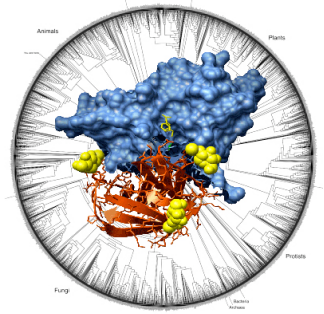
Information dans les séquences biologiques

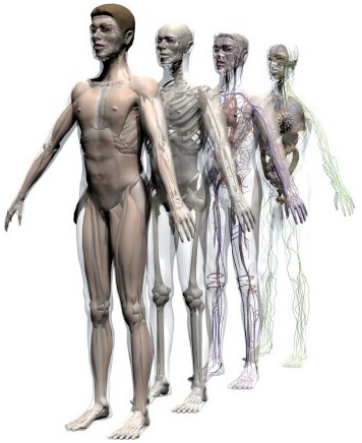


Alessandra Carbone - Université Pierre et Marie Curie & CNRS
Paris, France

Chercher les différences dans nos génomes

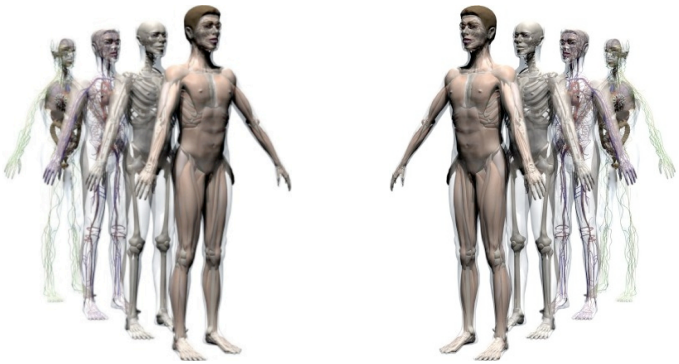
… dans les populations et les individus

Différences : Single Nucleotide Polymorphisms (SNPs)



10^9 nucléotides dans l'ADN et 10^6 SNPs

En comparant deux individus (de continents différents)



En moyenne, sur 1,331 bases, 1,330 bases seront identiques et une base (le site SNP) sera différente

SNPs nous rendent génétiquement uniques

différences physiques
prédisposition à certaines maladies
réponse aux médicaments

Et pleins d'autres attributs humains sont le résultats des variations de SNPs.

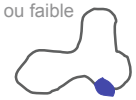
La forte majorité de SNPs n'a pas d'effet sur la fonction protéique
Ils constituent des "variations silentieuse". Parfois, pour qq SNPs, une mutation cause des problèmes de santé majeures.

	Séquence ADN
Individu 1	A A A T T T
Individu 2	A A T T T T
Individu 3	A A C T T T

Protéine normale

Certaines variations d'ADN ont aucun effet négatif

Fonctionnement protéique absent ou faible



Certaines variations amène à des maladies ou augmentent la prédisposition à une maladie

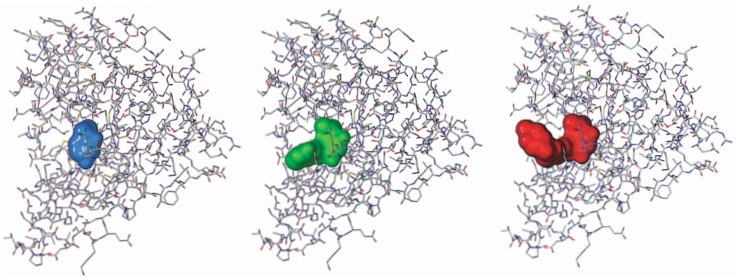
Mutations dans un seul gène

mucoviscidose, gène CFTR dans le chromosome 7

Mutations dans plusieurs gènes

Maladies cardiaques et cancer

Mutations dans les séquences protéiques :
<u>malfonctionnement dans l'interaction avec d'autres protéines</u>

