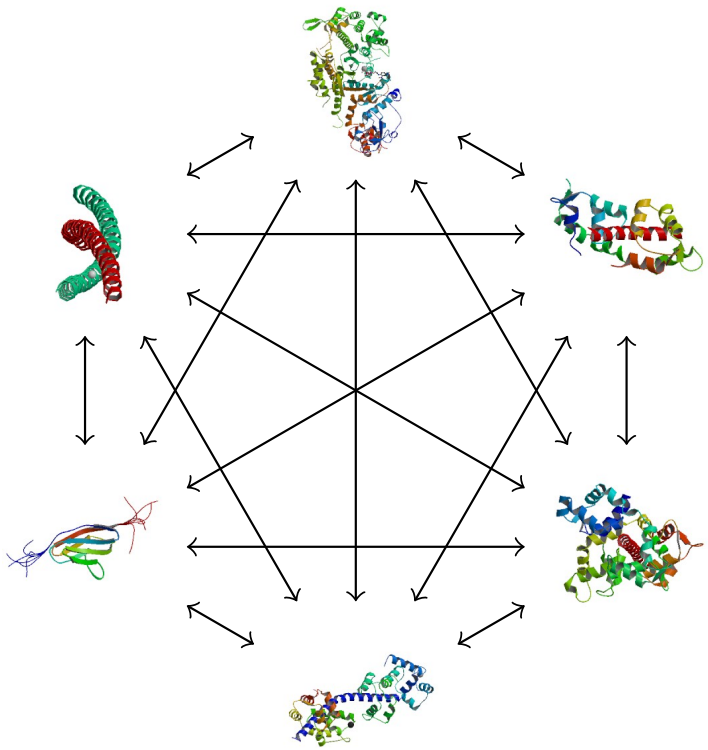


# Approche combinatoire

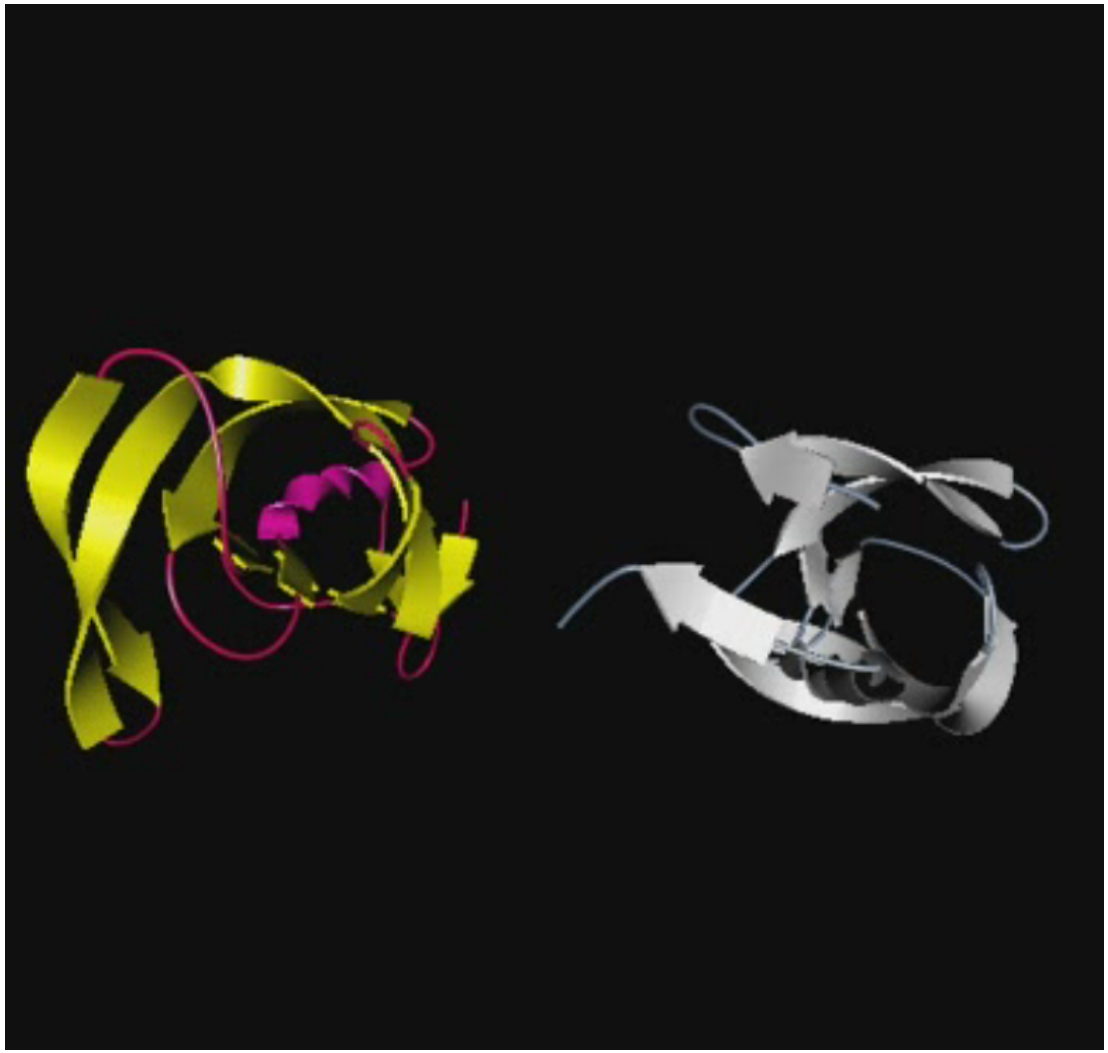
Docking croisé : Tester chaque couple de protéines, utiliser les résultats pour prédire les couples en interaction.

Données et approches disponibles :

- Biochimiques :
  Interaction fondée au niveau moléculaire
  $\Rightarrow$ Simulation de l'assemblage (Docking)
- Génomiques :
  Pression sélective pesant sur le réseau
  $\Rightarrow$ Conservation, mutations compensatoires
- Fonctionnelles : Protéines PDB connues
  $\Rightarrow$ Annotation (presque) systématique

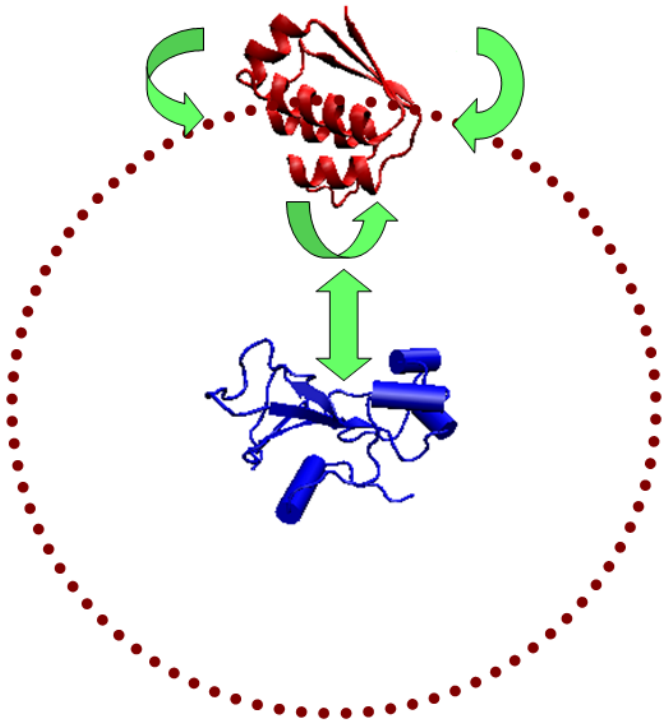


# Docking : Définition par l'exemple



O. Kreylos *et al*, IEEE Visualization 2003

# Docking : Principe général



*Recette classique* :

- Parcours d'un ensemble de positions relatives
- Rapprochement de la protéine mobile/Attribution d'un score
- Spécificités : Flexibilité/Représentation/Modèle d'énergie

# Docking : Temps de calcul

Exploration de $\sim$100 000–500 000 positions (1 pos. pour 10Å$^2$).
+ Évaluations multiples d'un score pour chaque position
$\Rightarrow$ 20–100 min CPU/Exploration dans des modèles rigides.

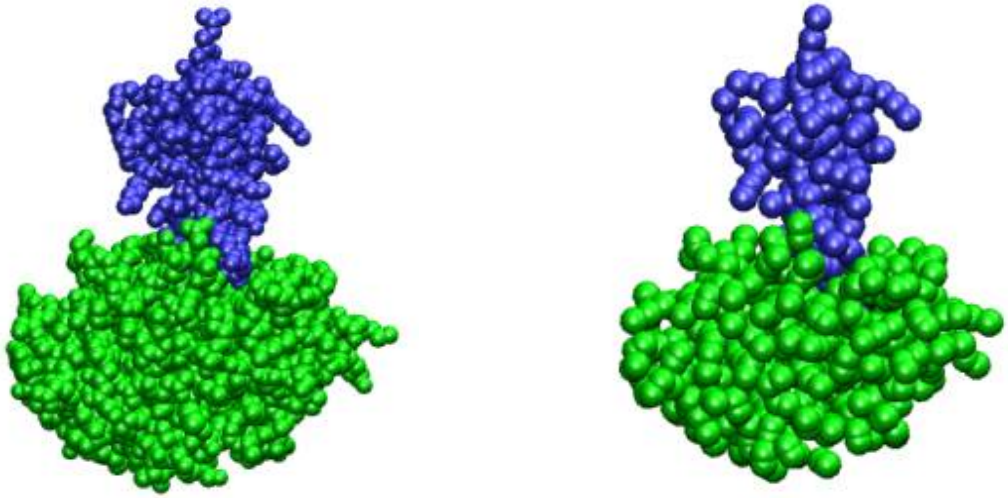
Temps de calcul rédhibitoires pour du docking croisé ...

