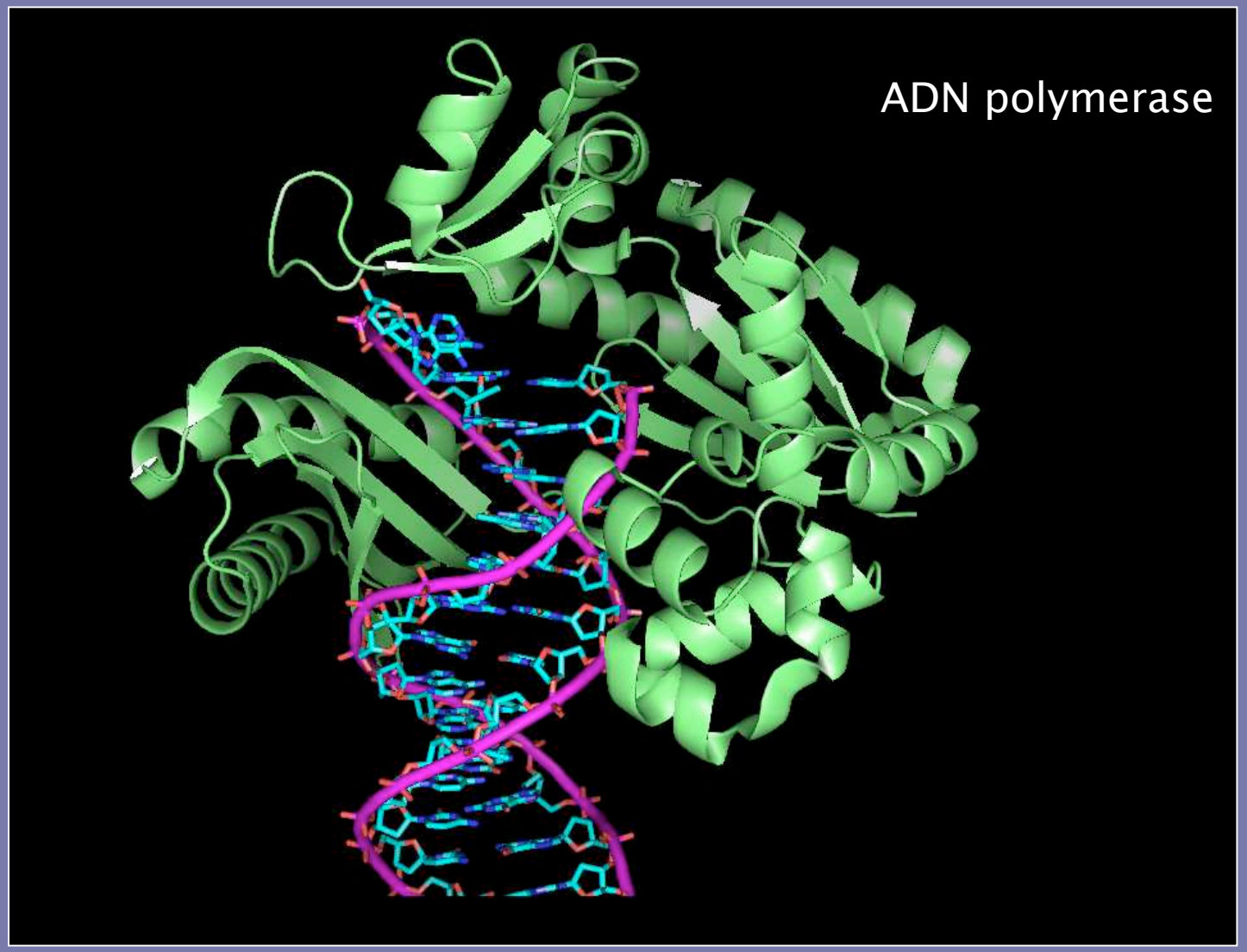
Utilisation des diagrammes de Voronoï et des algorithmes génétiques pour l'étude des complexes protéine-protéine.

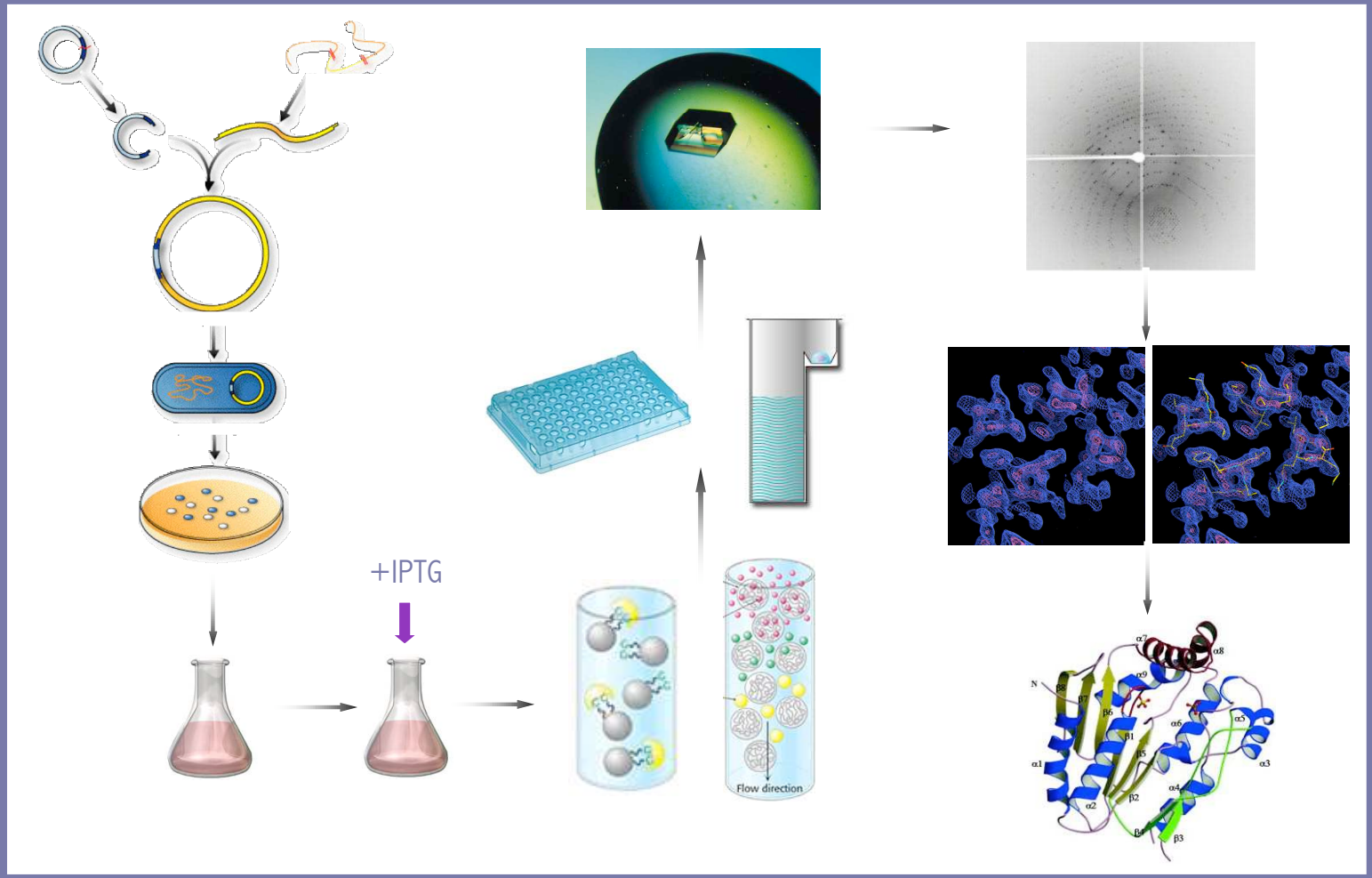
Anne Poupon
Biologie et Bioinformatique des Systèmes de Signalisation
INRA - Nouzilly
France

Pourquoi la structure 3D est-elle importante ?

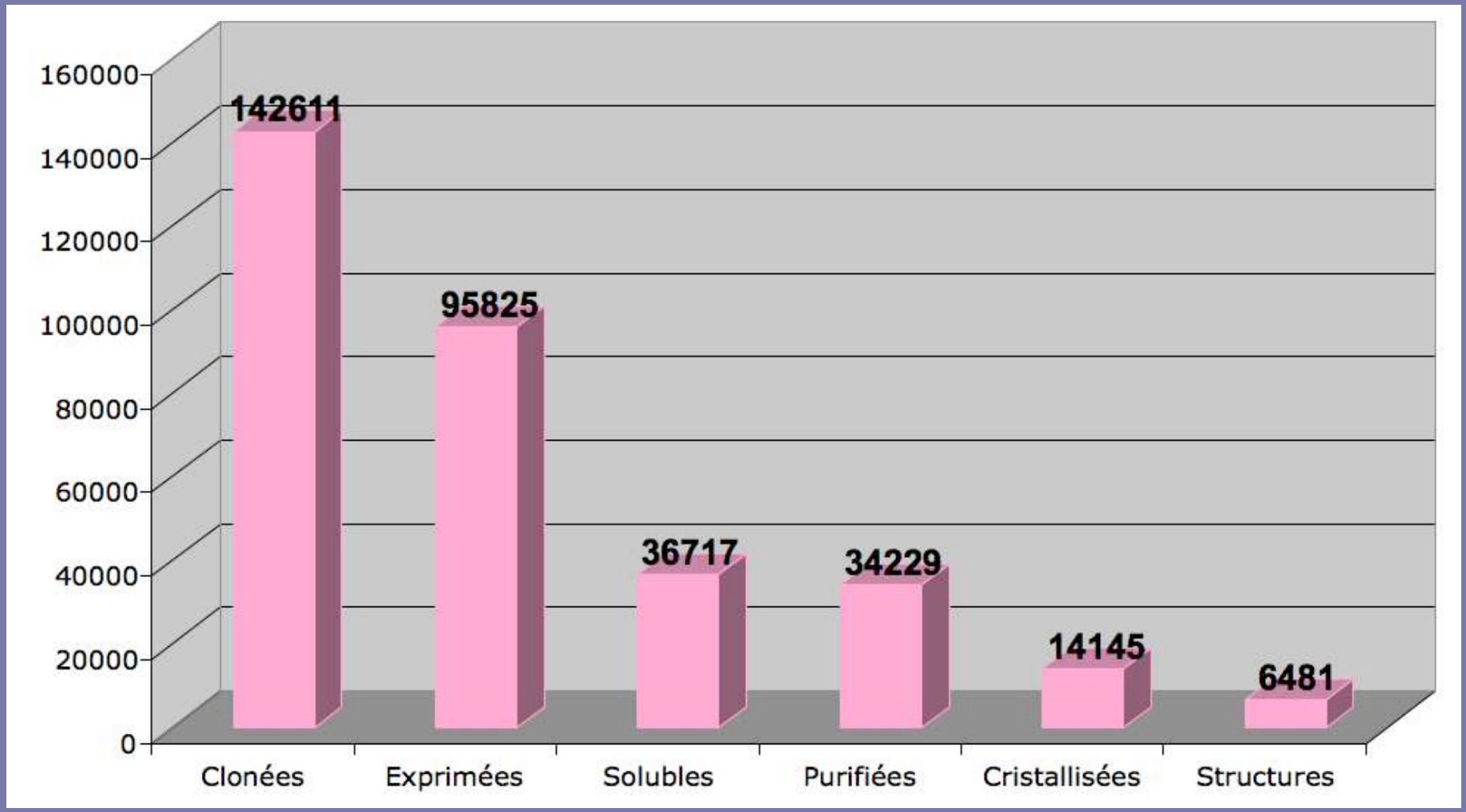
La fonction d'une protéine dépend de sa structure 3D.



Comment déterminer expérimentalement la structure 3D ?