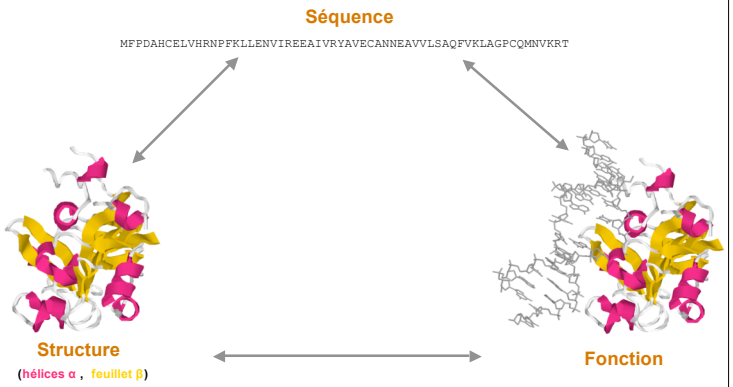


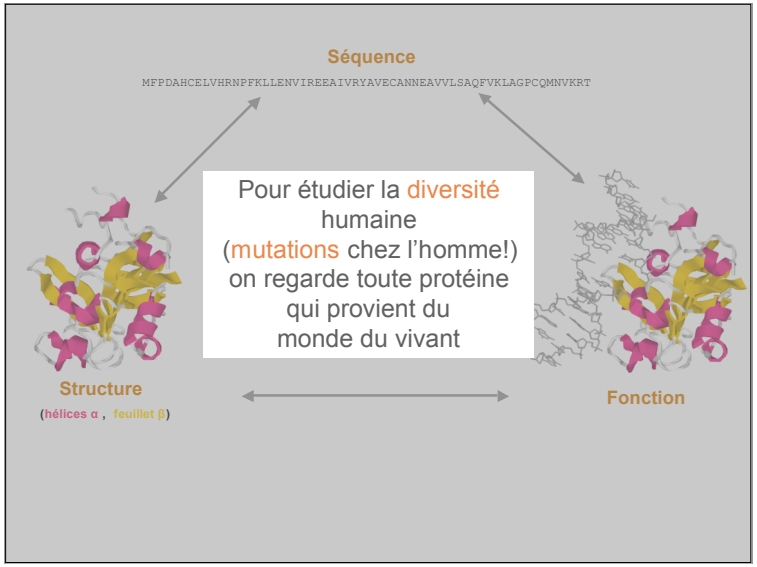
Les mutations peuvent induire une **instabilité** du complexe protéique

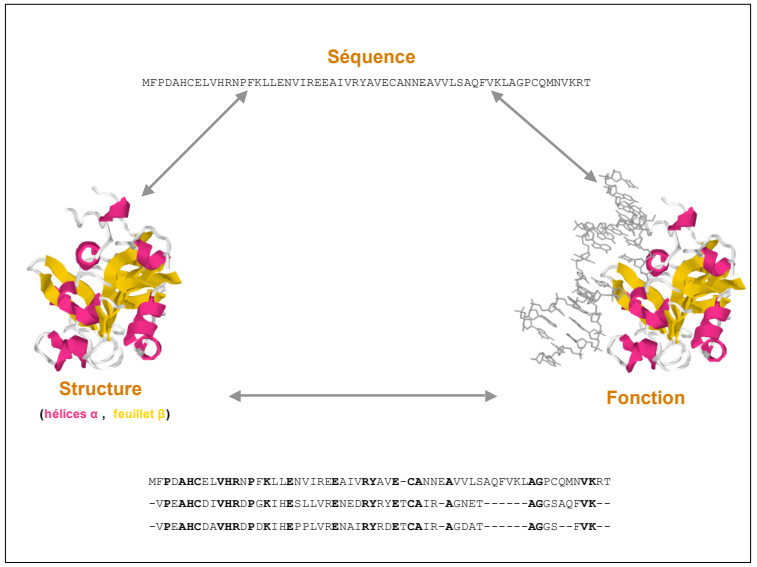
Plusieurs effets sont dûs aux mutations.
Un exemple complexe: la dystrophie musculaire

Légérement réduite

Normale: présente autour des fibres musculaires

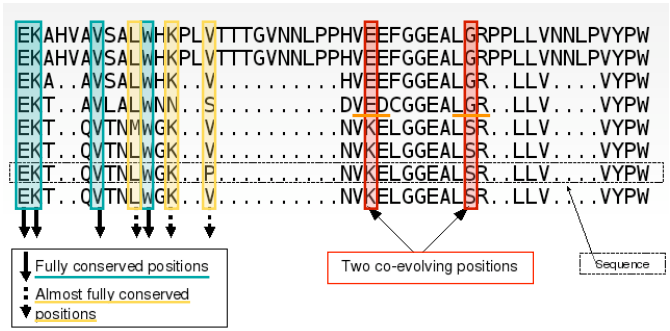


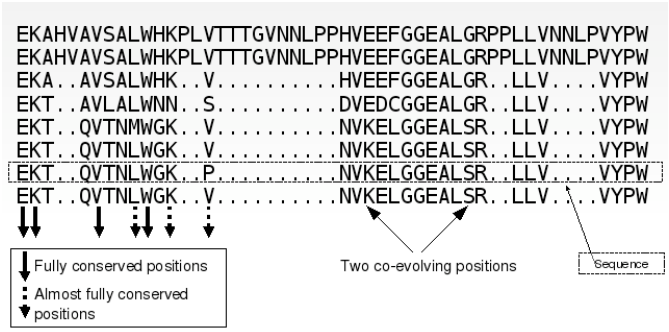




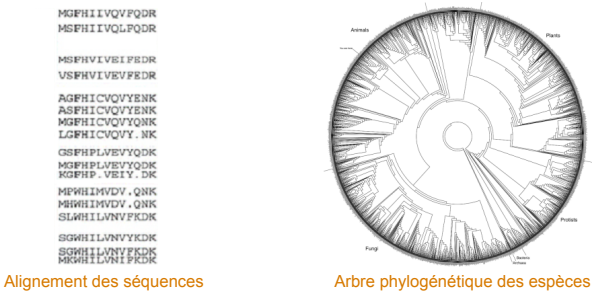
Deux signaux de codification de l'information :