Exemple :

- 168 protéines ($\Rightarrow$ $\sim$28000 couples)
- 200 000 positions par couple
- 400 secondes CPU par position

715 s. 98 a. 63 j. 8 h.
CPU *domestique*

# MAXDO : Docking simplifié



MAXDO (Sacquin-Mora, Carbone et Lavery, J Mol Biol 2008)

Représentation *gros grain* : Un à trois *pseudo-atomes* par résidu (Zacharias, Prot. Sci. 2003).

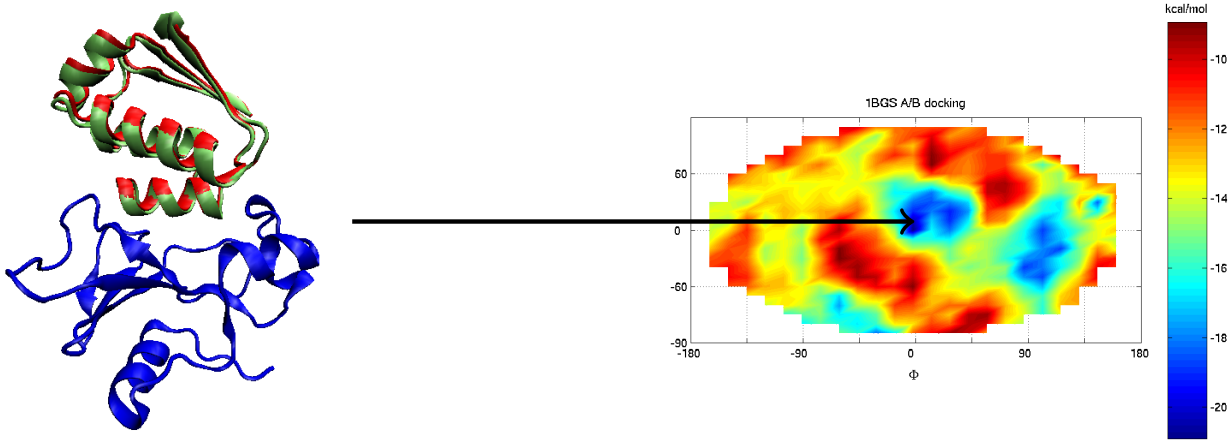
$\Rightarrow \sim$ 30 000 pos. par couple, $\sim$ 200 sec. par position.

Fonction d'énergie simple : $E_{ij} = \left(\frac{B_{ij}}{r_{ij}^8} - \frac{C_{ij}}{r_{ij}^6}\right) + \frac{q_i q_j}{15 r_{ij}^2}$

Où $r_{ij}$ : distance $ij$, $B_{ij}$ ($C_{ij}$) : potentiels LJ d'attraction (repulsion) et $q_i/q_j$ : charges électrostatiques des pseudo-atomes $i/j$.

Docking rigide

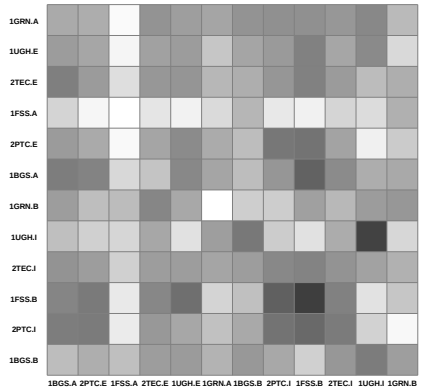
# MAXDO : Résultats




Bonne nouvelle : MAXDO approche à $\sim$ 1.5Å les assemblages connus*

Mais mauvaise nouvelle :

Énergie seule non-discriminante

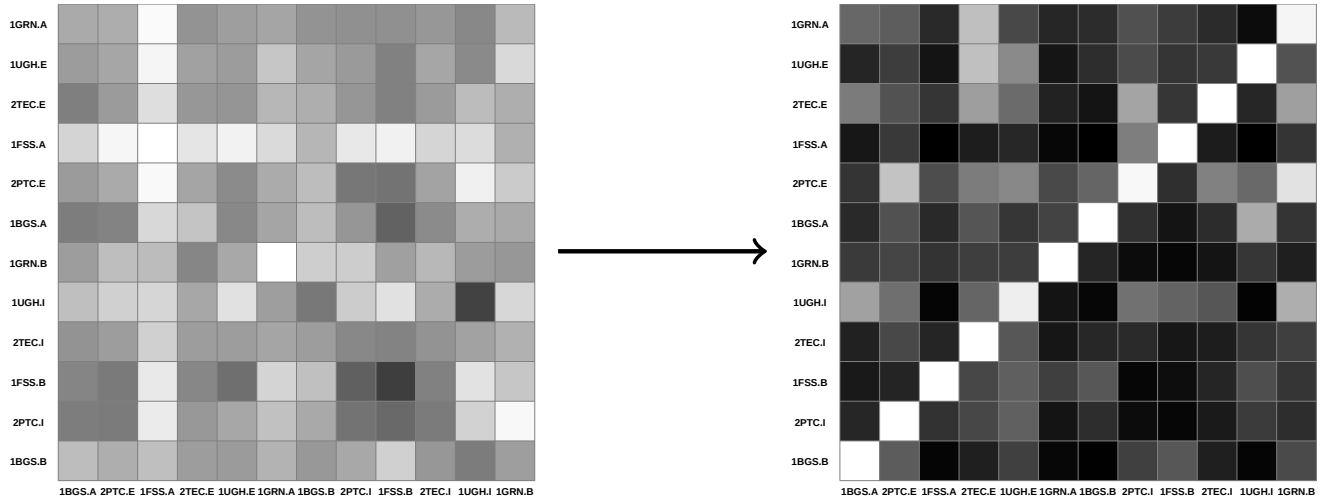


# MAXDO : Résultats

Mauvaise nouvelle : Énergie seule non-discriminante
Bonne nouvelle : Couple Énergie + Interface discriminant ! ! !

$$II_{\mathcal{P}_1\mathcal{P}_2} = \mathrm{Min}_{i \in I_{\mathcal{P}_1\mathcal{P}_2}}(E_i F_{i,\mathcal{P}_1} F_{i,\mathcal{P}_2}) \qquad NII_{\mathcal{P}_1\mathcal{P}_2} = \frac{II^2_{\mathcal{P}_1\mathcal{P}_2}}{\mathrm{Min}_{\mathcal{P}} II_{\mathcal{P}_1\mathcal{P}} \cdot \mathrm{Min}_{\mathcal{P}} II_{\mathcal{P}\mathcal{P}_2}}$$

Où $I$ : Ensembles des interfaces de $\mathcal{P}_1\mathcal{P}_2$ ; $E_i$ : Énergie d'interaction ;
$F_{i,\mathcal{P}_1}$ : Proportion de l'interface expérimentale *couverte* par l'interface $i$.



# WCG Phase 1 : Présentation

Pb éventuels de passage à l'échelle :

- Est ce que les *NII* résistent au bruit ?
- Calibration des (nombreux) paramètres
- Robustesse du code (Format PDB laxiste)

Utilisation d'un benchmark de docking (Mintseris *et al*, Proteins 2005) composé de 168 partenaires potentiels, dont :

(E) 23 couples enzyme/inhibiteur ou enzyme/substrat

(AB) 10 couples anticorps/antigène version liée

(A) 10 couples anticorps/antigène version non-liée

Problème de temps de calcul :
Malgré les gains de MAXDO, encore $\sim$**80 siècles** de temps de calcul

$\Rightarrow$ World Community Grid (IBM) :
$\sim$7 mois de calculs sur les $\sim$35 000 machines disponibles

# WCG Phase 1 : Résultats

**Observation 1 :** Même bruité par un grand nombre d'interactions potentielles, le critère *NII* reste discriminant (Courbe ROC).

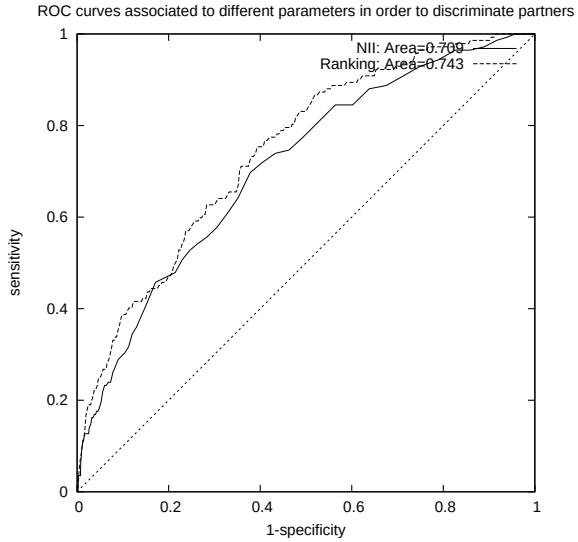
⇒ Comment remplacer des données d'interfaces expérimentales ? ? ?