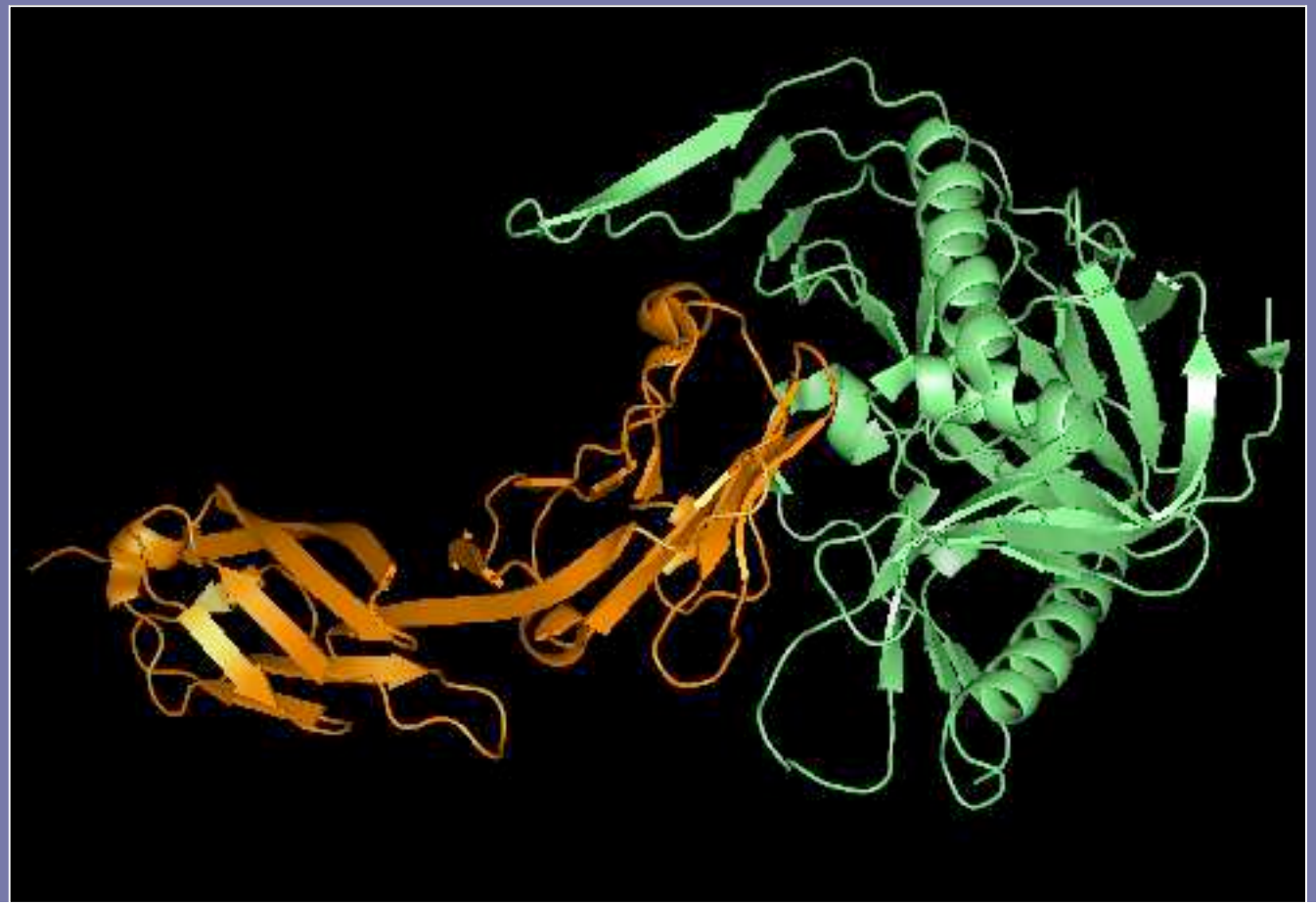
Comment déterminer expérimentalement la structure 3D ?



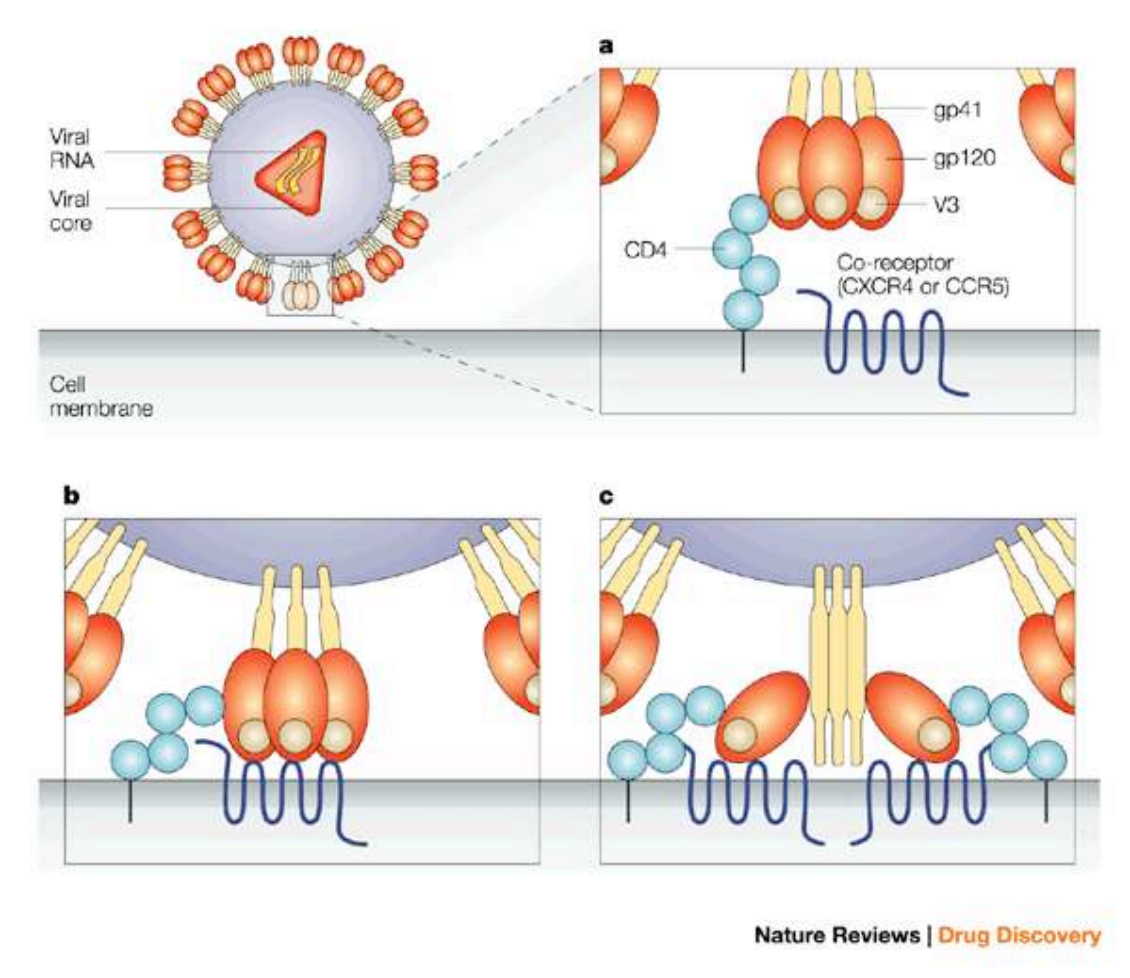
Quelques semaines, taux de succès : 4,5 %.

Les complexes protéine-protéine

Les complexes protéine-protéine sont partout ! Une large proportion des protéines accomplissent leur fonction à travers l'interaction avec d'autres macromolécules. Mais la détermination expérimentale de la structure 3D de l'assemblage est au mieux difficile !
Exemple : complexe GP120 (HIV) – CD4 (surface des lymphocytes T4)

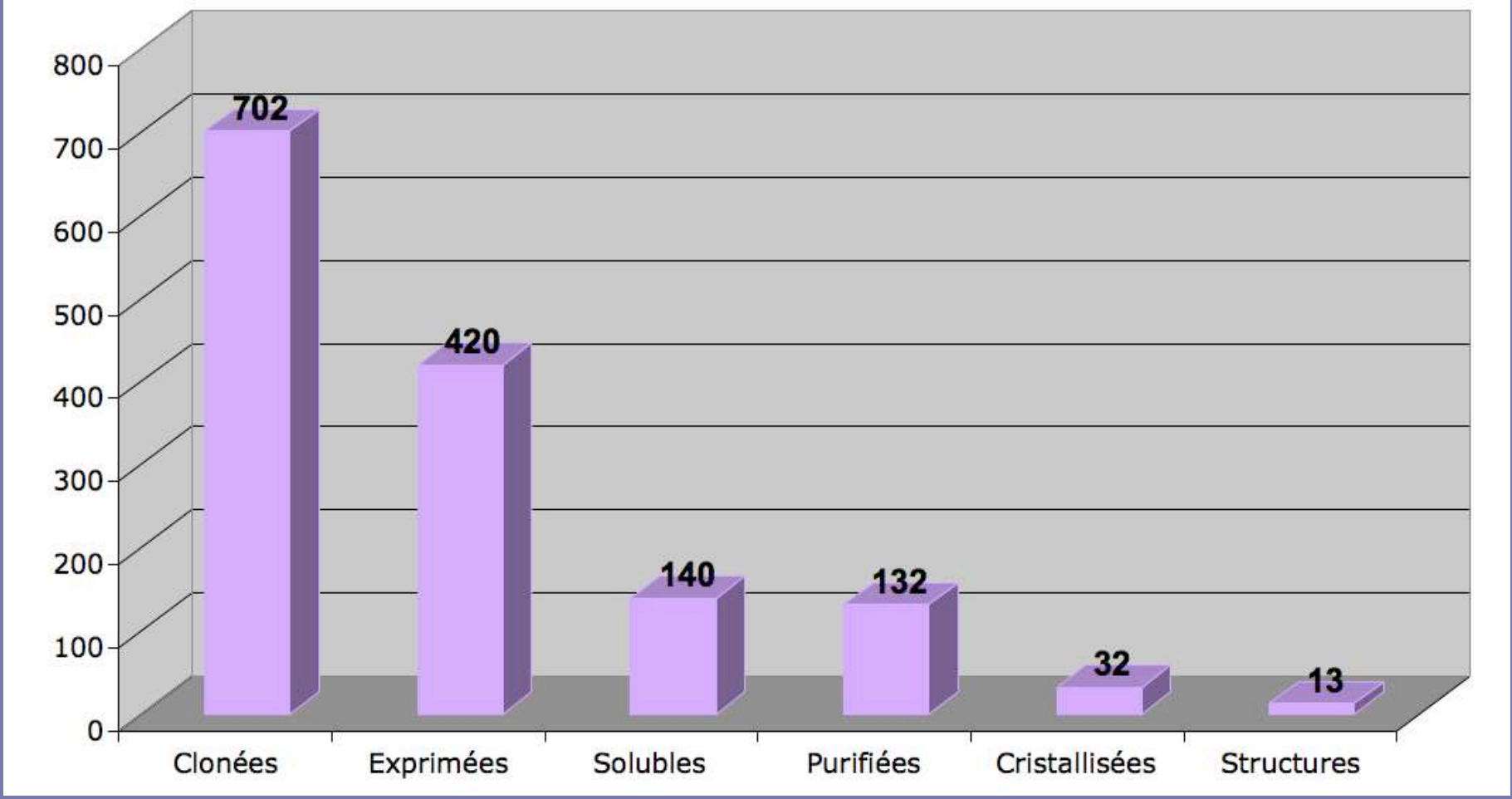


Les complexes protéine-protéine



Si on pouvait empêcher l'interaction entre GP120 et CD4, il serait peut-être possible d'empêcher la pénétration du virus dans le lymphocyte.