- conservation d'une position
- co-évolution de plusieurs positions





Conservation



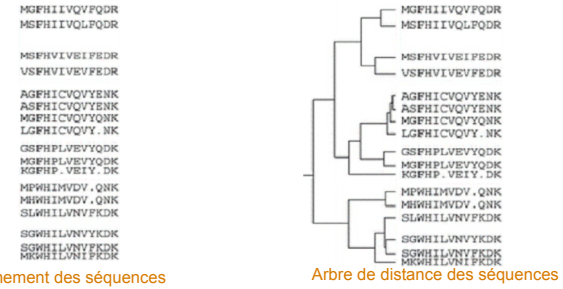
Alignement des séquences

Arbre phylogénétique des espèces

Contenu d'information (IC)

$$H(i) = - \sum_x p(x,i) \log_{20} p(x,i)$$

où p(x,i) est la fréquence d'occurrence de l'acide-aminé x à la position i



Alignement des séquences

Arbre de distance des séquences

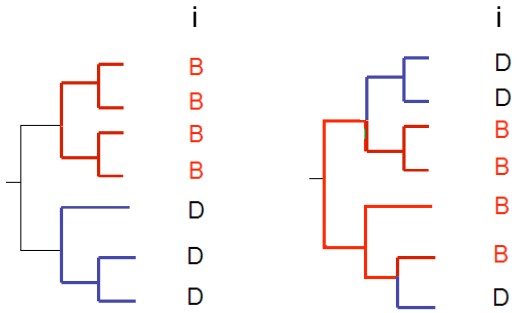
Contenu d'information (IC)

$$H(i) = - \sum_x p(x,i) \log_{20} p(x,i)$$

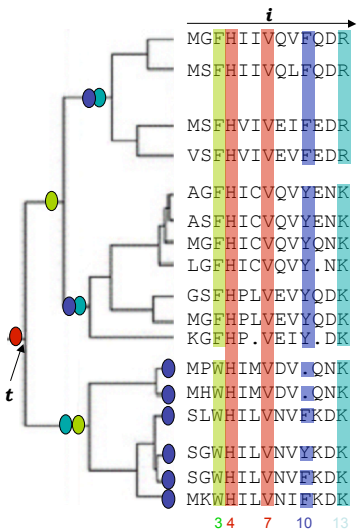
où p(x,i) est la fréquence d'occurrence de l'acide-aminé x à la position i

i	i
B	D
B	D
B	B
B	B
D	B
D	B
D	D

Même contenu d'information



Même contenu d'information



Pour quantifier l'histoire évolutive on considère la topologie de l'arbre

On compte, par exemple, le nombre de sous-arbres maximaux qui conservent une position

R(4,t)=R(7,t) = 1

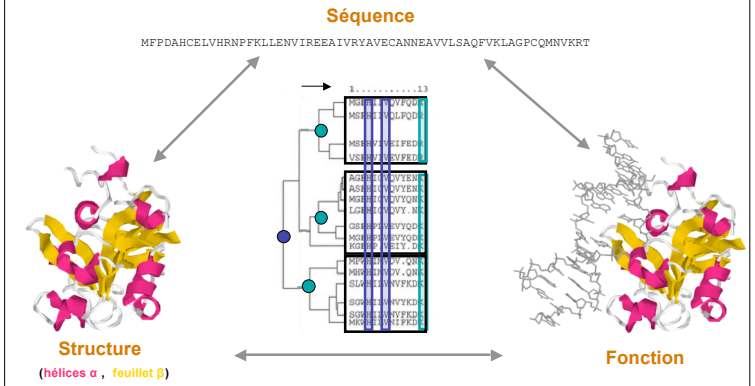
R(3,t) = 2

R(13,t) = 3

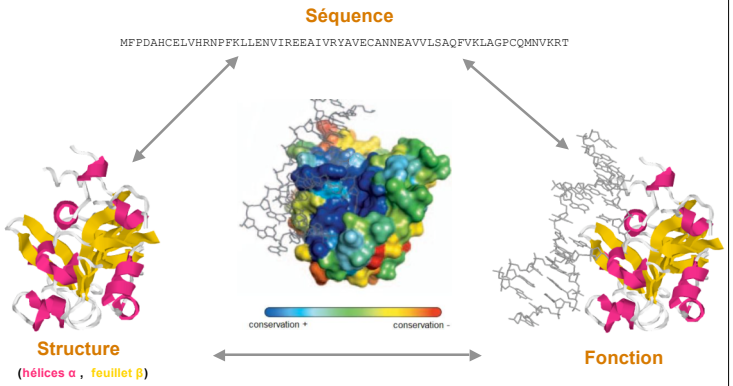
...