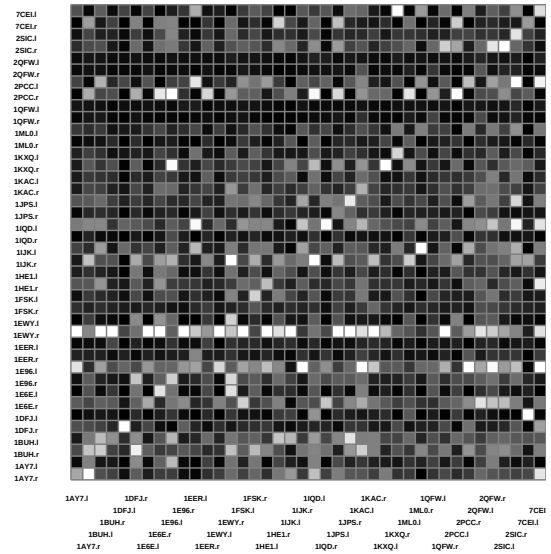
Observation 2 : Les pertes de performances sont souvent liées à la présence de chaines glissantes ou collantes.

⇒ Renormaliser en tenant compte du comportement moyen (*NII* utilise des min)

Observation 3 : Comportements différenciés selon les classes fonctionnelles

⇒ Tenir compte des annotations disponibles



# WCG Phase 1 : Résultats

Observation 1 : Même bruité par un grand nombre d'interactions potentielles, le critère *NII* reste discriminant (Courbe ROC).

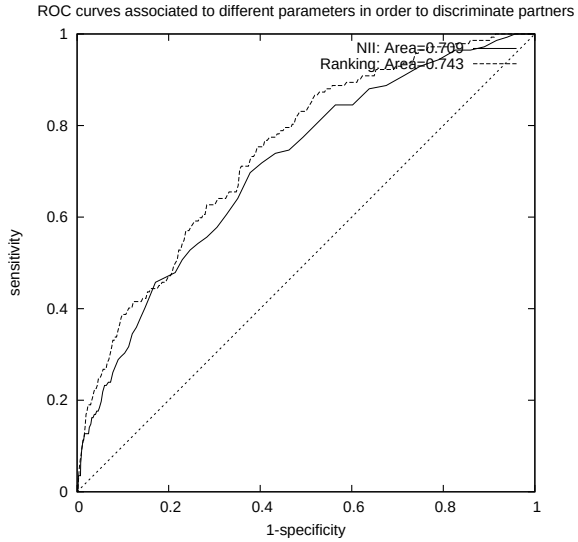
⇒ Comment remplacer des données d'interfaces expérimentales ? ? ?

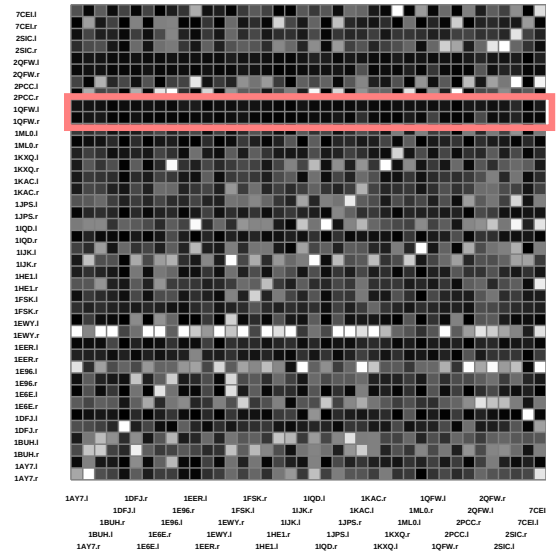
**Observation 2 :** Les pertes de performances sont souvent liées à la présence de chaines **glissantes** ou collantes.

⇒ Renormaliser en tenant compte du comportement moyen (*NII* utilise des min)

Observation 3 : Comportements différenciés selon les classes fonctionnelles

⇒ Tenir compte des annotations disponibles



# WCG Phase 1 : Résultats

Observation 1 : Même bruité par un grand nombre d'interactions potentielles, le critère *NII* reste discriminant (Courbe ROC).

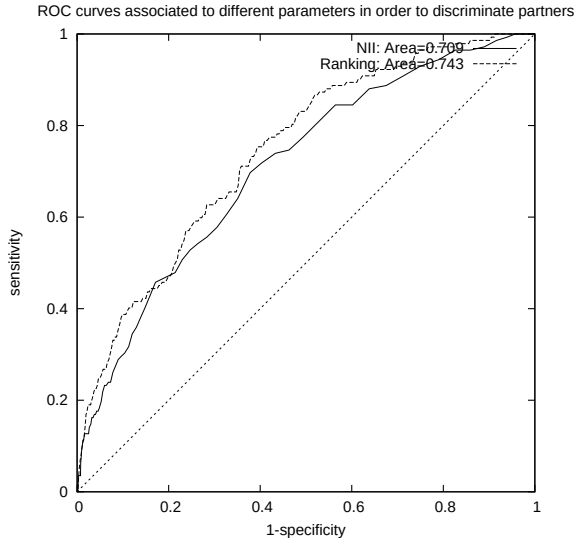
⇒ Comment remplacer des données d'interfaces expérimentales ? ? ?

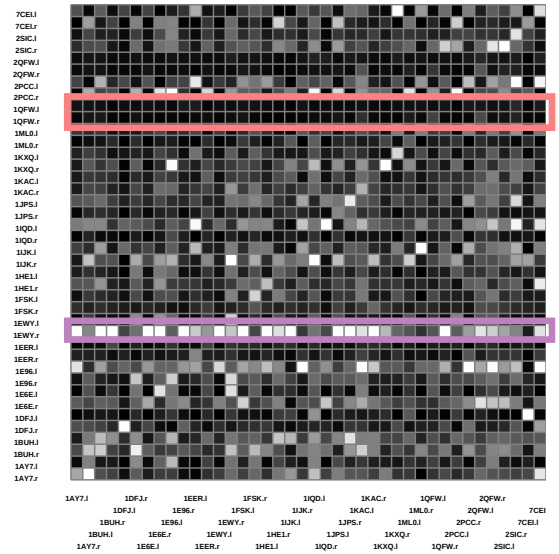
**Observation 2 :** Les pertes de performances sont souvent liées à la présence de chaines glissantes ou **collantes**.

⇒ Renormaliser en tenant compte du comportement moyen (*NII* utilise des min)

Observation 3 : Comportements différenciés selon les classes fonctionnelles

⇒ Tenir compte des annotations disponibles



# WCG Phase 1 : Résultats

Observation 1 : Même bruité par un grand nombre d'interactions potentielles, le critère *NII* reste discriminant (Courbe ROC).

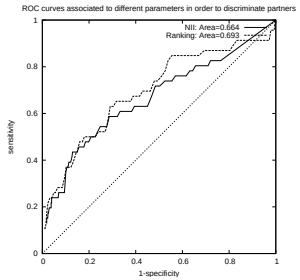
⇒ Comment remplacer des données d'interfaces expérimentales ? ? ?

Observation 2 : Les pertes de performances sont souvent liées à la présence de chaines glissantes ou collantes.

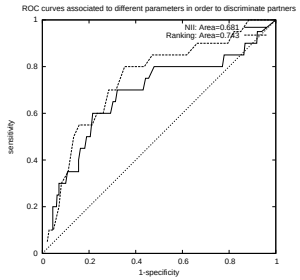
⇒ Renormaliser en tenant compte du comportement moyen (*NII* utilise des min)

**Observation 3 :** Comportements différenciés selon les classes fonctionnelles

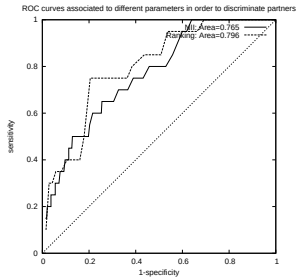
⇒ Tenir compte des annotations disponibles



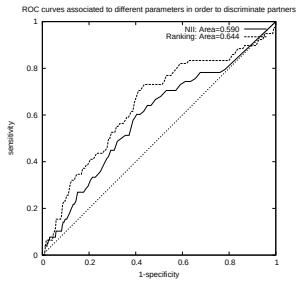

(E) : 69% AUC




(A) : 74% AUC




(AB) : 80% AUC




(O) : 64% AUC

# WCG Phase 1 : Résultats

Observation 1 : Même bruité par un grand nombre d'interactions potentielles, le critère *NII* reste discriminant (Courbe ROC).

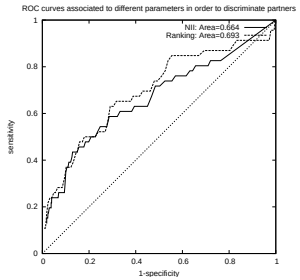
⇒ **Comment remplacer des données d'interfaces expérimentales ? ? ?**

Observation 2 : Les pertes de performances sont souvent liées à la présence de chaines glissantes ou collantes.