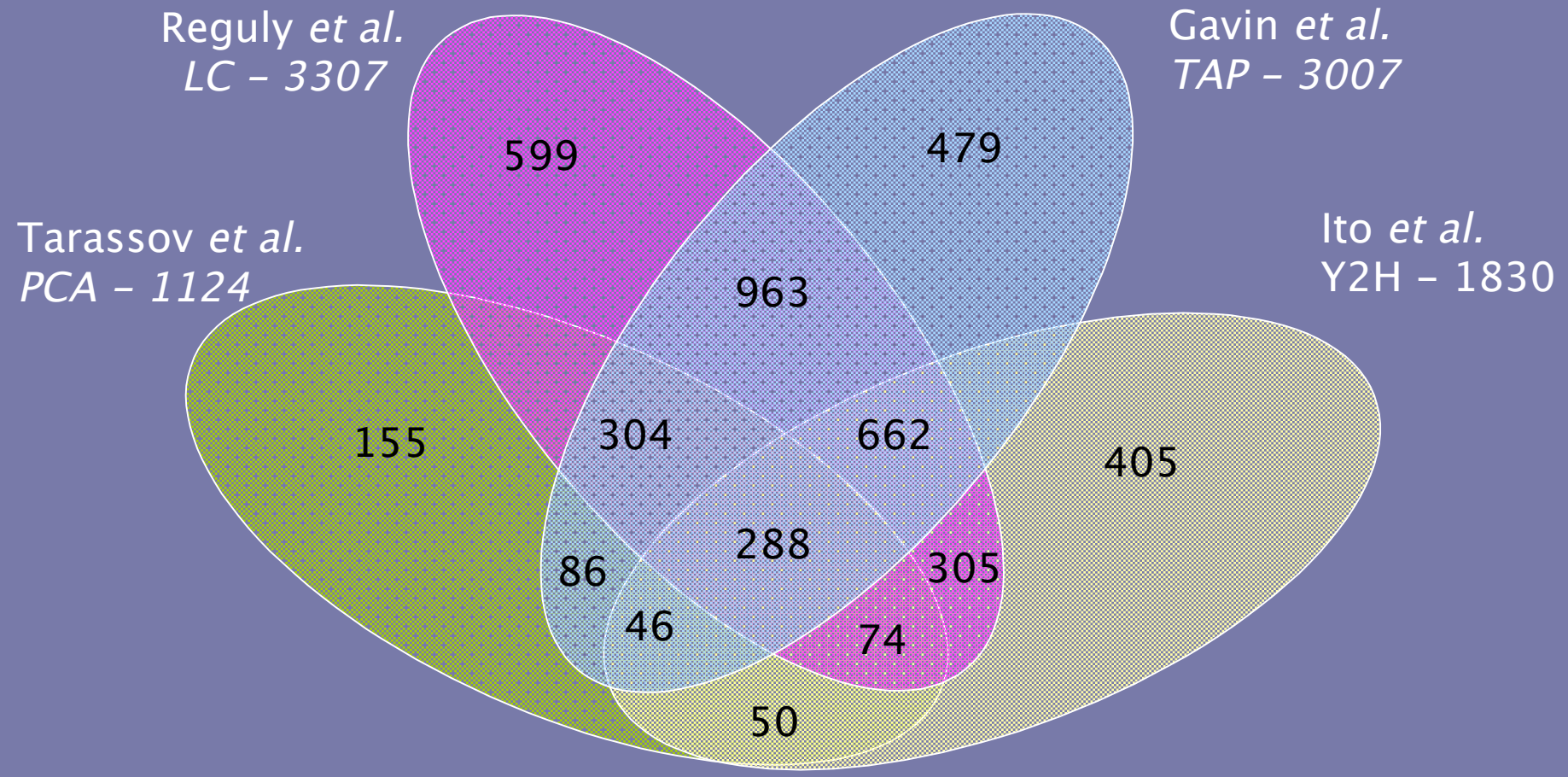
Les complexes protéine-protéine



Souvent plusieurs semaines de travail, succès : 1,8%.

Détection expérimentale

4416 complexes on été observés dans la levure
Très faible recouvrement entre les méthodes (50)
Beaucoup de complexes ne sont pas suffisemment stables



Modélisation des complexes protéine-protéine

La fiabilité des méthodes expérimentale n'est pas suffisante. Pour explorer l'interactome, il faudrait essayer « tout contre tout », soit environ 4 millions de couples pour la seule levure.

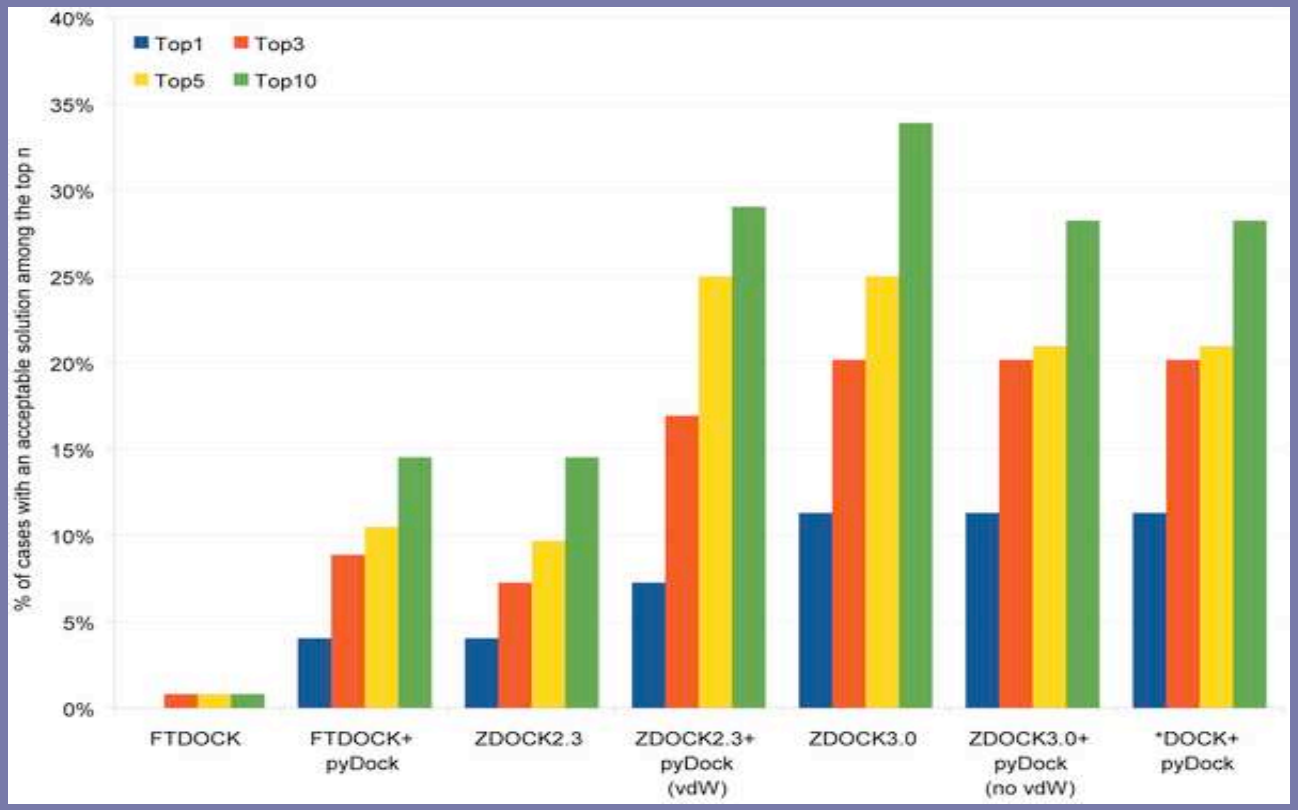
⇒ Pas accessible à l'expérience

Il faut prédire la structure 3D des complexes.

Même par modélisation, cela représente un défi puisque le temps de calcul pour un couple devra être de l'ordre de la seconde.

De plus, la précision doit être très bonne…

Modélisation des complexes protéine-protéine



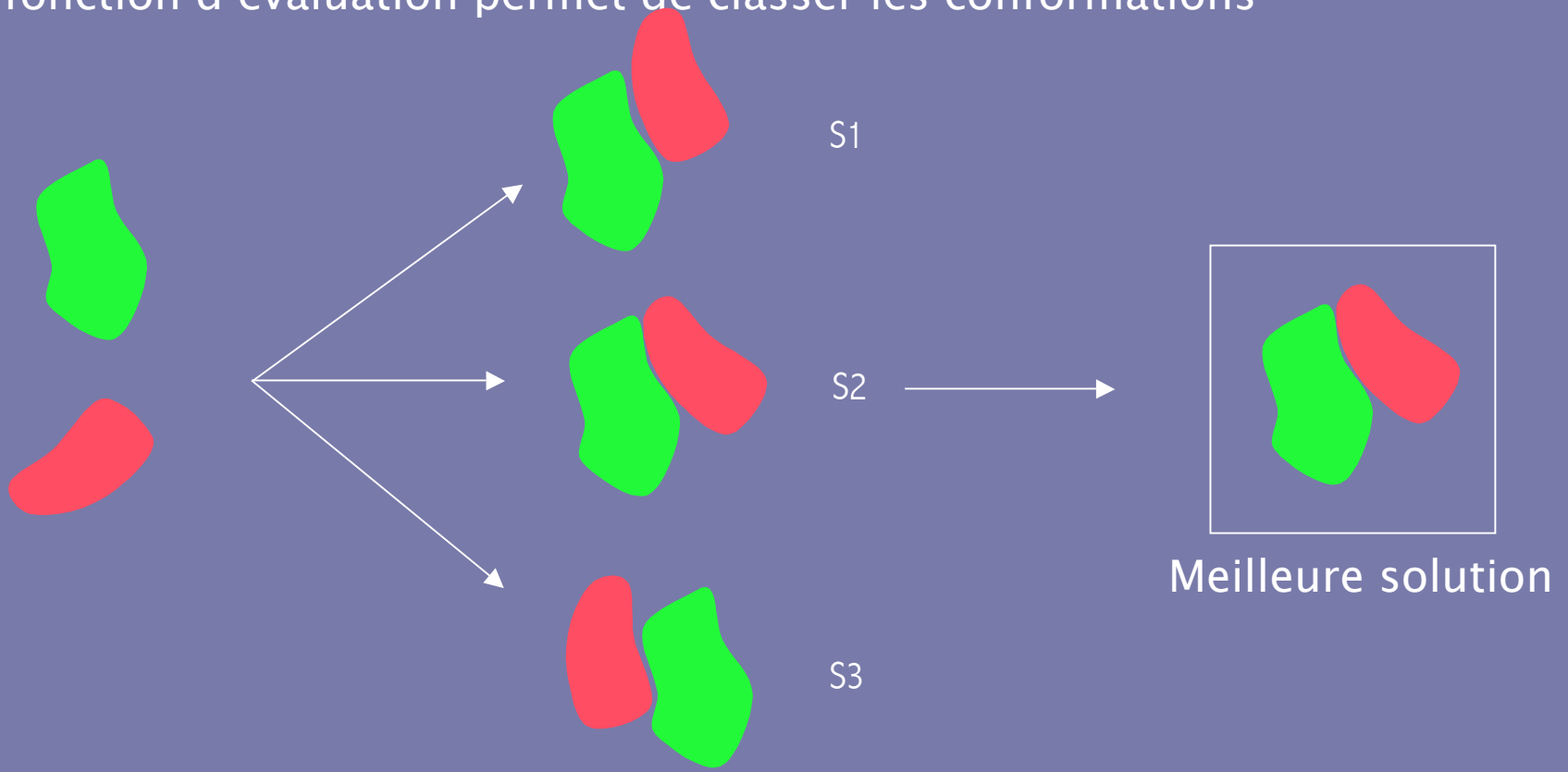
Si 10 solutions par couple sont explorées (expérimentalement), seulement 1 solution sur 30 est correcte. Si on explore 10 solutions par couple pour 20 couples, on aura une solution correcte seulement pour 7 couples.

Pas acceptable !!!