R(10,t) = 8

...

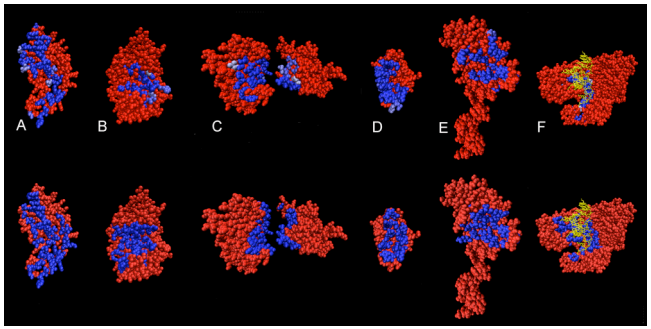


les résidus conservés se regroupent aux interfaces d'interaction



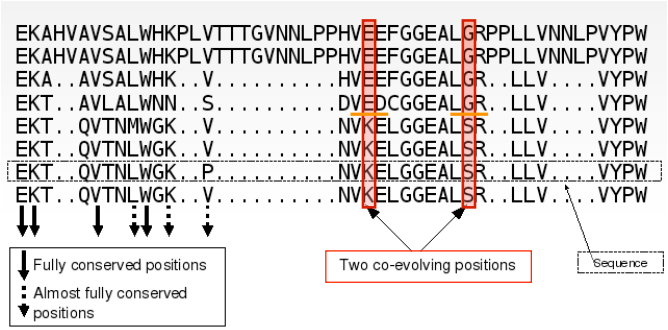
Proximité structurale
(O.Lichtarge 1996)

Analyses automatiques à large échelle



Prédictions (haut) et interfaces expérimentales (bas)

Prédictions réalisées en utilisant une variation de « sous-arbre maximaux »



Co-évolution

Information Mutuelle

$$H(i,j) = - \sum_{x,y} p(x,i,y,j) \log_{20} p(x,i,y,j)$$

$$MI(i,j) = H(i) + H(j) - H(i,j)$$

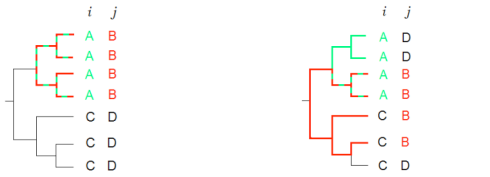
Des valeurs larges reflètent une plus forte interdépendances entre positions

Problèmes avec les séquences biologiques:

- Bruit de fond: nombre minimum de séquences pour garantir une séparation du signal avec le fond
- Composante phylogénétique: toutes paires de positions ont associées un bruit provenant d'une pauvre diversités de l'échantillon des séquences.

On regarde de près la combinatoire des sousarbres

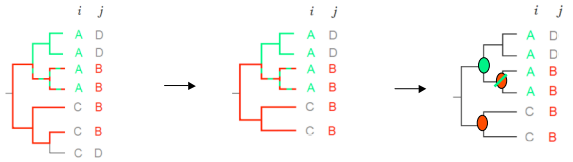
Si *i* et *j* sont des positions co-évoluées,
le sous-arbre conservant le résidu A_i (resp. B_j) inclut le sous-arbre conservant le résidu co-évolué B_j (resp. A_i)



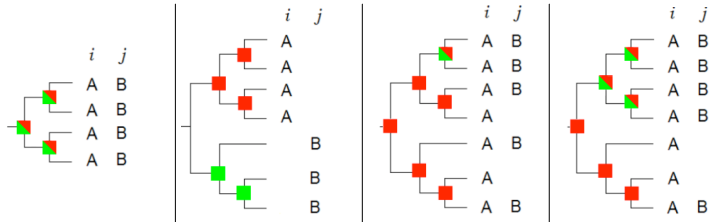
Résidus co-évolués Résidus non co-évolués

Évaluation de la co-évolution des paires i,j

Pour chaque couple de résidus A_i,B_j ⇨ Arbre spécifique du couple A_i,B_j



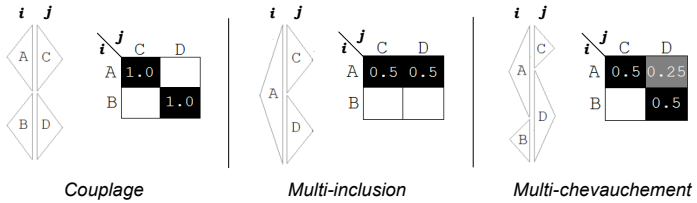
Score de correspondance $C(A_i,B_j)=$



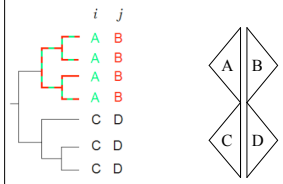
$C(Ai,Bj)=1$	$C(Ai,Bj)=0$	$C(Ai,Bj)=0.16$	$C(Ai,Bj)=0.5$
Inclusion parfaite	Exclusion parfaite	Chevauchement intermédiaire	