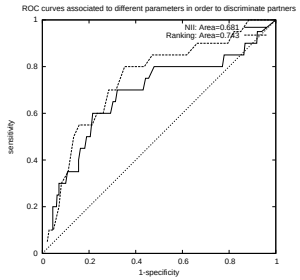
⇒ Renormaliser en tenant compte du comportement moyen (*NII* utilise des min)

Observation 3 : Comportements différenciés selon les classes fonctionnelles

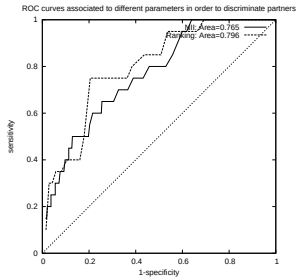
⇒ Tenir compte des annotations disponibles



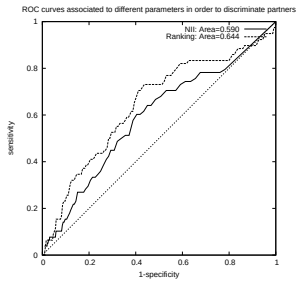

(E) : 69% AUC




(A) : 74% AUC

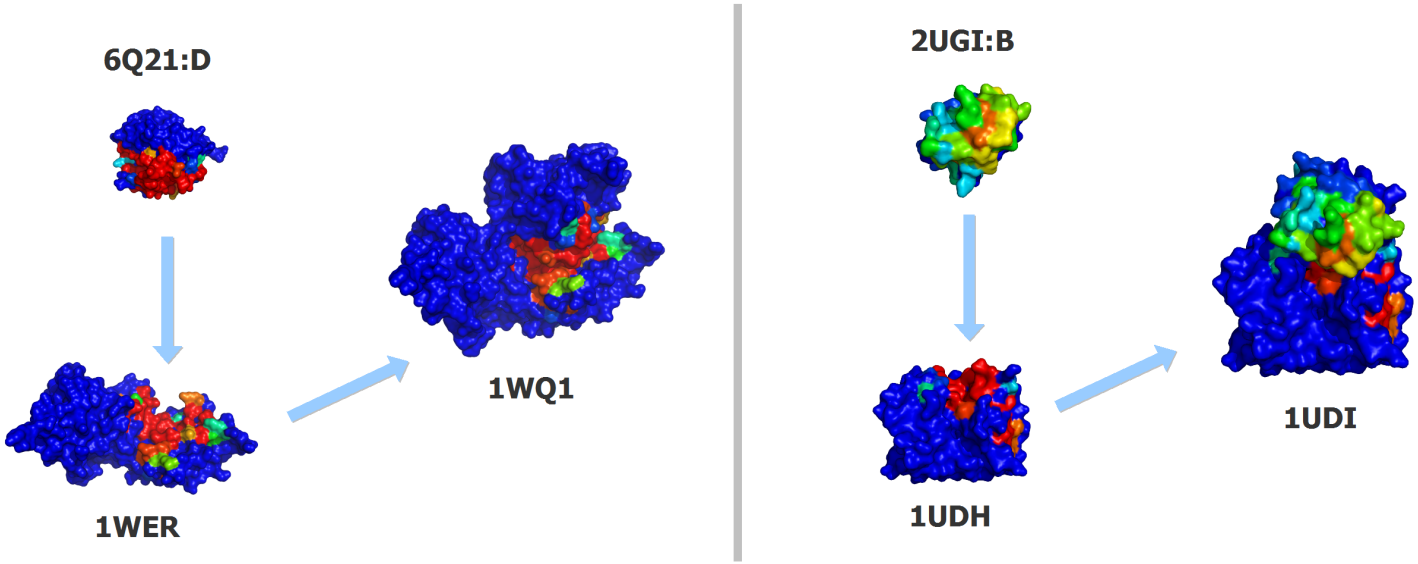



(AB) : 80% AUC




(O) : 64% AUC

# JET : Présentation




JET (Engelen *et al*, PLOS Comput Bio 09) est un outil robuste pour la détection des interfaces protéines/protéines par une approche évolutive.

Idées clés :

- Conservation des résidus aux interfaces (Lichtarge *et al*, J Mol Biol 96)
- Biais de composition des interfaces (Negi *et al*, J Mol Model 07)
- Relation #Résidus/taille de la protéine (Chen *et al*, Proteins 05)
- Interface = *Patch* centré sur un pic de conservation

# JET : Méthode

Méthode générale :

A Rapatriement de séquences homologues $S$ (PsiBlast)

B Regroupement des séquences par homologie $S_{20\text{-}40}, \ldots, S_{80\text{-}98}$

C Échantillonnage de $N_T = \sqrt{|S|}$ sous-ensembles de $S_T = \sqrt{|S|}$ séquences empruntant également aux différentes classes.

D Construction de $N_T$ arbres phylogénétiques par alignement multiple (ClustalW) puis Neighbor Joining.

E Association d'une **trace** évolutive aux résidus. ▸ Exemple

F Attribution d'un score de **propriété physico-chimique** (PC).

G Construction de *coeurs* de trace **ou** PC significativement élevée.

H Extension des coeurs par élargissement aux résidus connexes de forte conservation et/ou PC.

I Itération (Typiquement 10 fois) et sélection des résidus récurrents.

# JET : Méthode

Méthode générale :

A Rapatriement de séquences homologues $S$ (PsiBlast)

B Regroupement des séquences par homologie $S_{20\text{-}40}, \ldots, S_{80\text{-}98}$

C Échantillonnage de $N_T = \sqrt{|S|}$ sous-ensembles de $S_T = \sqrt{|S|}$ séquences empruntant également aux différentes classes.

D Construction de $N_T$ arbres phylogénétiques par alignement multiple (ClustalW) puis Neighbor Joining.

E Association d'une **trace** évolutive aux résidus. ▸ Exemple

F Attribution d'un score de **propriété physico-chimique** (PC).

G Construction de *coeurs* de trace **ou** PC significativement élevée.

H Extension des coeurs par élargissement aux résidus connexes de forte conservation et/ou PC.

I Itération (Typiquement 10 fois) et sélection des résidus récurrents.

# JET : Méthode

Méthode générale :

A Rapatriement de séquences homologues $S$ (PsiBlast)

B Regroupement des séquences par homologie $S_{20\text{-}40}, \ldots, S_{80\text{-}98}$

C Échantillonnage de $N_T = \sqrt{|S|}$ sous-ensembles de $S_T = \sqrt{|S|}$ séquences empruntant également aux différentes classes.

D Construction de $N_T$ arbres phylogénétiques par alignement multiple (ClustalW) puis Neighbor Joining.

E Association d'une **trace** évolutive aux résidus. ▸ Exemple

F Attribution d'un score de **propriété physico-chimique** (PC).

G Construction de *coeurs* de trace **ou** PC significativement élevée.

H Extension des coeurs par élargissement aux résidus connexes de forte conservation et/ou PC.

I Itération (Typiquement 10 fois) et sélection des résidus récurrents.

# JET : Méthode

Méthode générale :

A Rapatriement de séquences homologues $S$ (PsiBlast)

B Regroupement des séquences par homologie $S_{20\text{-}40}, \ldots, S_{80\text{-}98}$

C Échantillonnage de $N_T = \sqrt{|S|}$ sous-ensembles de $S_T = \sqrt{|S|}$ séquences empruntant également aux différentes classes.

D Construction de $N_T$ arbres phylogénétiques par alignement multiple (ClustalW) puis Neighbor Joining.

E Association d'une **trace** évolutive aux résidus. ▸ Exemple

F Attribution d'un score de **propriété physico-chimique** (PC).

G Construction de *coeurs* de trace **ou** PC significativement élevée.

H Extension des coeurs par élargissement aux résidus connexes de forte conservation et/ou PC.

I Itération (Typiquement 10 fois) et sélection des résidus récurrents.

# JET : Méthode

Méthode générale :

A Rapatriement de séquences homologues $S$ (PsiBlast)

B Regroupement des séquences par homologie $S_{20\text{-}40}, \ldots, S_{80\text{-}98}$

C Échantillonnage de $N_T = \sqrt{|S|}$ sous-ensembles de $S_T = \sqrt{|S|}$ séquences empruntant également aux différentes classes.

D Construction de $N_T$ arbres phylogénétiques par alignement multiple (ClustalW) puis Neighbor Joining.

E Association d'une **trace** évolutive aux résidus. ▸ Exemple

F Attribution d'un score de **propriété physico-chimique** (PC).

G Construction de *coeurs* de trace **ou** PC significativement élevée.

H Extension des coeurs par élargissement aux résidus connexes de forte conservation et/ou PC.

I Itération (Typiquement 10 fois) et sélection des résidus récurrents.

# JET : Méthode

Méthode générale :

A Rapatriement de séquences homologues $S$ (PsiBlast)

B Regroupement des séquences par homologie $S_{20\text{-}40}, \ldots, S_{80\text{-}98}$

C Échantillonnage de $N_T = \sqrt{|S|}$ sous-ensembles de $S_T = \sqrt{|S|}$ séquences empruntant également aux différentes classes.

D Construction de $N_T$ arbres phylogénétiques par alignement multiple (ClustalW) puis Neighbor Joining.

E Association d'une **trace** évolutive aux résidus. ▸ Exemple

F Attribution d'un score de **propriété physico-chimique** (PC).