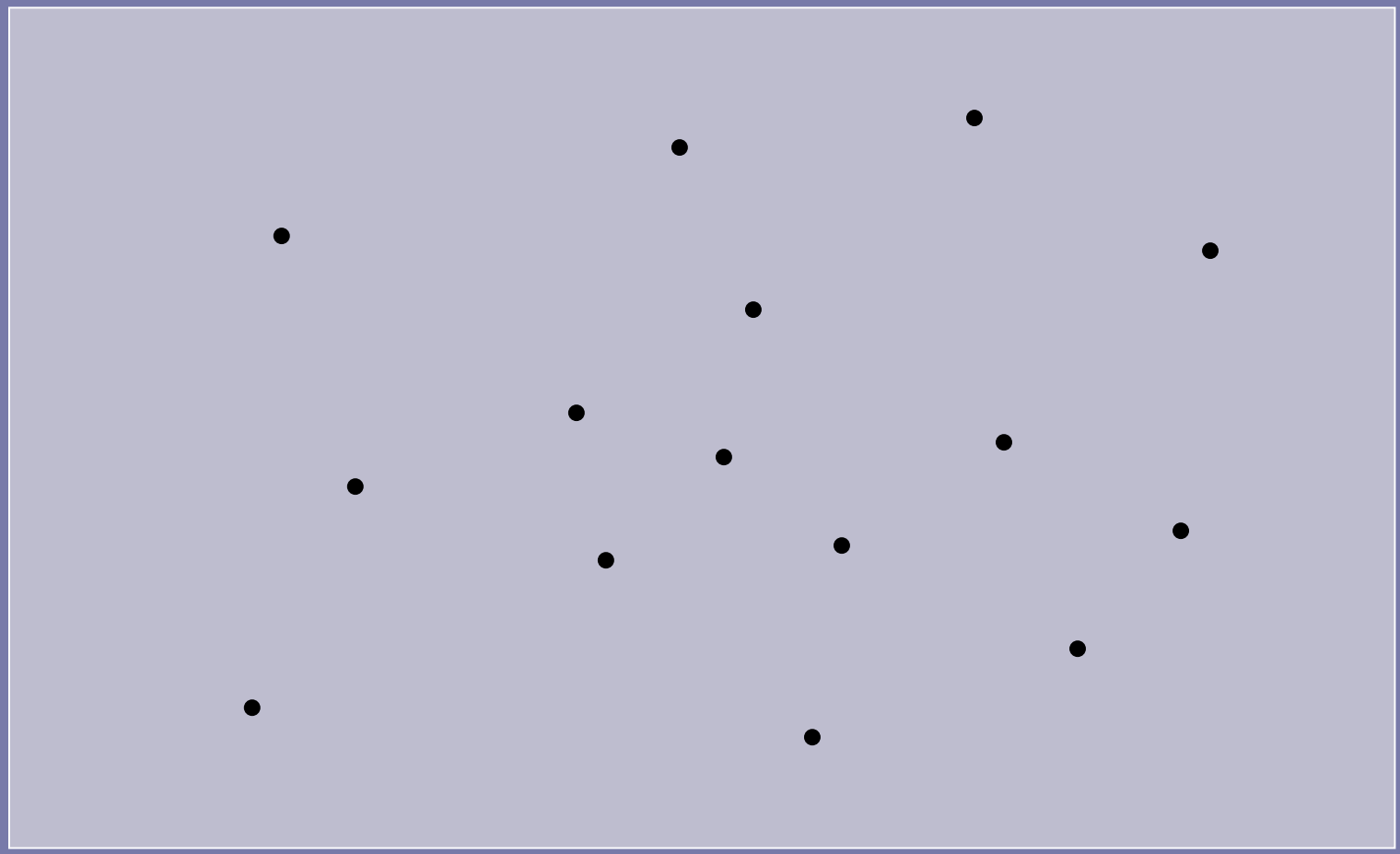
Modélisation des complexes protéine-protéine

Principe : étant données les structures 3D de 2 protéines A et B, quelle est la meilleure conformation possible pour l'assemblage ? Est-elle suffisamment bonne pour que le complexe existe *in vivo* ?
Le problème est généralement traité en deux étapes successives :

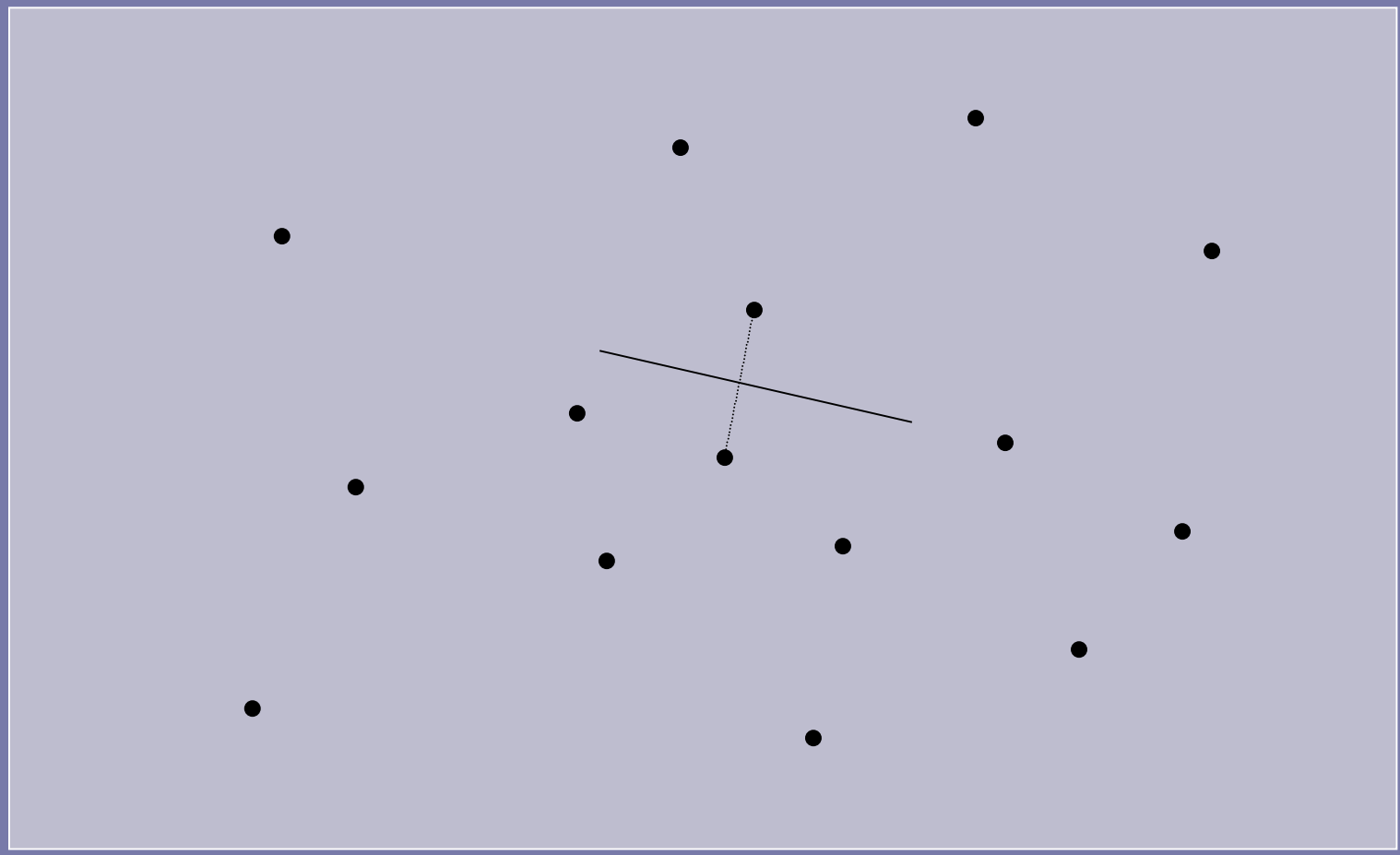
- une échantillonnage des conformations possibles est généré (plusieurs millions)
- une fonction d'évaluation permet de classer les conformations