⇨ Matrice de correspondance $C_{i,j}$

De la matrice de correspondance au score de co-évolution

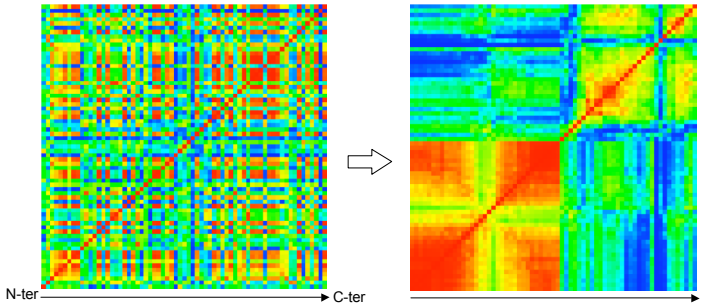


Couplage *Multi-inclusion* *Multi-chevauchement*



- Combinatoire des sous-arbres
- Motifs de matrice

Reconstruction des réseaux



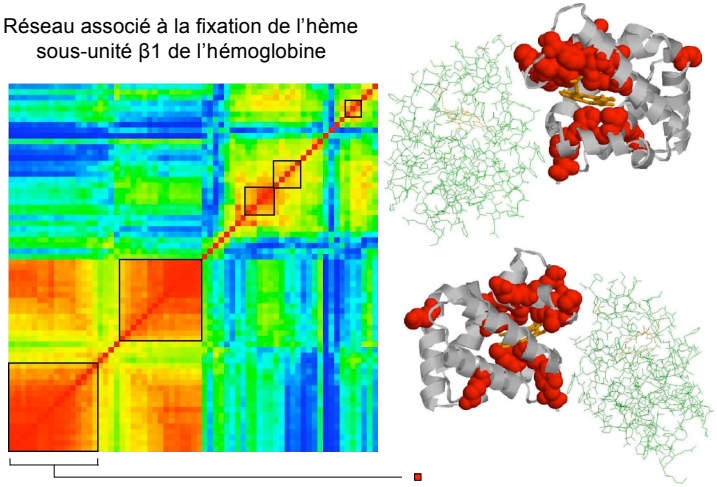
Ordre des positions dans la séquence

Ordre des positions après *clusterisation*

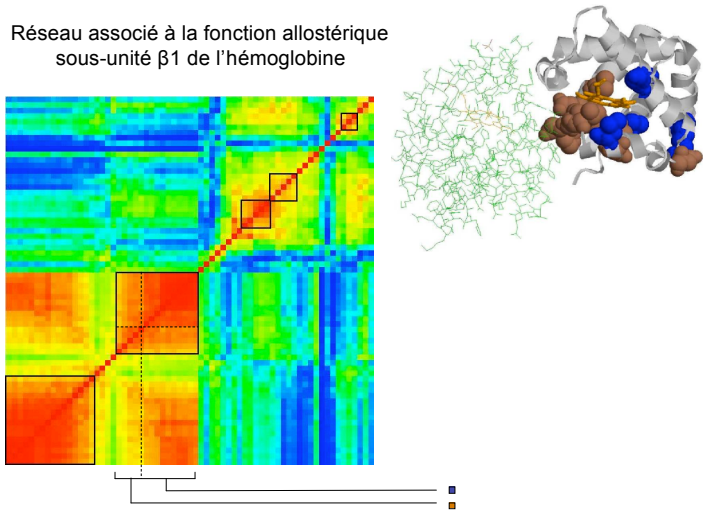
Regroupement des positions obtenant ensemble leur meilleur score de co-évolution

La famille des hémoglobines

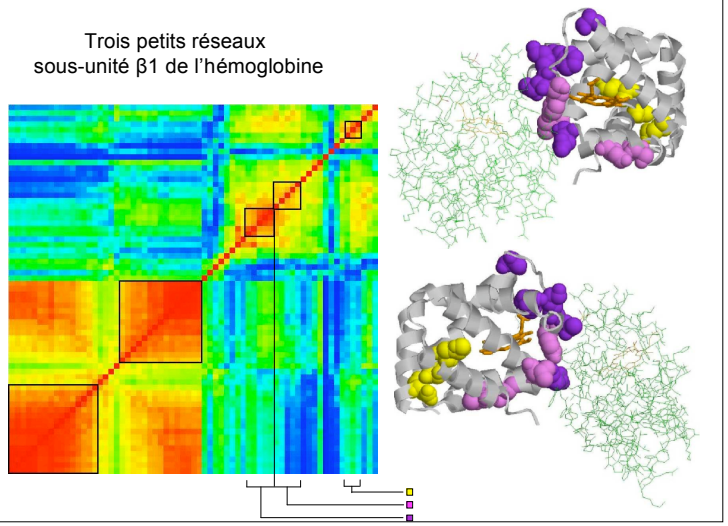
Réseau associé à la fixation de l'hème
sous-unité β1 de l'hémoglobine