G Construction de *coeurs* de trace **ou** PC significativement élevée.

H Extension des coeurs par élargissement aux résidus connexes de forte conservation et/ou PC.

I Itération (Typiquement 10 fois) et sélection des résidus récurrents.

# JET : Méthode

Méthode générale :

A Rapatriement de séquences homologues $S$ (PsiBlast)

B Regroupement des séquences par homologie $S_{20\text{-}40}, \ldots, S_{80\text{-}98}$

C Échantillonnage de $N_T = \sqrt{|S|}$ sous-ensembles de $S_T = \sqrt{|S|}$ séquences empruntant également aux différentes classes.

D Construction de $N_T$ arbres phylogénétiques par alignement multiple (ClustalW) puis Neighbor Joining.

E Association d'une **trace** évolutive aux résidus. ▸ Exemple

F Attribution d'un score de **propriété physico-chimique** (PC).

G Construction de *coeurs* de trace **ou** PC significativement élevée.

H Extension des coeurs par élargissement aux résidus connexes de forte conservation et/ou PC.

I Itération (Typiquement 10 fois) et sélection des résidus récurrents.

# JET : Méthode

Méthode générale :

A Rapatriement de séquences homologues $S$ (PsiBlast)

B Regroupement des séquences par homologie $S_{20\text{-}40}, \ldots, S_{80\text{-}98}$

C Échantillonnage de $N_T = \sqrt{|S|}$ sous-ensembles de $S_T = \sqrt{|S|}$ séquences empruntant également aux différentes classes.

D Construction de $N_T$ arbres phylogénétiques par alignement multiple (ClustalW) puis Neighbor Joining.

E Association d'une **trace** évolutive aux résidus. ▸ Exemple

F Attribution d'un score de **propriété physico-chimique** (PC).

G Construction de *coeurs* de trace **ou** PC significativement élevée.

H Extension des coeurs par élargissement aux résidus connexes de forte conservation et/ou PC.

I Itération (Typiquement 10 fois) et sélection des résidus récurrents.

# JET : Méthode

Méthode générale :

A Rapatriement de séquences homologues $S$ (PsiBlast)

B Regroupement des séquences par homologie $S_{20\text{-}40}, \ldots, S_{80\text{-}98}$

C Échantillonnage de $N_T = \sqrt{|S|}$ sous-ensembles de $S_T = \sqrt{|S|}$ séquences empruntant également aux différentes classes.

D Construction de $N_T$ arbres phylogénétiques par alignement multiple (ClustalW) puis Neighbor Joining.

E Association d'une **trace** évolutive aux résidus. ▸ Exemple

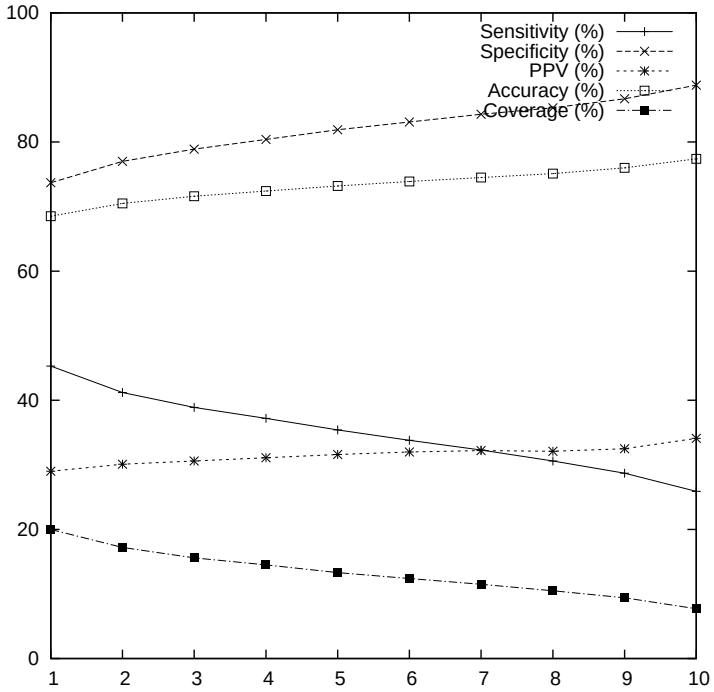
F Attribution d'un score de **propriété physico-chimique** (PC).

G Construction de *coeurs* de trace **ou** PC significativement élevée.

H Extension des coeurs par élargissement aux résidus connexes de forte conservation et/ou PC.

I Itération (Typiquement 10 fois) et sélection des résidus récurrents.

# JET : Résultats sur la base Mintseris



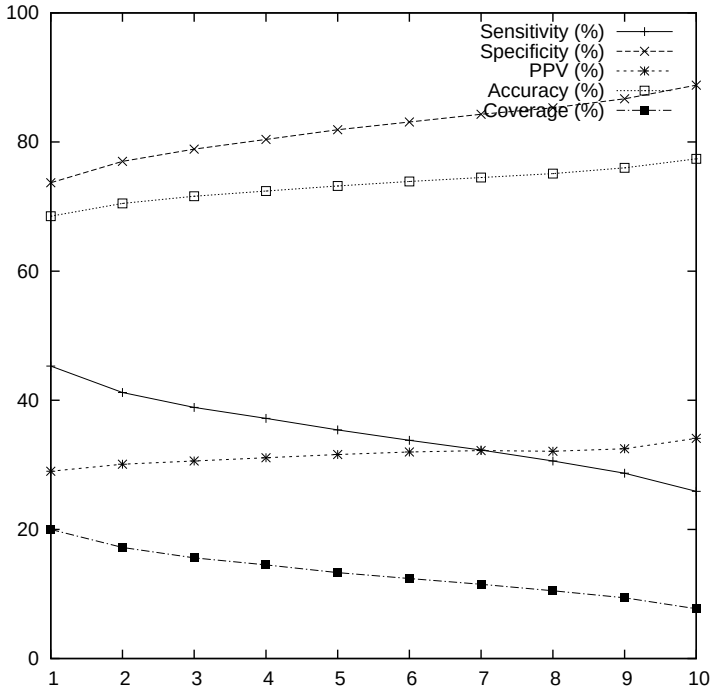

| Classe | Sen% | Spe% | PPV% | Cov% |
|---|---|---|---|---|
| A | 12.6 | 84.4 | 22.5 | 10.6 |
| AB | 15.9 | 86.6 | 28.0 | 9.6 |
| E | 49.3 | 79.3 | 40.8 | 14.8 |
| O | 33.0 | 86.4 | 29.1 | 11.2 |

(Haut) Résultats obtenus au moins $k = 7$ occurrences parmi 10 itérations.

(Gauche) Évolution des performances pour différents seuils $k$.

- Spécificité remarquable : Typiquement $\sim$80%
- Sensibilité variable : $>$20% nécessaires pour *diriger* MAXDO
- PPV stable sur $k$ : Proportion de *bons* résidus constante sur $k$
  $\Rightarrow$ Baisser $k \rightarrow 1$ pour les classes difficiles
- Couverture $<$20% : Restreindre les positions de MAXDO.

# JET : Résultats sur la base Mintseris



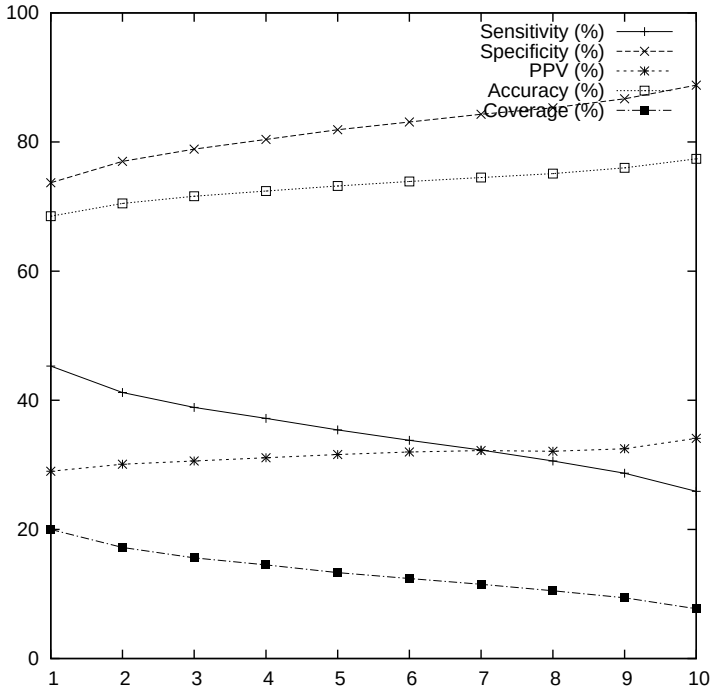

| Classe | Sen% | Spe% | PPV% | Cov% |
|---|---|---|---|---|
| A | 12.6 | 84.4 | 22.5 | 10.6 |
| AB | 15.9 | 86.6 | 28.0 | 9.6 |
| E | 49.3 | 79.3 | 40.8 | 14.8 |
| O | 33.0 | 86.4 | 29.1 | 11.2 |

(Haut) Résultats obtenus au moins $k = 7$ occurrences parmi 10 itérations.

(Gauche) Évolution des performances pour différents seuils $k$.