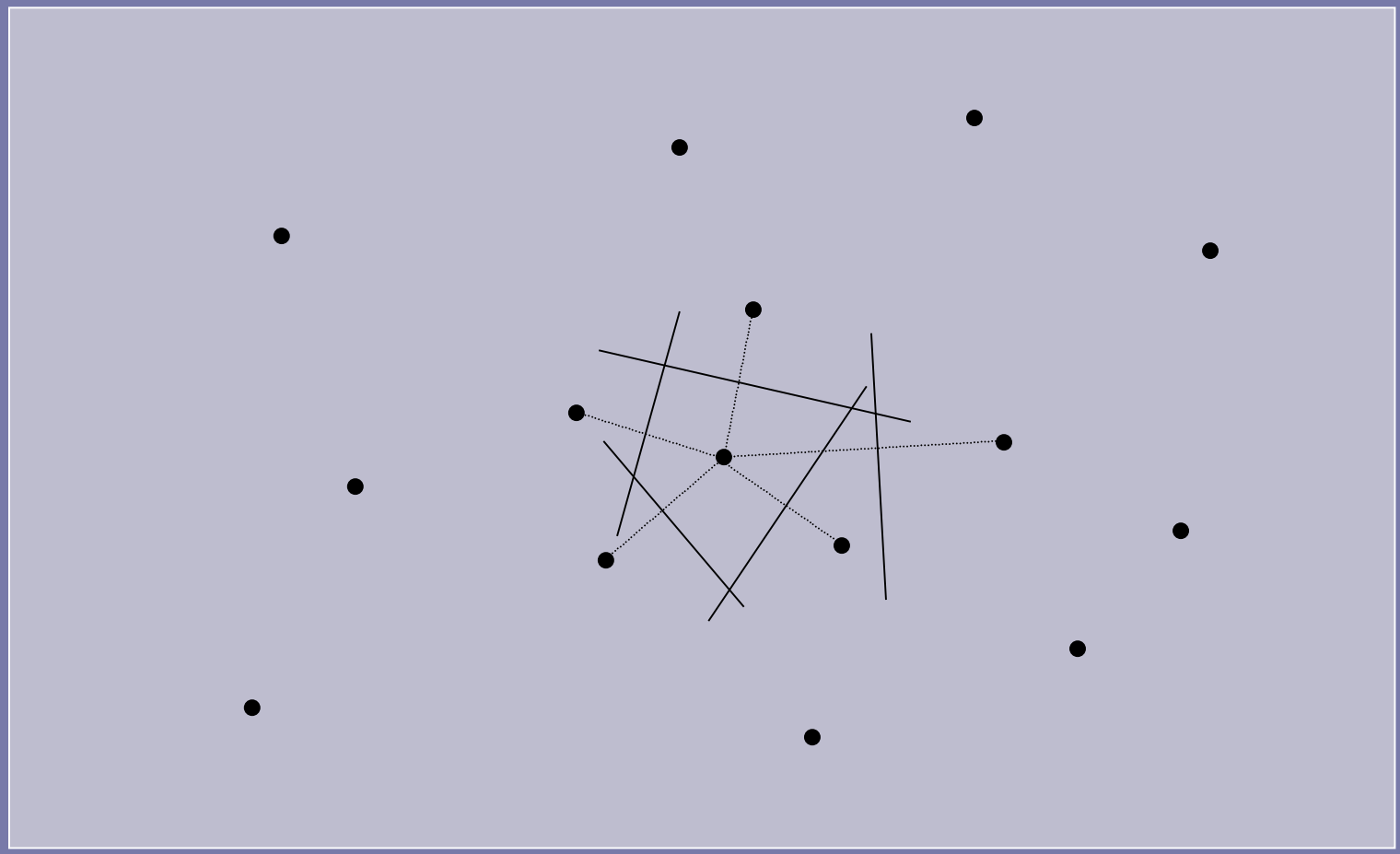
Le diagramme de Voronoï



Le diagramme de Voronoï

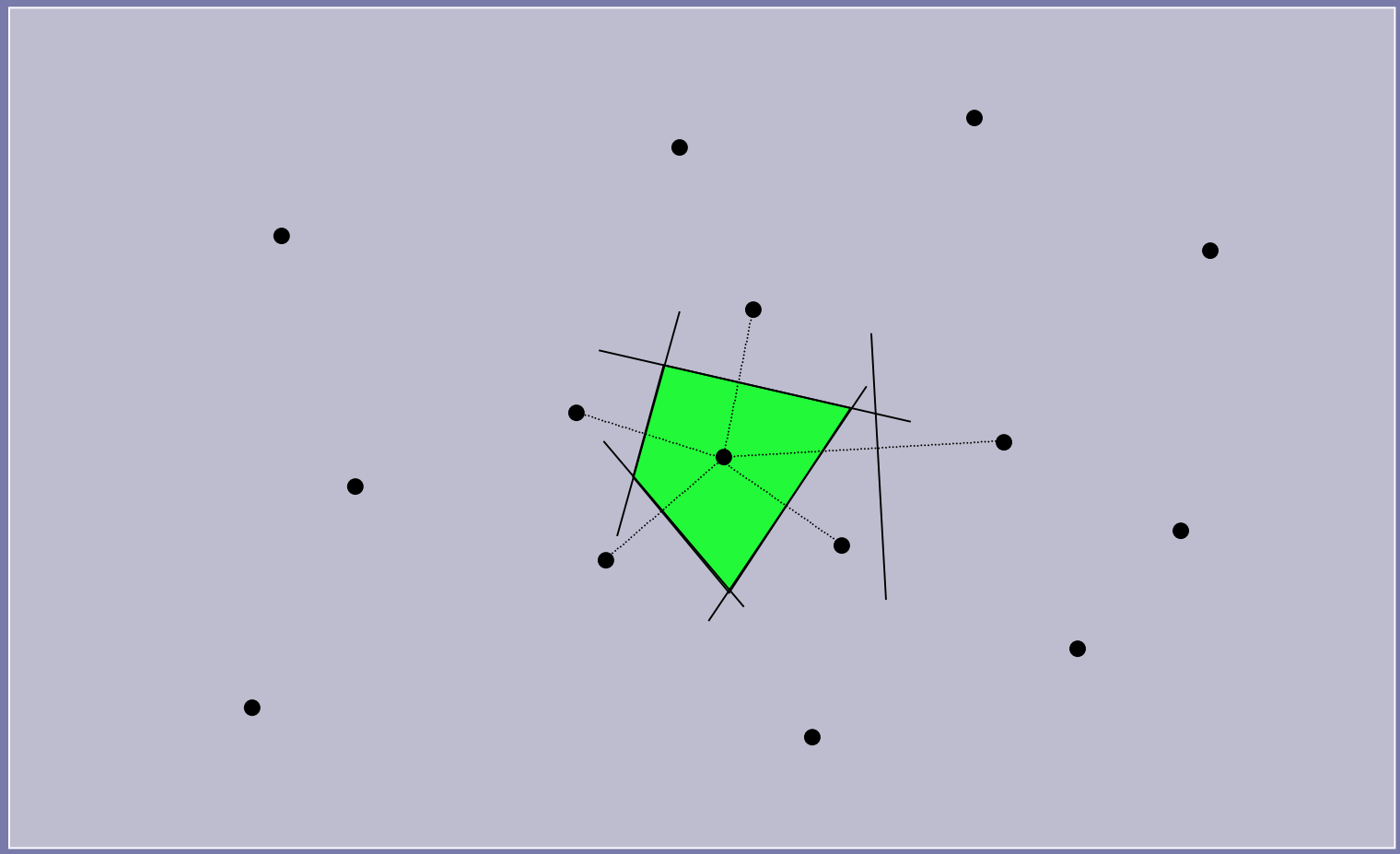


Le diagramme de Voronoï



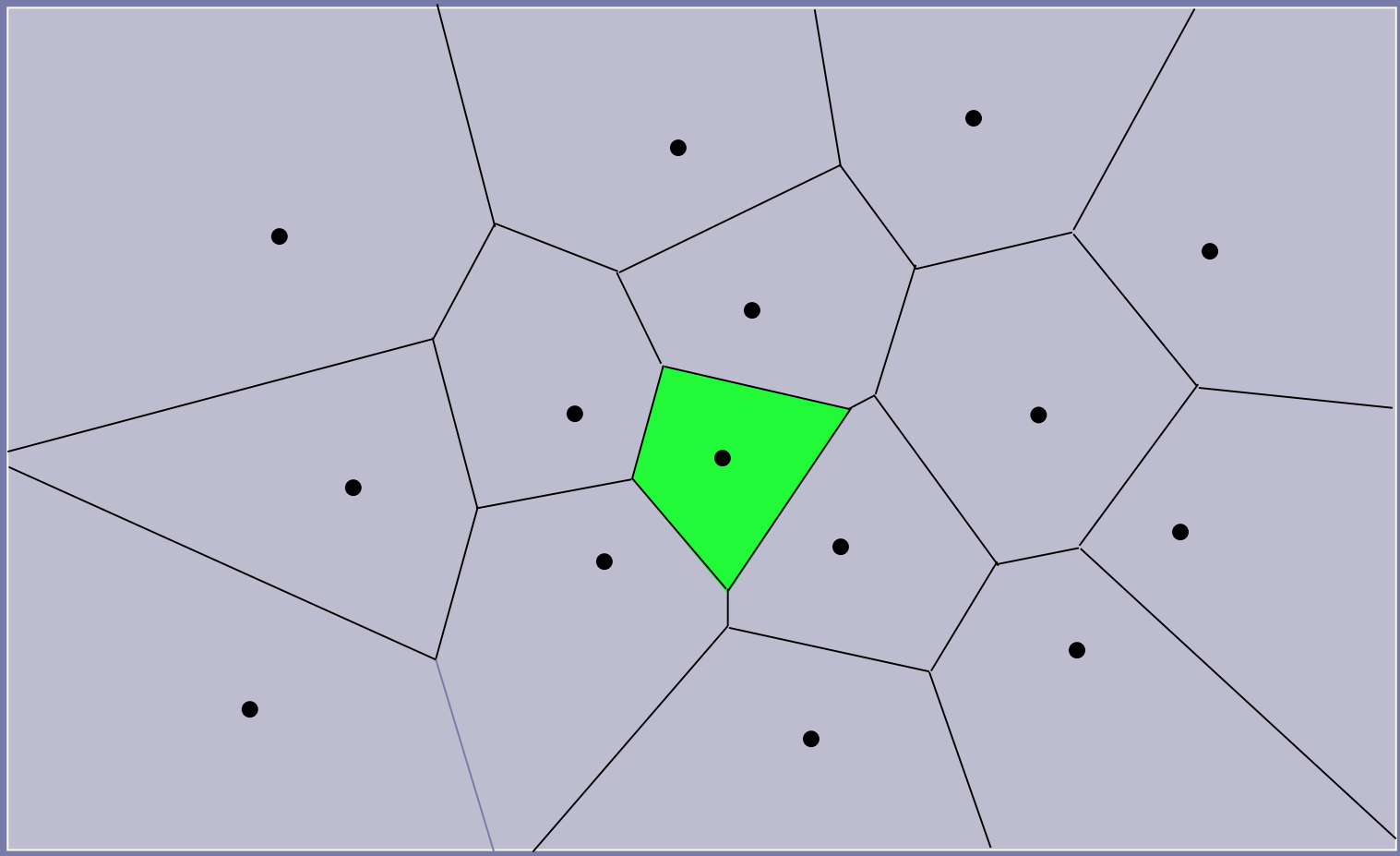
Le diagramme de Voronoï

La région verte est la cellule de Voronoï du centroïde.



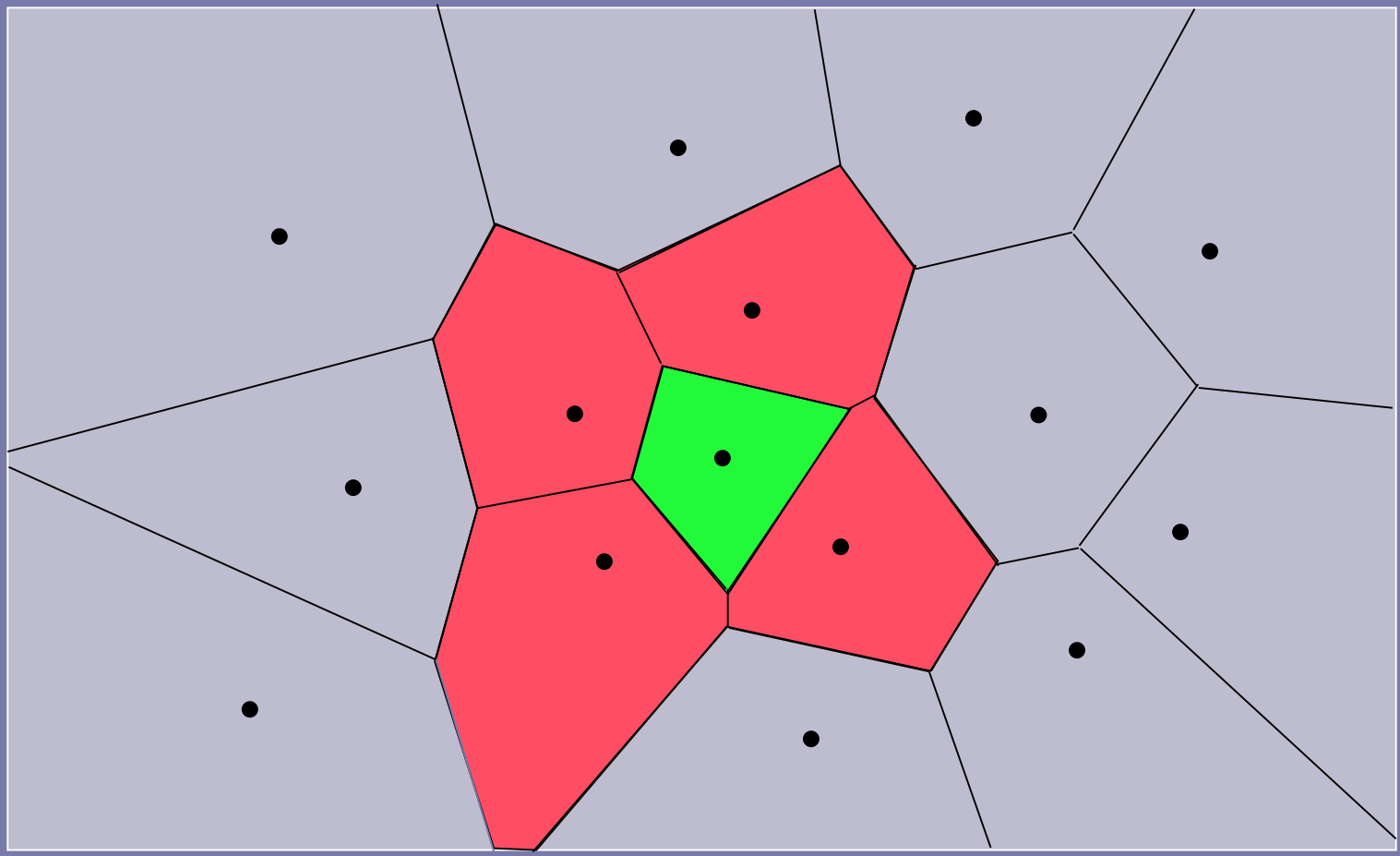
Le diagramme de Voronoï

Le diagramme de Voronoï est un pavage.



Le diagramme de Voronoï

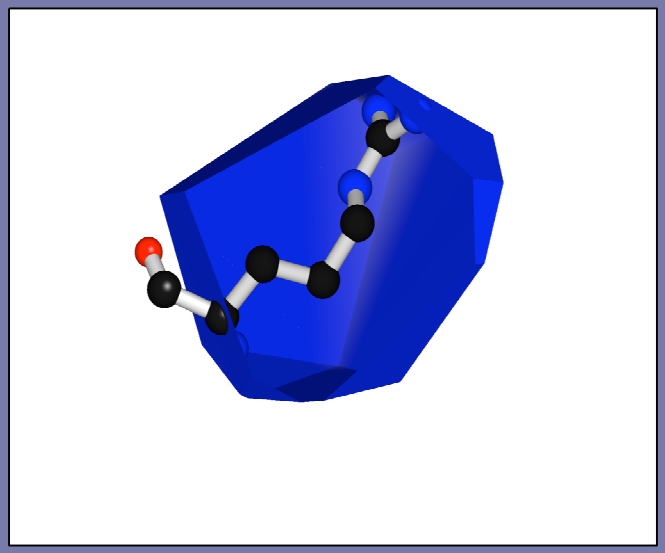
Les voisins peuvent être définis de manière non-ambiguë.



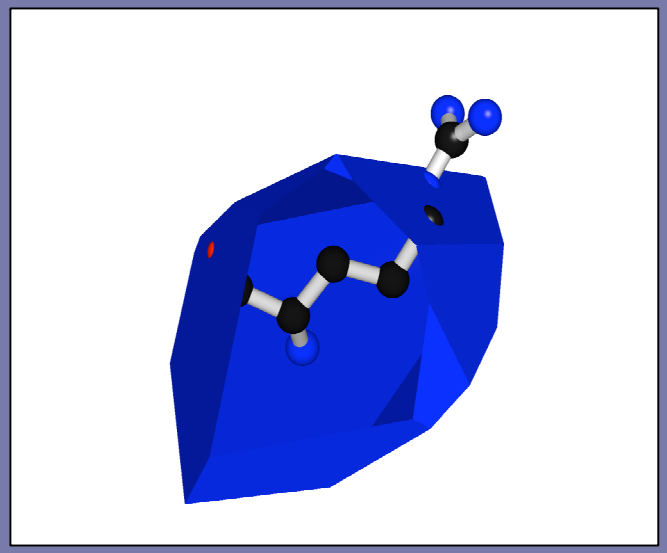
Le diagramme de Voronoï d'une protéine

On choisit un noeud par acide aminé.