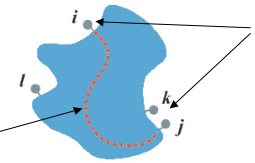
Réseau associé à la fonction allostérique
sous-unité β1 de l'hémoglobine



Trois petits réseaux
sous-unité β1 de l'hémoglobine



Chemins de communication entre sites d'interaction:



Reconnaissance
de ligands

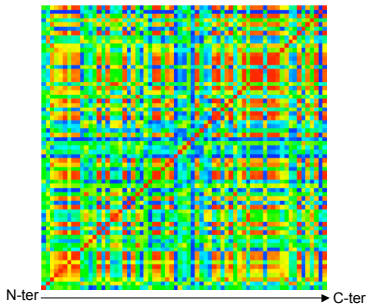
Communication
entre les sites

⇩

Contraintes
évolutives
mutuelles

- allostérie
- sites fonctionnels

Revenons à la reconstruction des réseaux



Ordre des positions dans la séquence

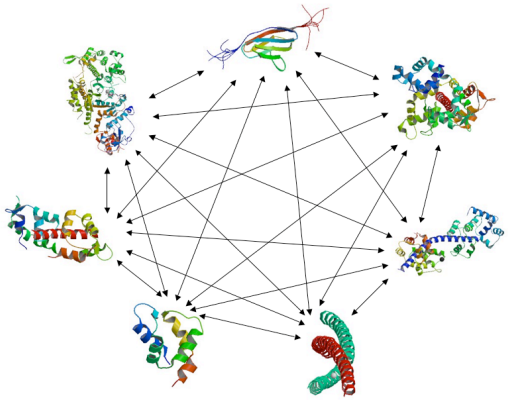
On regarde des **blocs de résidus adjacents**

- Repliement
- Identification d'entités ayant subis des pressions évolutive différentes

Travail en cours

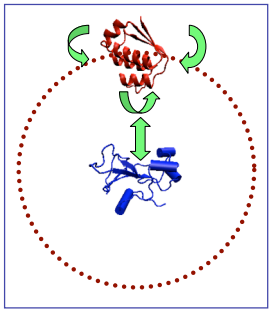
Une application bioinformatique

Help Cure Muscular Dystrophy



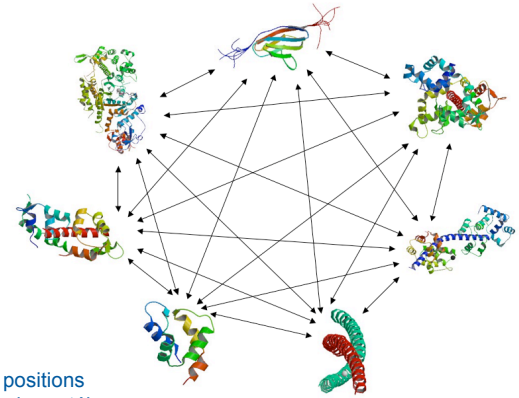
Qui sont les partenaires dans la cellule ?

Modélisation moléculaire : docking protéine-protéine



Exploration de la surface d'interaction : 100.000-500.000 positions différentes du ligand (1 pos. par 10Å^2 & 210 orientations du ligand par position)

15sec par minimization d'énérgie energy minimization calculation on 2GHz processor



200 000 positions
par paire de protéines
~14 siècles

150 x 150 = 22.500

1000 x 1000 = 1.000.000

beaucoup trop nombreuses

Toutes les positions peuvent être calculées indépendamment l'une de l'autre et le parallélisme peut être employé

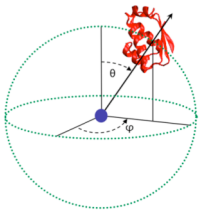
100 seconds CPU par position (sur grande grille)
~80 siècles