- Spécificité remarquable : Typiquement $\sim$80%
- Sensibilité variable : $>$20% nécessaires pour *diriger* MAXDO
- PPV stable sur $k$ : Proportion de *bons* résidus constante sur $k$
  $\Rightarrow$ Baisser $k \rightarrow 1$ pour les classes difficiles
- Couverture $<$20% : Restreindre les positions de MAXDO.

# JET : Résultats sur la base Mintseris



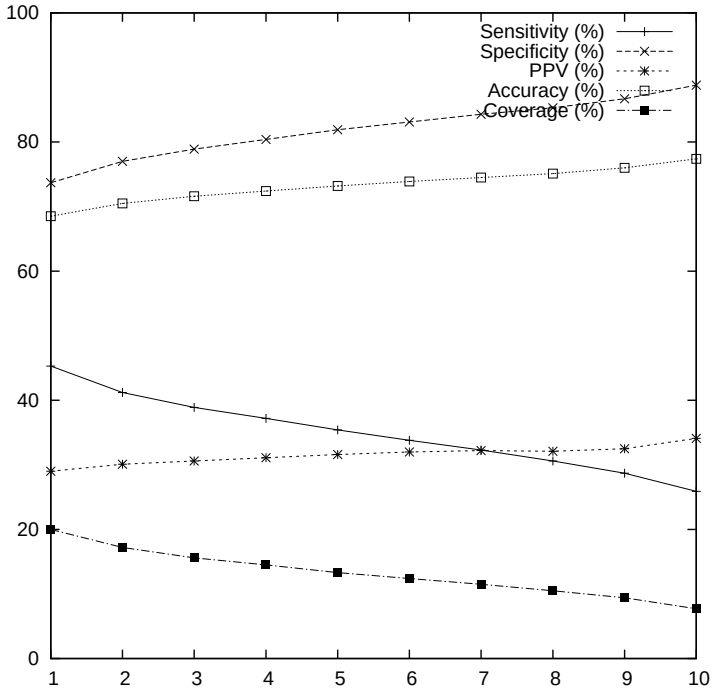

| Classe | Sen% | Spe% | PPV% | Cov% |
|---|---|---|---|---|
| A | 12.6 | 84.4 | 22.5 | 10.6 |
| AB | 15.9 | 86.6 | 28.0 | 9.6 |
| E | 49.3 | 79.3 | 40.8 | 14.8 |
| O | 33.0 | 86.4 | 29.1 | 11.2 |

(Haut) Résultats obtenus au moins $k = 7$ occurrences parmi 10 itérations.

(Gauche) Évolution des performances pour différents seuils $k$.

- Spécificité remarquable : Typiquement $\sim$80%
- Sensibilité variable : >20% nécessaires pour *diriger* MAXDO
- PPV stable sur $k$ : Proportion de *bons* résidus constante sur $k$
  $\Rightarrow$ Baisser $k \rightarrow 1$ pour les classes difficiles
- Couverture <20% : Restreindre les positions de MAXDO.

# JET : Résultats sur la base Mintseris



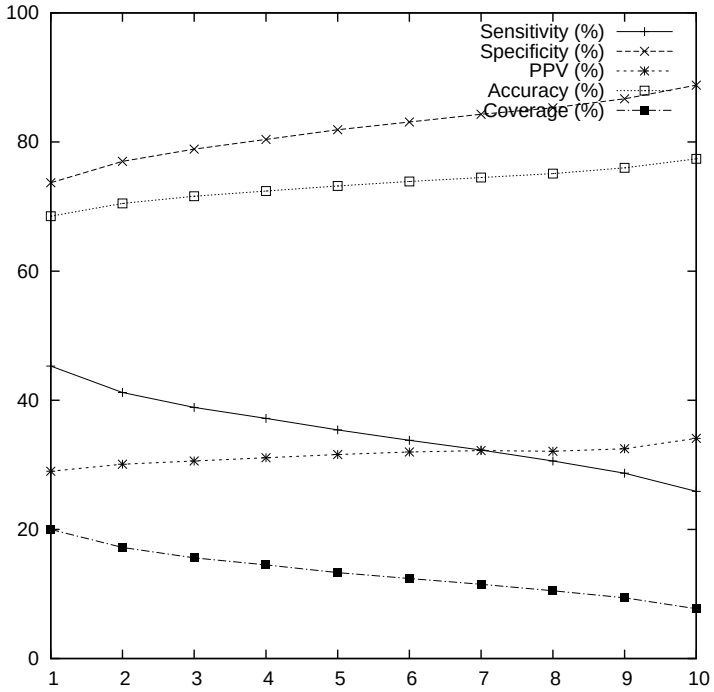

| Classe | Sen% | Spe% | PPV% | Cov% |
|---|---|---|---|---|
| A | 12.6 | 84.4 | 22.5 | 10.6 |
| AB | 15.9 | 86.6 | 28.0 | 9.6 |
| E | 49.3 | 79.3 | 40.8 | 14.8 |
| O | 33.0 | 86.4 | 29.1 | 11.2 |

(Haut) Résultats obtenus au moins $k = 7$ occurrences parmi 10 itérations.

(Gauche) Évolution des performances pour différents seuils $k$.

- Spécificité remarquable : Typiquement $\sim$80%
- Sensibilité variable : $>$20% nécessaires pour *diriger* MAXDO
- PPV stable sur $k$ : Proportion de *bons* résidus constante sur $k$
  $\Rightarrow$ Baisser $k \rightarrow 1$ pour les classes difficiles
- Couverture $<$20% : Restreindre les positions de MAXDO.

# JET : Résultats sur la base Mintseris



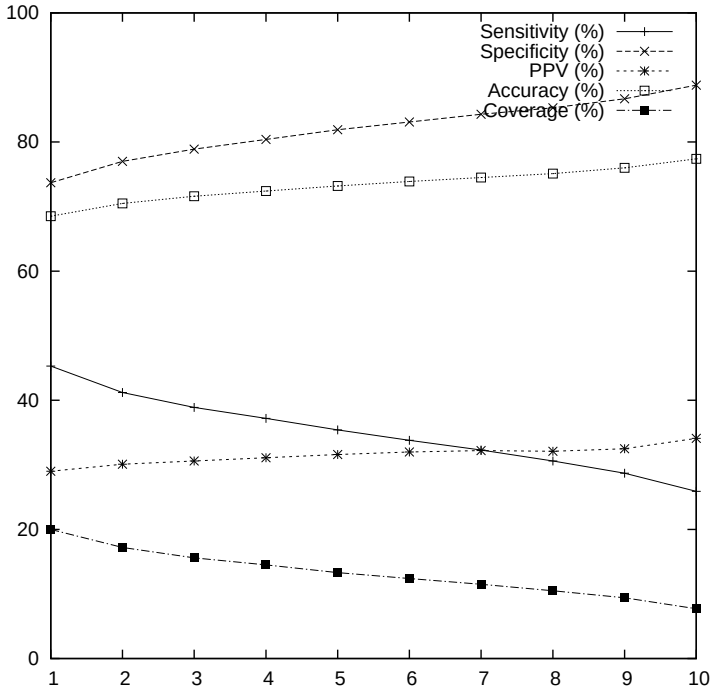

| Classe | Sen% | Spe% | PPV% | Cov% |
|---|---|---|---|---|
| A | 22.6 | 74.1 | 22.7 | 19.1 |
| AB | 26.9 | 76.6 | 25.9 | 19.1 |
| E | 49.3 | 79.3 | 40.8 | 14.8 |
| O | 33.0 | 86.4 | 29.1 | 11.2 |

(Haut) Résultats obtenus au moins $k = 7$ ou $k = 1$ occurrences parmi 10 itérations.

(Gauche) Évolution des performances pour différents seuils $k$.

- Spécificité remarquable : Typiquement $\sim$80%
- Sensibilité variable : >20% nécessaires pour *diriger* MAXDO
- PPV stable sur $k$ : Proportion de *bons* résidus constante sur $k$
  $\Rightarrow$ Baisser $k \rightarrow 1$ pour les classes difficiles
- Couverture <20% : Restreindre les positions de MAXDO.