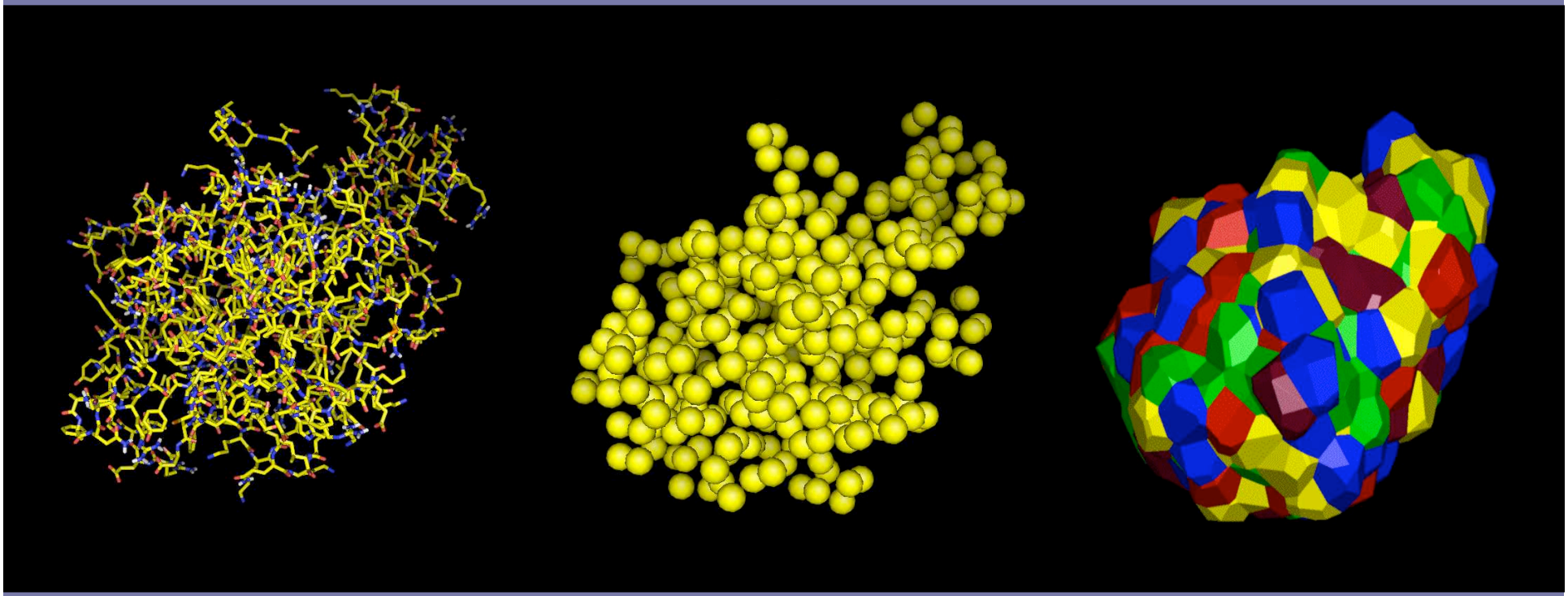
Noeud : centre géométrique de la chaîne latérale



Noeud : Cα



Le diagramme de Voronoï d'une protéine

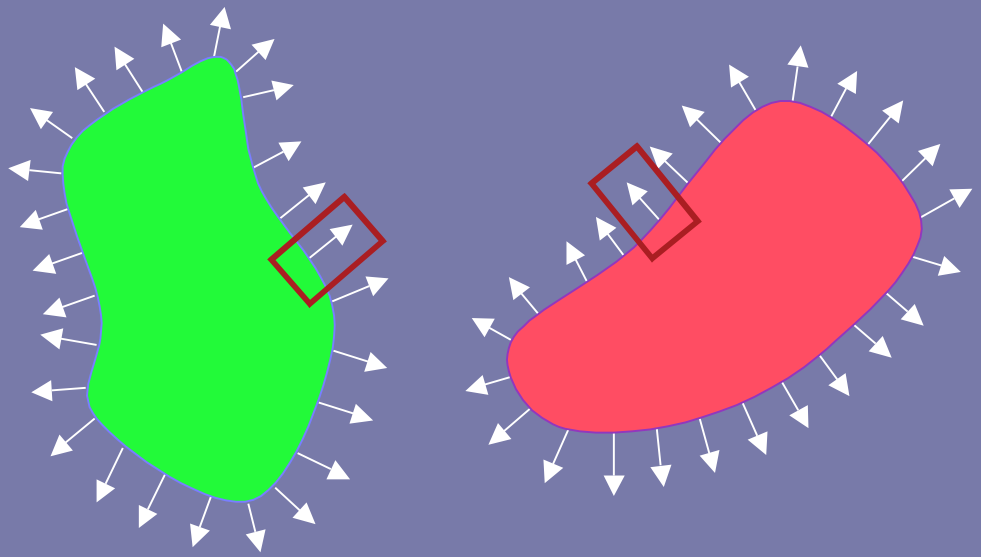


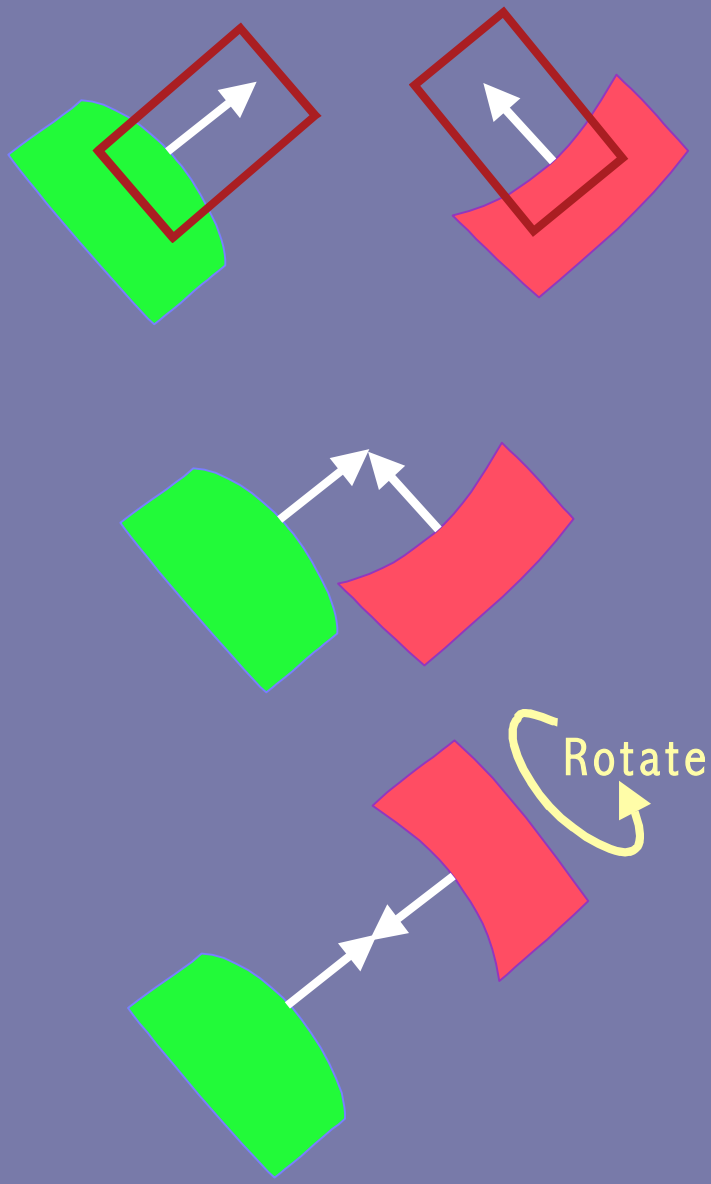
1 – Chaque acide aminé est remplacé par un centroïde
2 – Chaque centroïde est remplacé par sa cellule

L'objet ainsi construit est moins précis que la structure atomique, mais plus « computer-friendly », et moins sensible à la flexibilité des chaînes latérales.

Génération des conformations

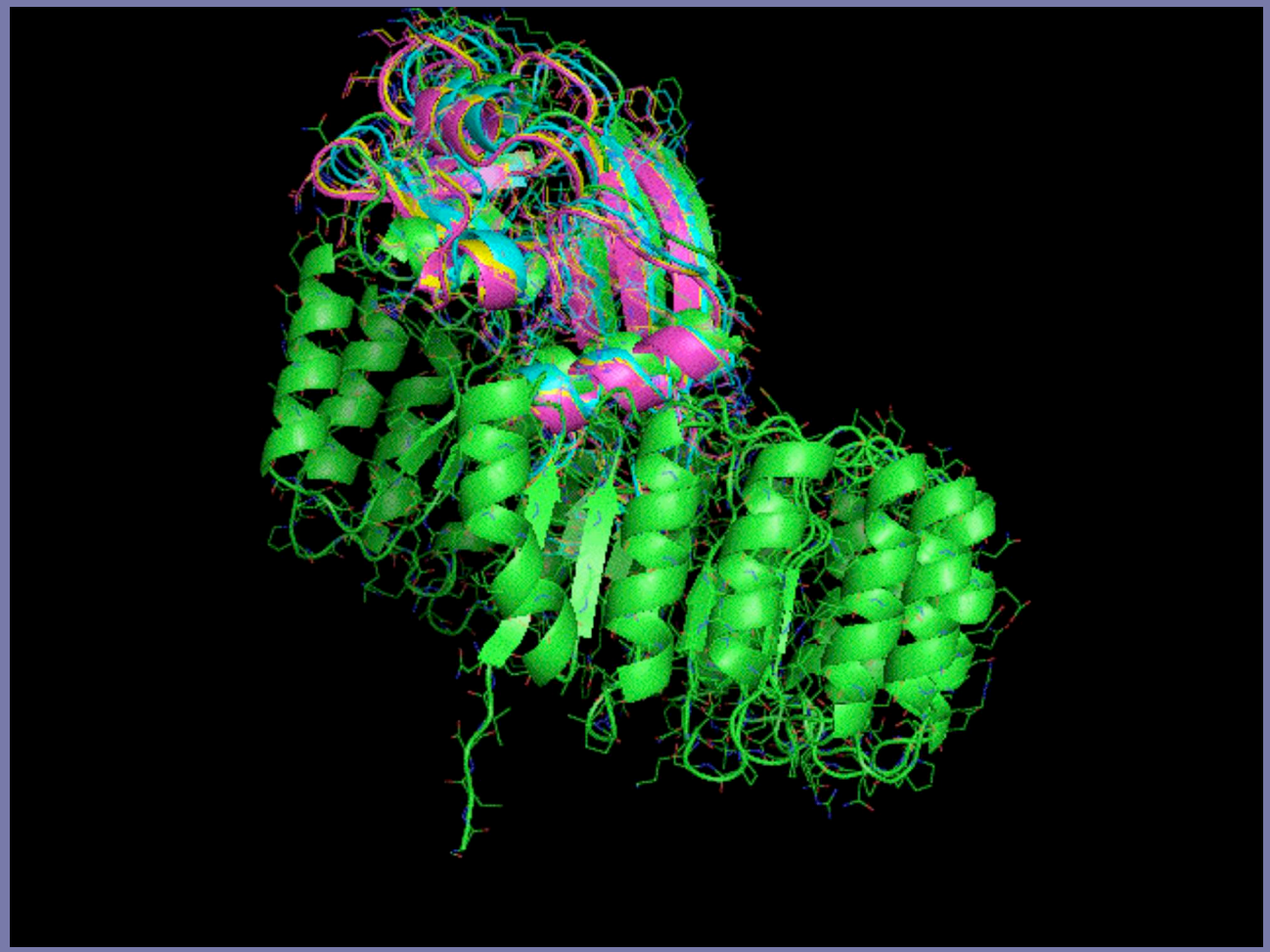
- Remplacer les acides aminés par les noeuds
- Calcul de la triangulation de Delaunay
- Pour chaque noeud on calcule un vecteur « normal » à partir des positions des voisins, et de longueur fixe
- Pour chaque paire de nœuds, on superpose les extrémités des vecteurs, on les aligne, on fait une rotation suivant l'axe.





Génération des conformations

Il y a bien de « bonnes » solutions dans celles qui sont générées.
Cette méthode génère de l'ordre de 1 million de conformations.