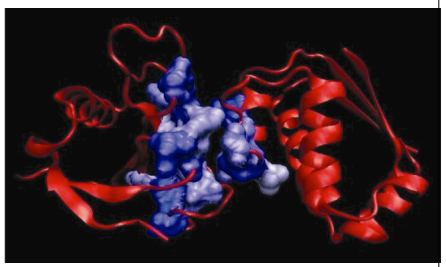
Grande grille d'ordinateurs

www.worldcommunitygrid.org

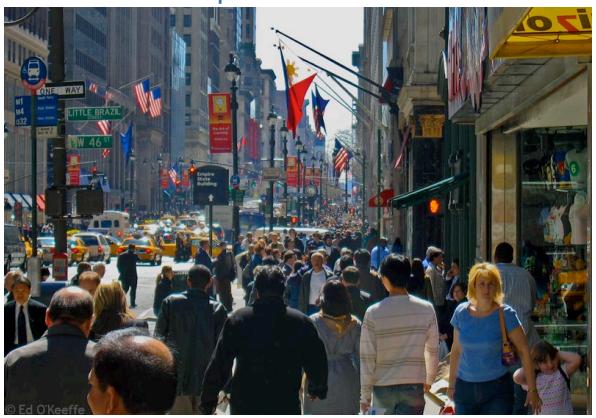
7 months with 6000-8000 machines

Prédiction des sites d'interaction

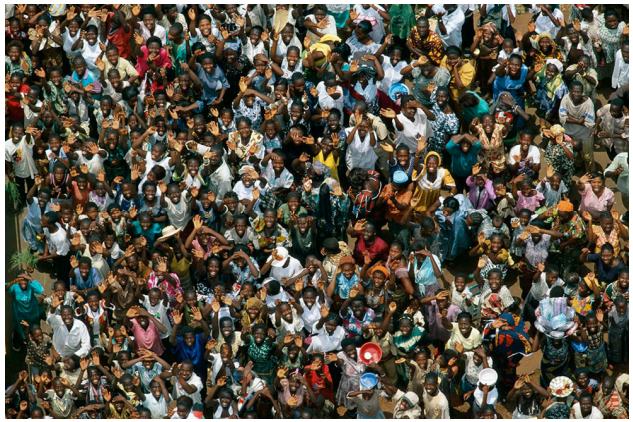




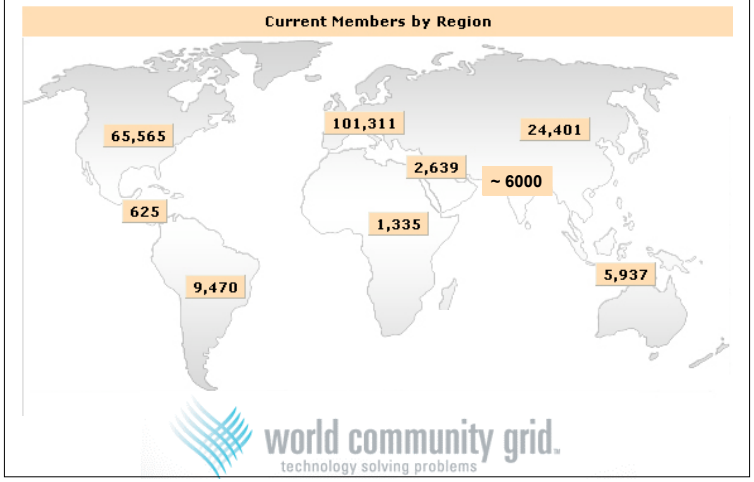
Mai 2009, HCMD a été lancée sur WCG :
2200 protéines humaines



fin attendue en juin 2011



Inscrivez vous sur www.worldcommunitygrid.org
et aidez les projets scientifiques !



	Help Cure Muscular Dystrophy	UPMC, France
	Help Cure Childhood Cancer	Ciba University, Japan
	Nutritious rice for the world	University of Washington US
	Human Proteome Folding	New York University US
	Help Conquer Cancer	Ontario Cancer Institute Canada
	FightAids@Home	Scripps Research Institute US

Co-évolution

Julie Baussand
EMBO postdoctoral fellow
Mathematical Biology Lab
National Institute for Medical Research
London, England

Linda Dib
Doctorande, UPMC

Conservation : HCMD project

Stefan Engelen (Génoscope)
Yann Ponty (CNRS-LIX)
Sophie Sacquin Mora (CNRS-IBPC)

IBCP (Lyon) & IBPC (Paris):
Richard Lavery
Sophie Sacquin-Mora

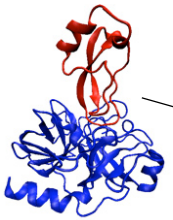
CNRS-UPMC:
Stefan Engelen
Yann Ponty
Ladislas Trojan

INSERM - Institut de Myologie:
Pascale Guichenay

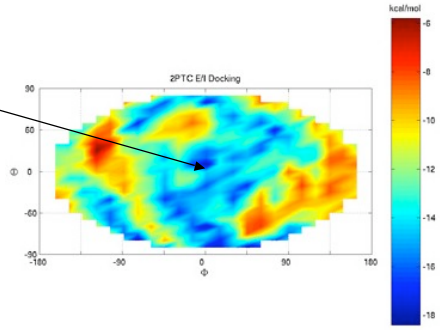
LIP (Lyon)
Raphael Bolze
Michaël Heymann
Nicolas Bard