# JET : Résultats sur la base Mintseris



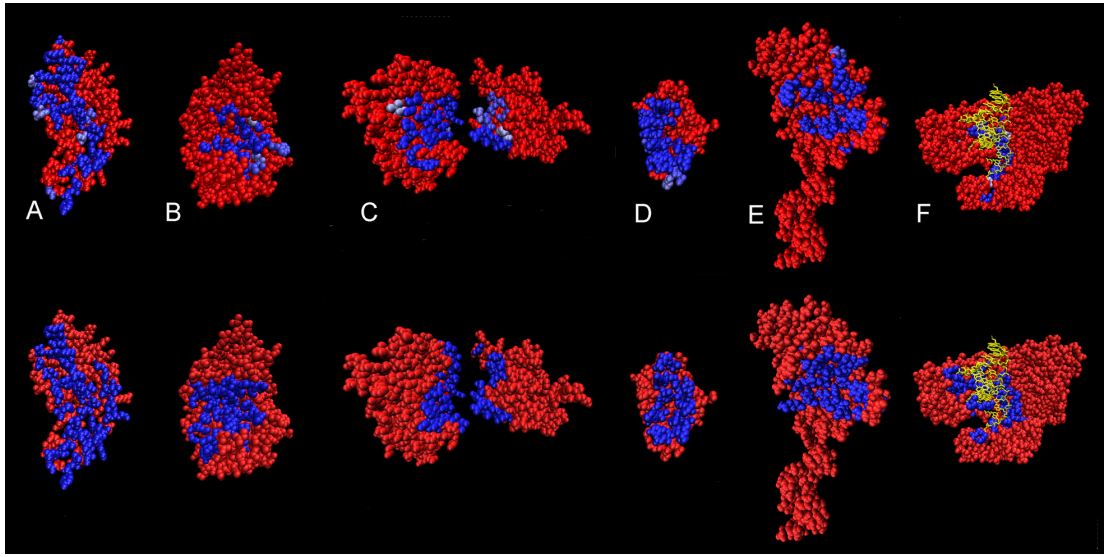

| Classe | Sen% | Spe% | PPV% | Cov% |
|---|---|---|---|---|
| A | 22.6 | 74.1 | 22.7 | 19.1 |
| AB | 26.9 | 76.6 | 25.9 | 19.1 |
| E | 49.3 | 79.3 | 40.8 | 14.8 |
| O | 33.0 | 86.4 | 29.1 | 11.2 |

(Haut) Résultats obtenus au moins $k = 7$ ou $k = 1$ occurrences parmi 10 itérations.

(Gauche) Évolution des performances pour différents seuils $k$.

- Spécificité remarquable : Typiquement $\sim$80%
- Sensibilité variable : $>$20% nécessaires pour *diriger* MAXDO
- PPV stable sur $k$ : Proportion de *bons* résidus constante sur $k$
  $\Rightarrow$ Baisser $k \rightarrow 1$ pour les classes difficiles
- Couverture $<$20% : Restreindre les positions de MAXDO.

# Croisement JET/MAXDO




Prédictions JET (Haut) vs interfaces expérimentales (Bas)

En résumé :

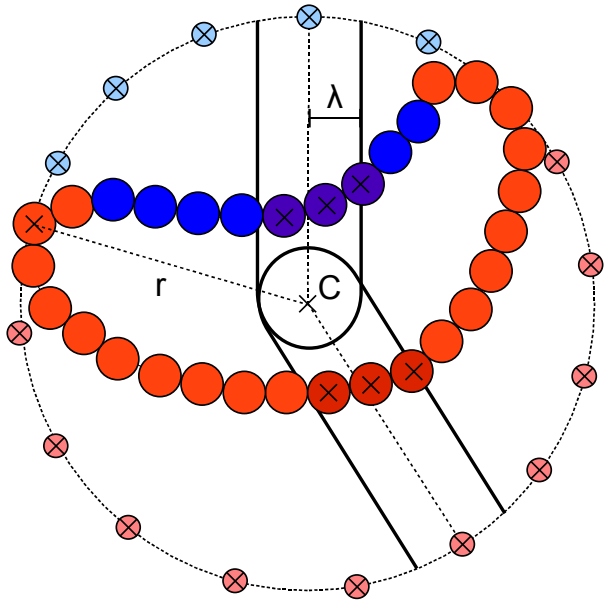
- MAXDO prédit correctement la fixation de partenaires
- Couple énergie d'interaction/Résidus à l'interface discriminant
- JET prédit correctement 20-30% les interfaces (+connexes)

Croiser JET et MAXDO afin de :

A Restreindre les positions explorées (Limiter calculs inutiles)

B Remplacer les couvertures expérimentales par celles prédites par JET

# Croisement JET/MAXDO : Méthode de restriction




Residues:

- JET-predicted
- JET-unpredicted

MAXDO initial docking sites:

- Selected by JET
- Discarded by JET

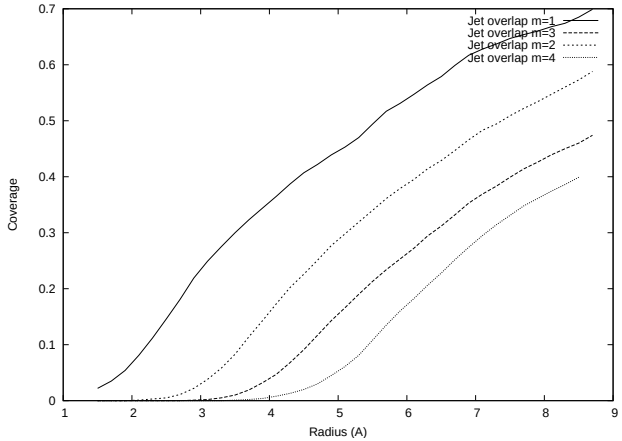
On se limite aux positions *projetées* à moins de $\lambda$ Å de $>m$ résidus JET.

Alt. : On garde une position $P$ si plus de $m$ résidus JET appartiennent au demi-tube de rayon $\lambda$ basé sur la demi-droite $[CS)$.

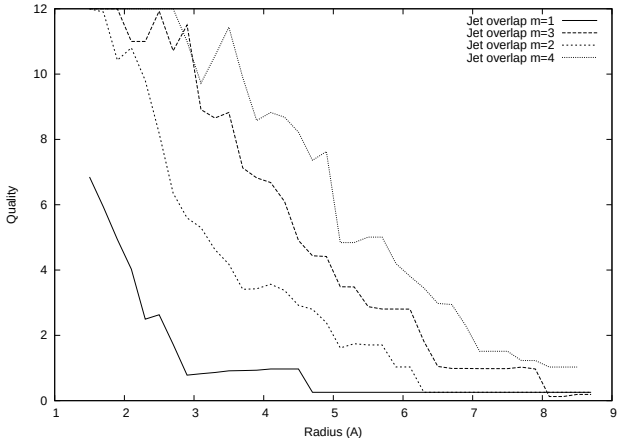
Score de dégradation $\mathcal{D}$ pour le calibrage des $(m, \lambda)$ :

$$\mathcal{D}_{m,\lambda} = \sum_{(\mathcal{P},\mathcal{P}^{-1})\in \text{Part.}} (1 - NII_{\mathcal{P},\mathcal{P}^{-1}})^2$$

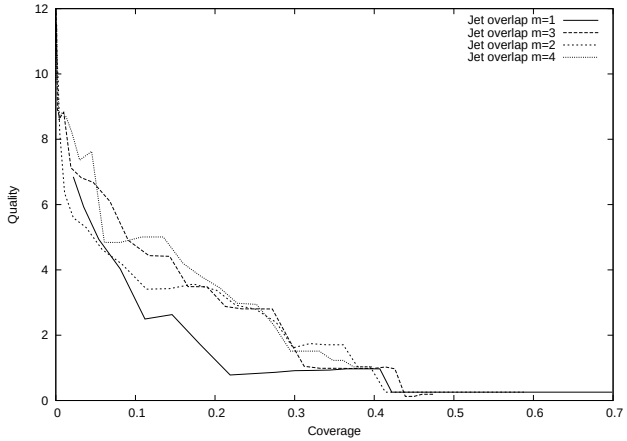
# Résultats restriction



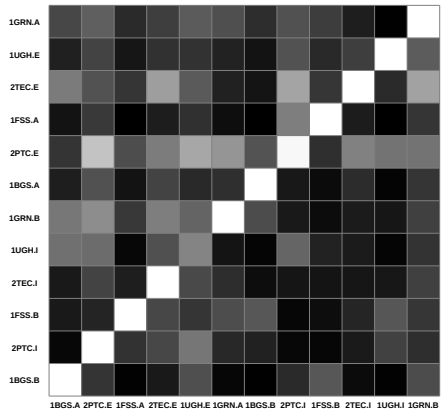

%Pos. explorées pour $\lambda$ croissant




Dégradation pour $\lambda$ croissant




Dégradation sur %pos. explorées




Meilleur compromis : $\lambda$=2.9 Å, $m$=1
$\Rightarrow \mathcal{D}_{m,\lambda} = 0.782$ et %Pos = 22%

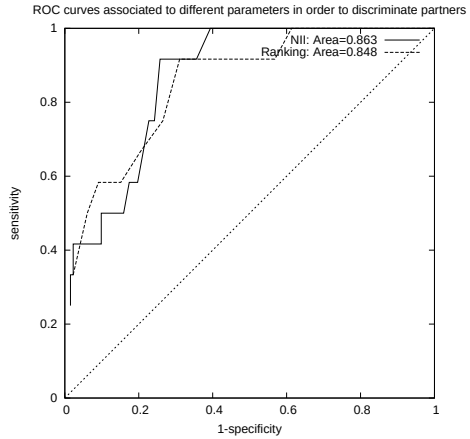
$\Rightarrow$ On divise par 5 les positions MAXDO sans perte critique de sensibilité

# Résultats prédiction

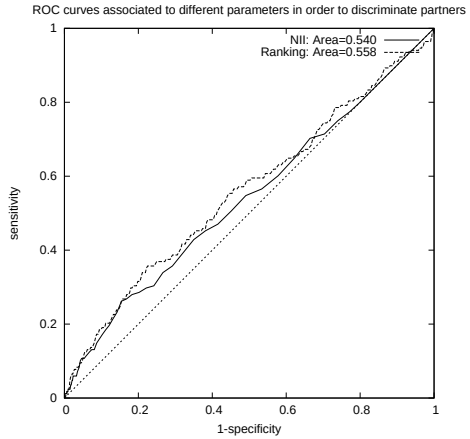
Méthode : Recalcul des $F_{i,\mathcal{P}}$ intervenant dans le calcul des *NII*

$$F_{i,\mathcal{P}} = \frac{|\mathcal{J}_\mathcal{P} \cap \mathcal{I}_i|}{|\mathcal{J}_\mathcal{P}|}$$

où $J_\mathcal{P}$ : Résidus prédits par JET pour $\mathcal{P}$ ; $\mathcal{I}_i$ : Résidus de $i$ selon MAXDO.