Paramètres

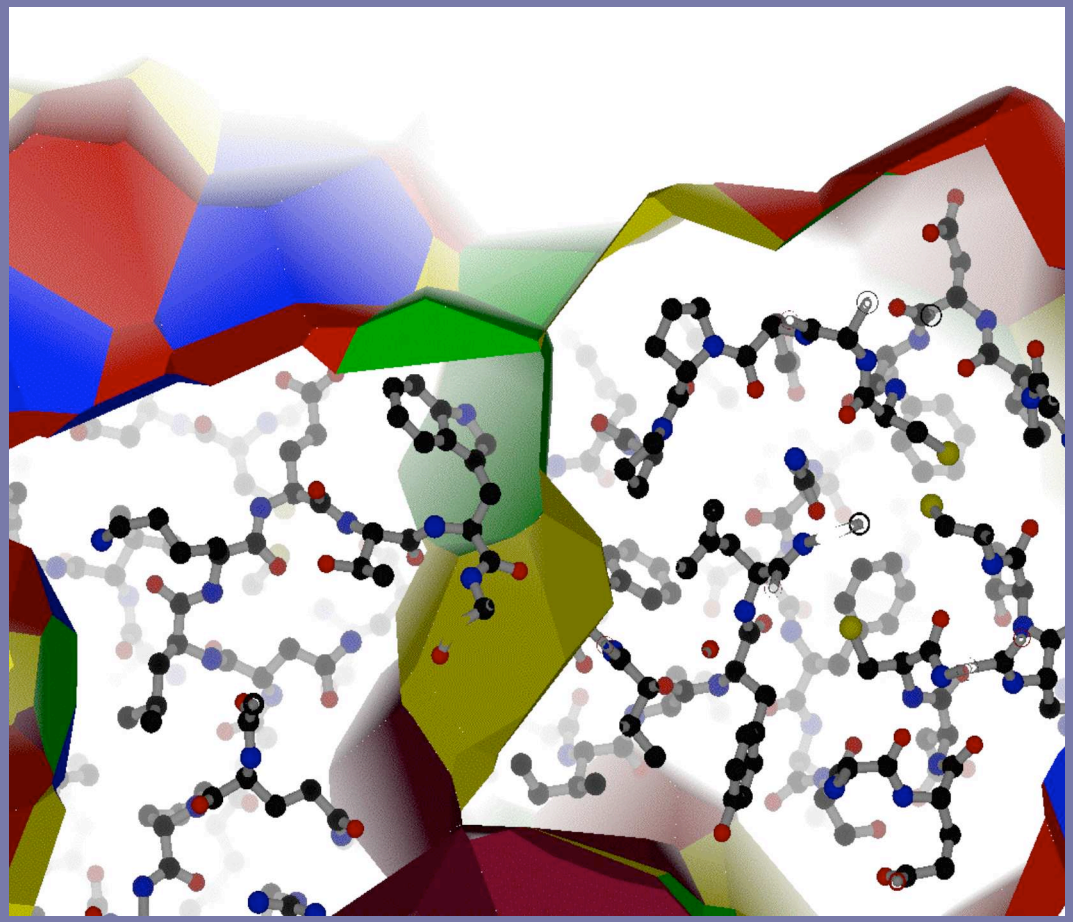
On « observe » l'interface entre les deux partenaires dans des complexes natifs et des complexes non-natifs.

On mesure différents paramètres :

- fréquence des AA
- volumes occupé
- fréquence des paires

...

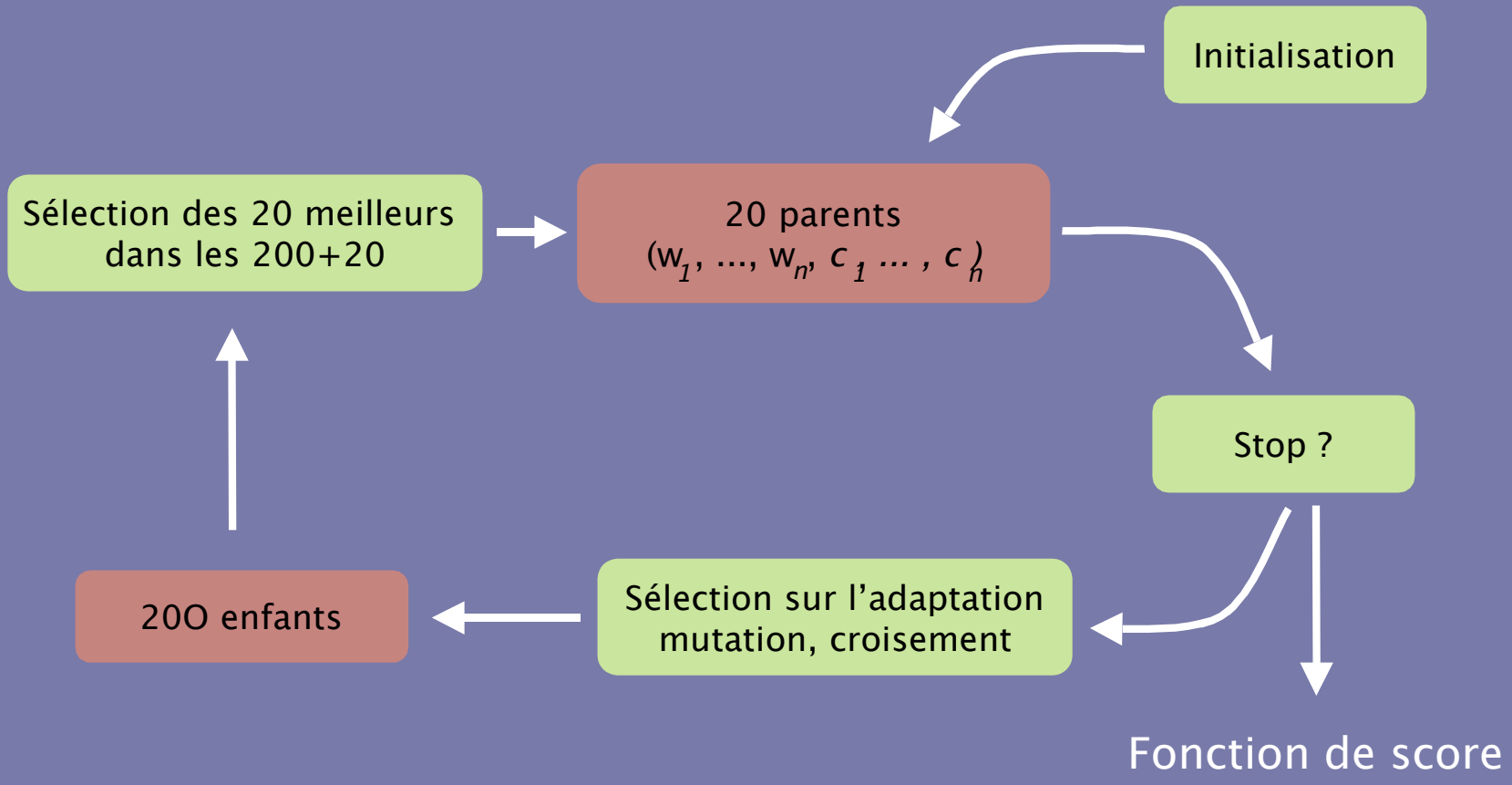
Puis apprentissage machine



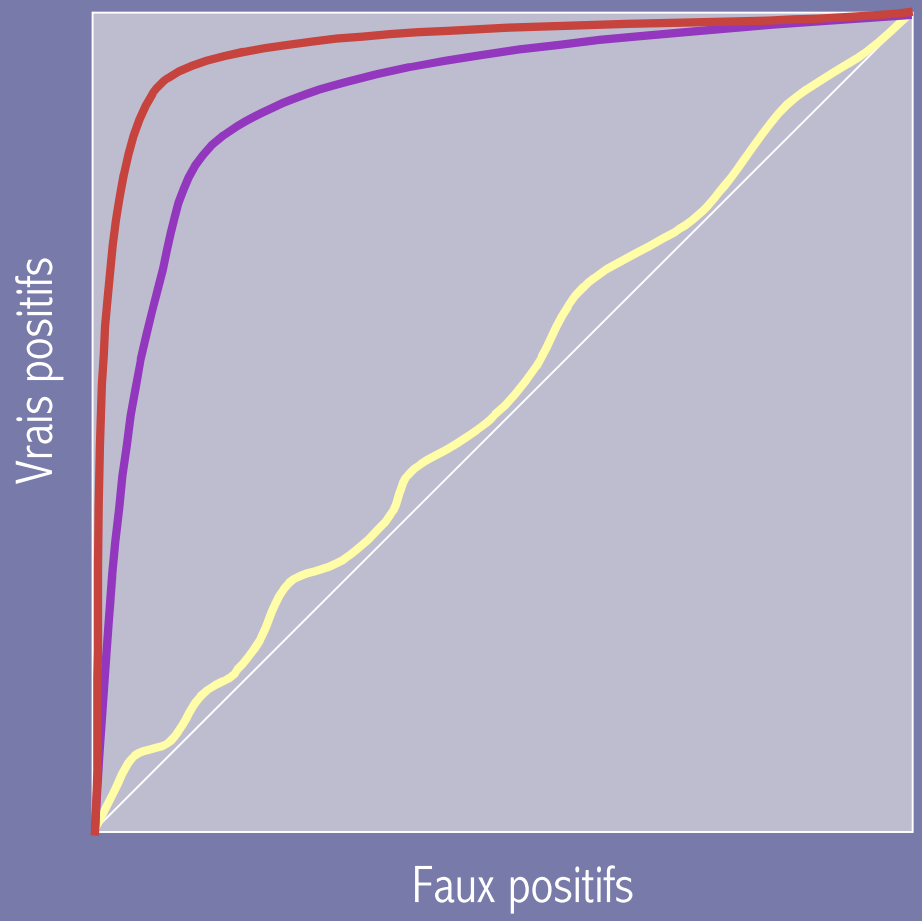
Algorithme génétique

On utilise un algorithme génétique qui optimise l'aire sous la courbe de ROC.

$$S = \sum_i \omega_i |x_i - c_i|$$



L'amarrage protéine-protéine



Ecart quadratique moyen
Presque aléatoire

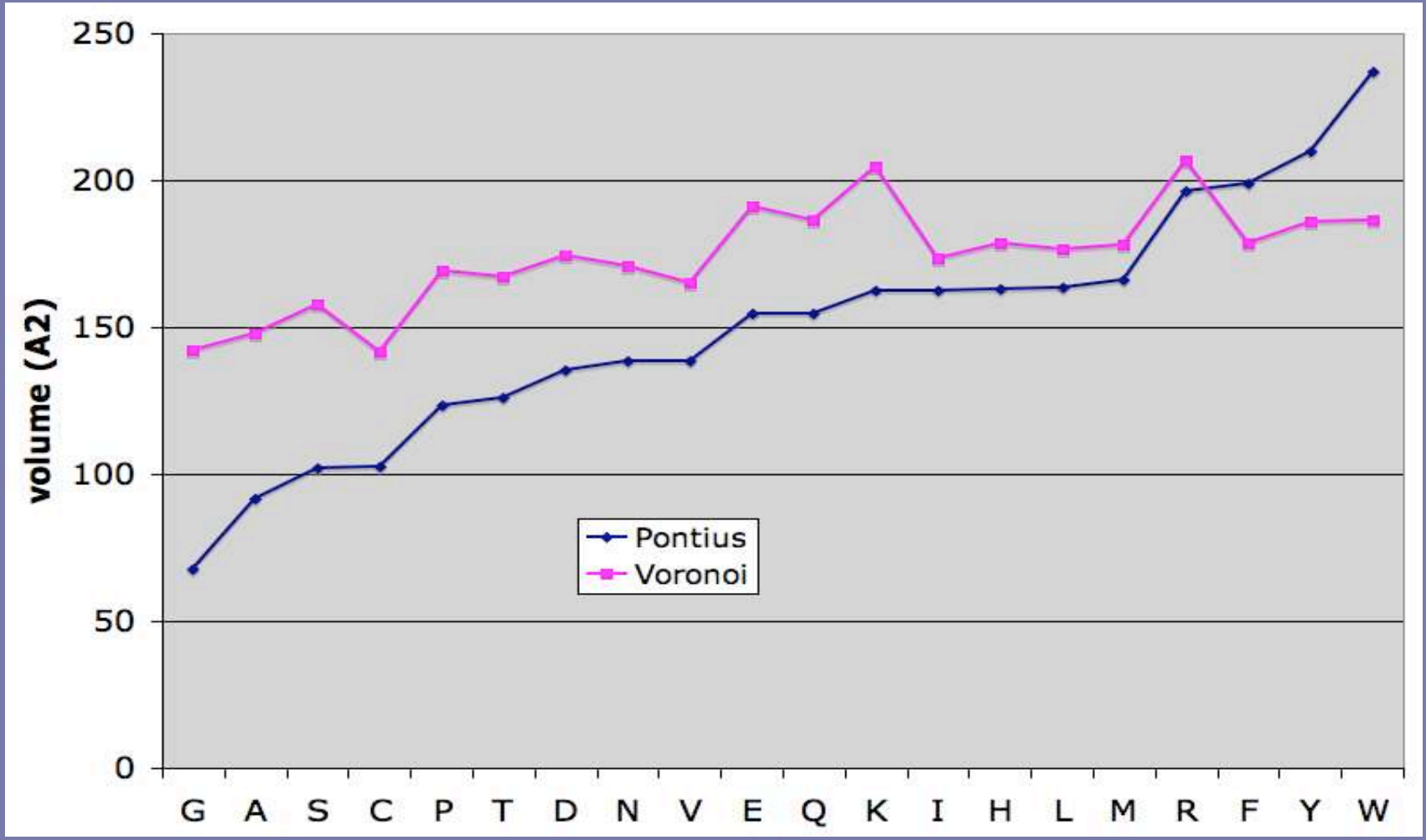
SVM (support vector machine)
Aire 0,85

ROGER (algorithme génétique)
Aire 0,98
Et surtout, beaucoup de vrai positifs dans les solutions les mieux classées.