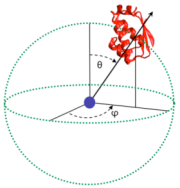
WCG:
Bill Bovermann
WCG team

Molecular docking results



Crystallographic complex





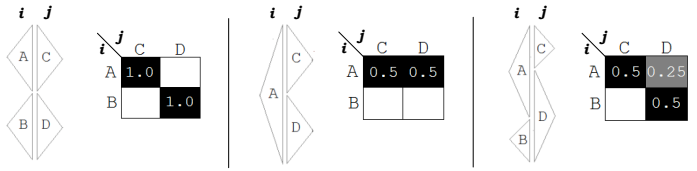
Potential map obtained after docking
(angles Θ et Φ determine the position of the ligand around the receptor)

Docking searches for optimal geometries of the interaction between two proteins

Scoring function : the free energy is minimized

De la matrice de correspondance au score de co-évolution

- Combinatoire des sous-arbres & Motifs de matrice



Couplage *Multi-inclusion* *Multi-chevauchement*

↳ *Cas idéal de co-évolution*

- Score de co-évolution

Max *Specificité* *Interference*

$Score(A \triangleright j) = AC \times \frac{AC}{AC + AD} \times \left(1 - \frac{AD}{BD + AD}\right)$

$Score(C \triangleright i) = CA \times \frac{CA}{CA + CB} \times \left(1 - \frac{BC}{BD + BC}\right)$

$CoE(i,j) = \sum score(X_i \triangleright j) + \sum score(X_j \triangleright i) = \{0,2\}$

Vue globale des réseaux
sous-unité β1 de l'hémoglobine

(5 réseaux → 29% des positions)

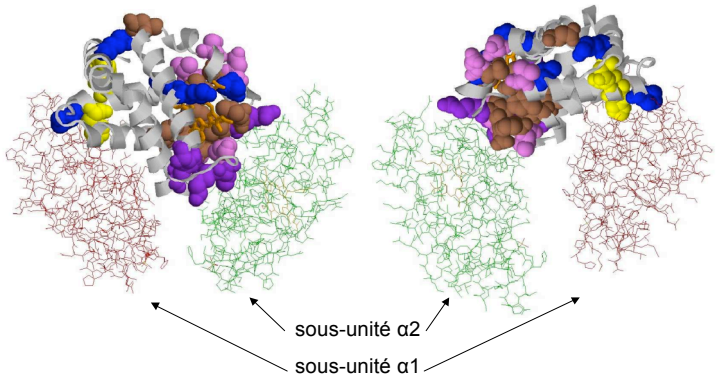
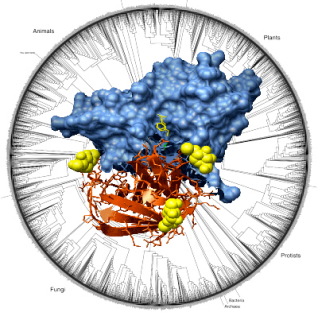


Table 2. JET, iJET, and ET evaluation on the Huang dataset.

	Sen	ScSen	PPV	ScPPV	Spe	ScSpe	Acc	ScAcc
No Clustering								
ET	35.4	14.0	50.4	1.6	83.9	5.1	68.4	7.8
JET – cons	34.7	13.3	50.3	1.6	84.5	5.8	67.6	7.5
With Clustering								
Homodimers								
ET	35.7	15.1	51.9	1.7	86.0	6.6	69.9	8.8
JET - cons	34.7	17.1	60.4	2.0	90.5	8.1	73.5	10.5
JET - cons+pc	37.9	18.9	61.2	2.1	89.5	8.4	73.1	11.0
iJET	**36.2**	**18.6**	**62.7**	**2.1**	**90.8**	**8.4**	**73.9**	**10.9**
Heterodimers								
ET	35.5	15.4	55.6	1.8	**86.3**	6.4	68.2	8.7
JET - cons	41.7	17.5	60.8	2.0	84.2	8.4	70.2	10.5
JET - cons+pc	47.9	21.1	59.6	1.9	83.8	10.5	71.2	12.6
iJET	**46.6**	**21.2**	**62.1**	**2.0**	**85.6**	**11.0**	**72.7**	**13.0**
Transients								
ET	29.7	10.5	50.7	1.5	**82.2**	1.4	63.8	5.7
JET - cons	38.9	12.9	57.5	1.8	77.8	3.9	67.0	6.6
JET - cons+pc	39.7	12.9	55.9	1.7	78.3	5.1	67.3	7.0
iJET	**37.5**	**14.0**	**59.1**	**1.9**	81.7	**5.1**	**68.2**	**7.6**
All Confounded								
ET	34.2	14.2	51.8	1.7	85.1	5.2	68.5	7.9
JET – cons	37.9	16.4	59.0	2.0	85.6	7.1	71.4	9.5
JET - cons+pc	41.6	18.4	58.4	1.9	84.9	8.1	71.6	10.5
iJET	**39.8**	**18.5**	**60.5**	**2.0**	**86.9**	**8.2**	**72.6**	**10.6**

Comment vaincre la dystrophie musculaire à l'aide de nos ordinateurs



Alessandra Carbone - Université Pierre et Marie Curie & CNRS
Paris, France

www.worldcommunitygrid.org