Base réduite de 12 protéines
AUC : 85%




Base Mintseris de 168 protéines
AUC : 56%

Signal réél, mais *n*oyé quand trop de partenaires potentiels

En cours :

- Utiliser d'autres paramètres pour la discrimination
- Réaliser des analyses partielles basées sur annotations

# WCG Phase 2 : Base de données

Constitution d'une base de données composée de :

- $\sim$90 prot. étudiées par P. Guicheney (Institut de Myologie, Paris)
- $\sim$160 prot. issues de modélisation (IGBMC, Strasbourg)
- $\sim$1560 prot. PDB liées à la *myopathie*
- $\sim$7160 prot. PDB liées aux mécanismes cardiaques ou cérébraux

$\Rightarrow$ $\sim$14000 chaînes protéiques

Filtrage nécessaire :

- Duplicata exacts (RMN, ...) $\Rightarrow$ $\sim$11300 chaînes
- Redondance structurale (Base ASTRAL/SCOP) $\Rightarrow$ $\sim$6500 chaînes
- Redondance séquentielle (PDBSelect95) $\Rightarrow$ 3200 chaînes
- Petites chaines non-classées (<15 res.) $\Rightarrow$ 2246 chaînes

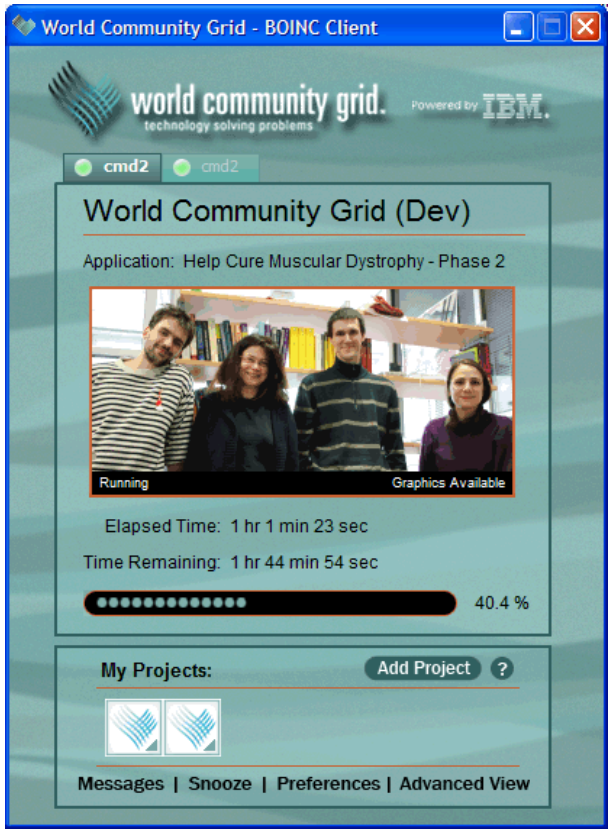
Réannotation JET par croisement avec annotations GO

# WCG Phase 2 : Lancement

Déc 08 Prédiction JET des interfaces

Fév 09 Annotation GO/Extension des interfaces prédites

Avr 09 Lancement sur la grille communautaire IBM du programme



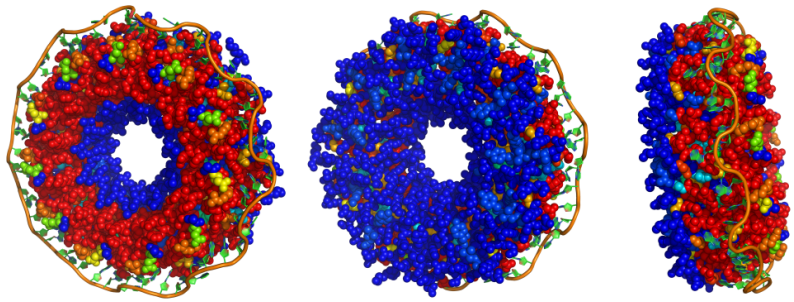

*Help cure muscular dystrophy*

Stats :

- 2466 753 paires de protéines
- 913 627 781 945 positions
- JET économise 87% de l'exploration
- 11.43 fois plus de calculs que la phase 1
  $\sim$ 880 siècles !
- Mais 441 737 inscrits à la WCG
- Cycles disponibles sur 1 242 326 machines
- $\sim$35 000 CPU à tout instant

# Perspectives

- Cross-docking localisé (Colocalisation, fonction)
  $\Rightarrow$ Classe GO vs performance du pipeline
- Intégrer davantage de paramètres (Séquences, géométrie, aire)
  $\Rightarrow$ Support Vector Machine
  ($\exists$ Outil générique classification GO/Données expérimentales ? ? ?)
- JET : Protéines/RNA et Protéines/DNA



- Nombreux problèmes techniques :
  - Stockage : 365 Go envoyés/1.5 To reçus (Compression $\sim \times 30$)
    $\Rightarrow$ Comment exploiter les données ?
  - Quelle granularité pour les données analysées ?
  - Codage binaire *ad hoc* ...
- Projeter les prédictions sur des protéines similaires

# Collaborateurs/Agences

LBT-IBPC :
Richard Lavery
Sophie Sacquin-Mora

Institut de Myologie :
Pascale Guicheney

INSERM-UPMC :
Alessandra Carbone
Stefan Engelen
Ladislas Trojan

Graal LIP-ENS Lyon :
Raphael Bolze
Michaël Heymann
Nicolas Bard




# Questions ?

Charles C. Ebbets, 1932

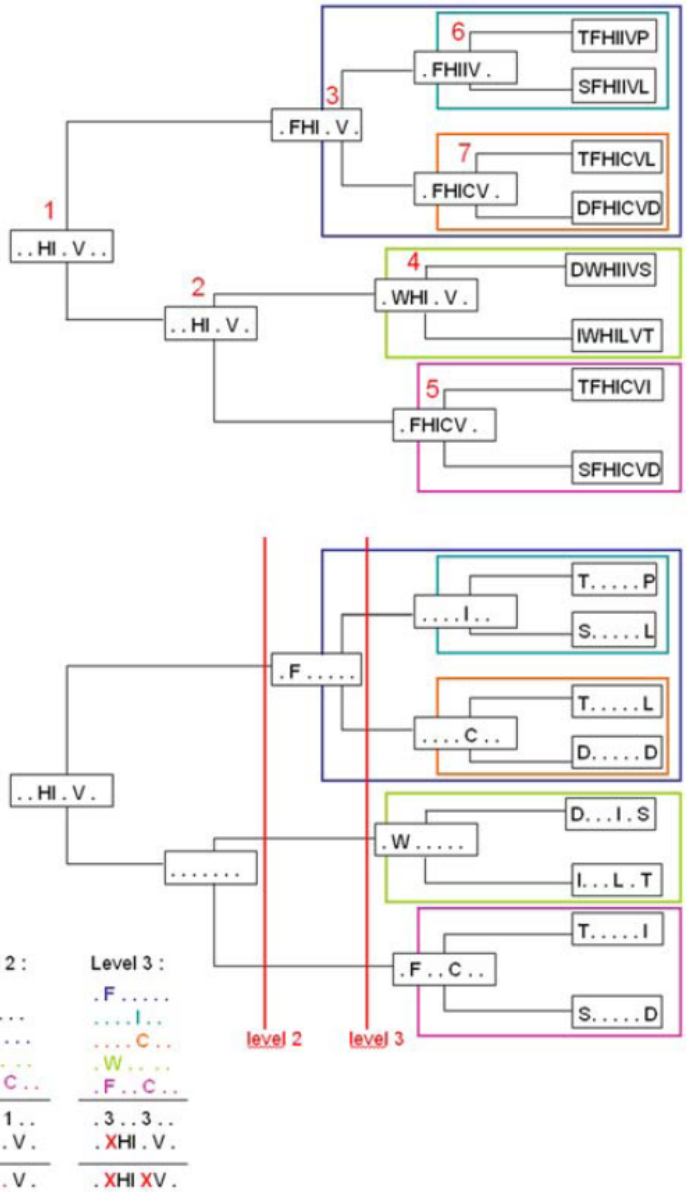
# Exemple JET

Trace relative $d_j$ du résidu $j$ :

$$d_j = \frac{1}{M_j} \sum_{i=1}^{M_j} \frac{L_{T_i} - l_{j,T_i}}{T_i}$$

Où $M_j$ : #Arbres où résidu $j$ apparaît ; $L_{T_i}$ : Profondeur max de $T_i$ ; $l_{j,T_i}$ : Niveau où res. $j$ apparaît dans $T_i$.

Trace $trace(j)$ : Moyenne pondérée des traces relatives de $j$ et des résidus à distance <5Å.




Retour