Bernauer *et al.* (2005) actes de JOBIM 2005 ; Bernauer *et al.* (2007) Bioinformatics

Le problème des volumes

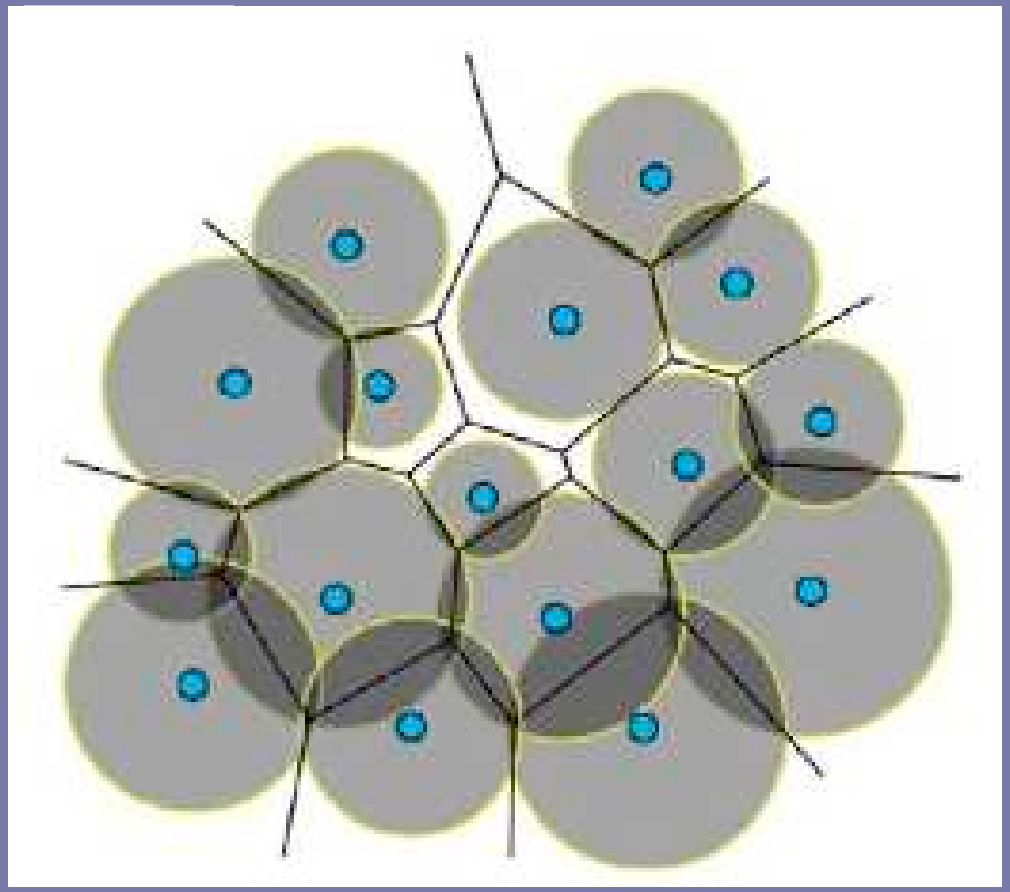
Le diagramme de Voronoï n'est pas pondéré. Les volumes des petits aa sont sur-estimés, ceux des gros aa sont sous-estimés.



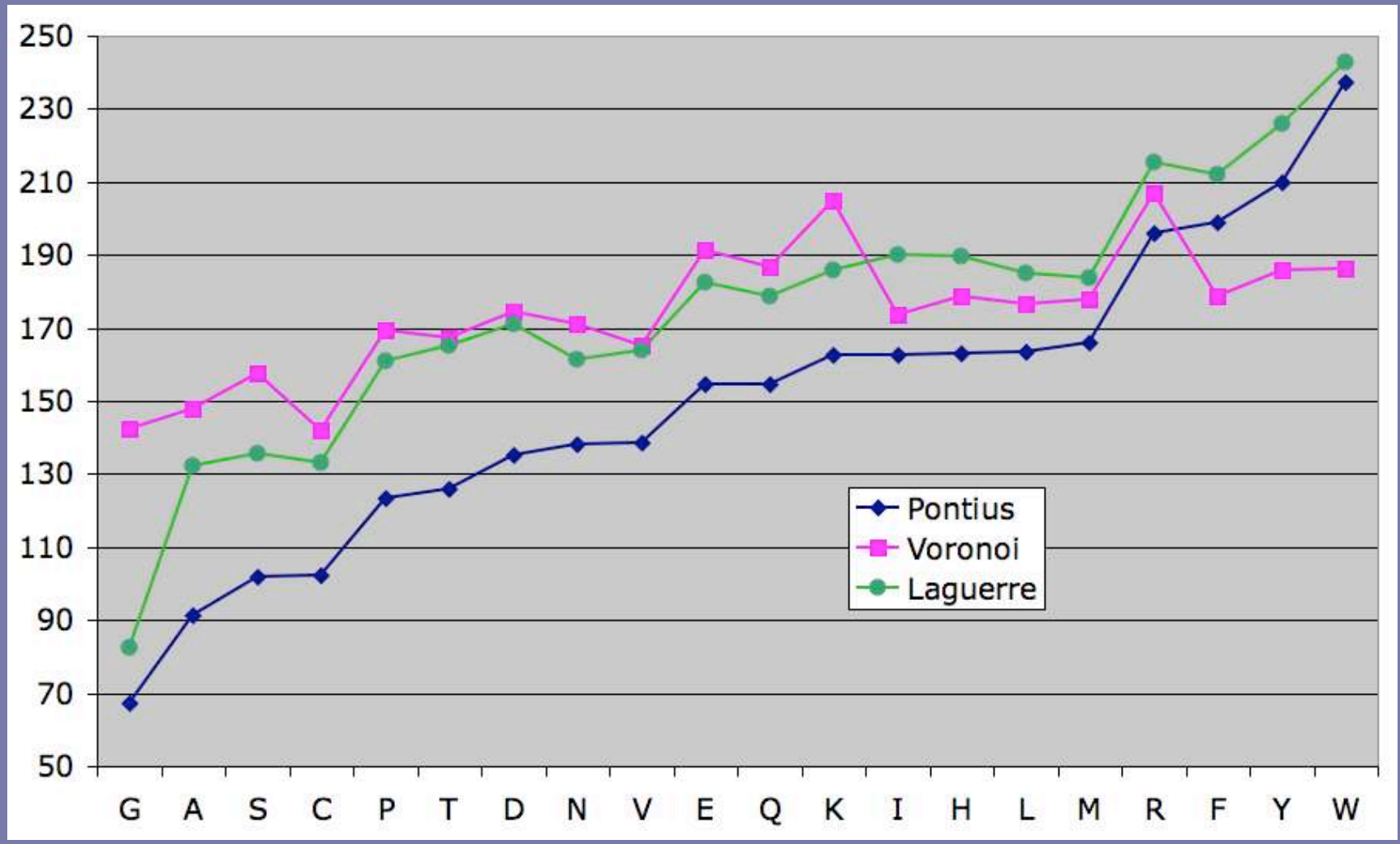
Le problème des volumes

BIOS

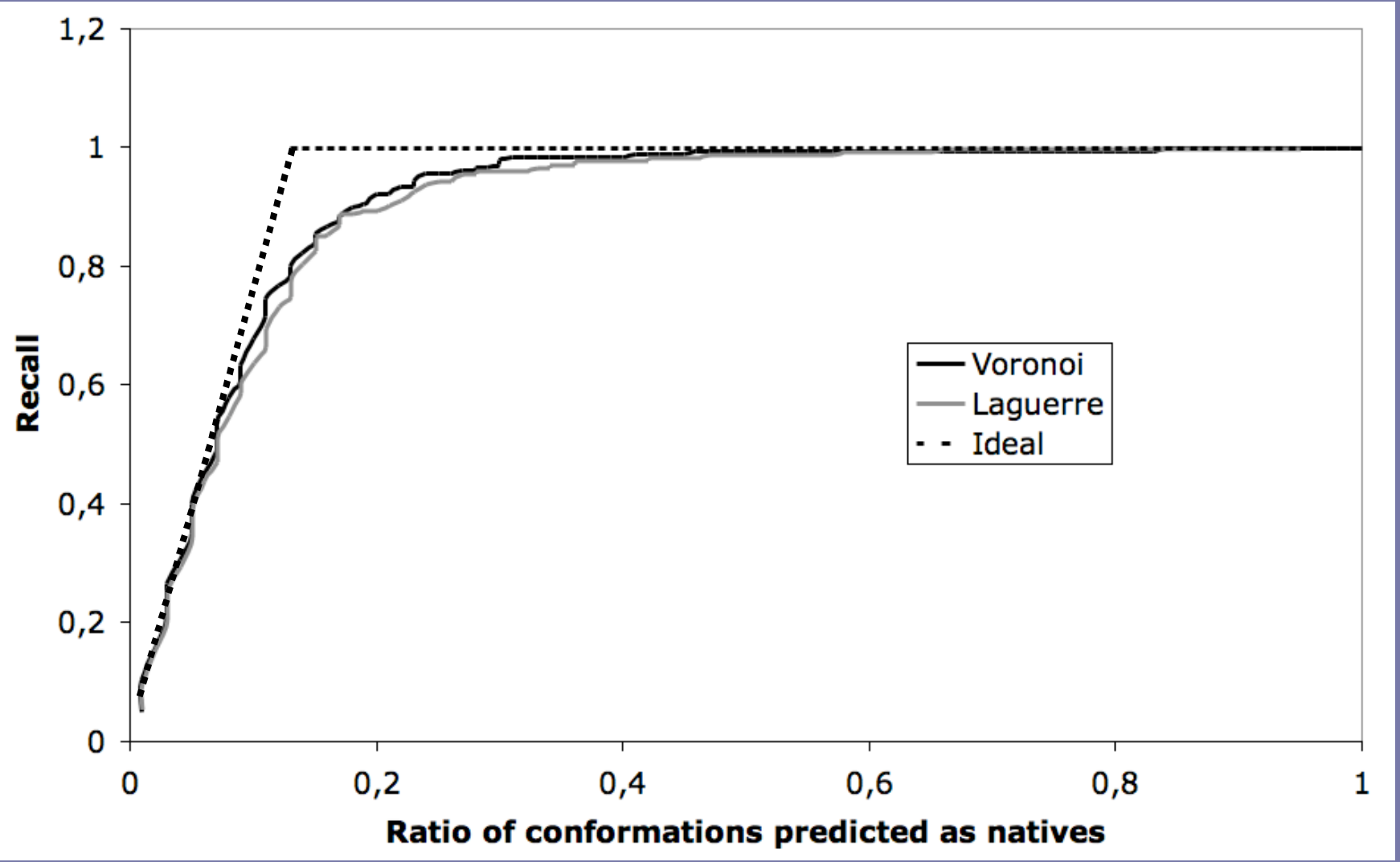
Utilisation d'un diagramme de Laguerre



Le problème des volumes

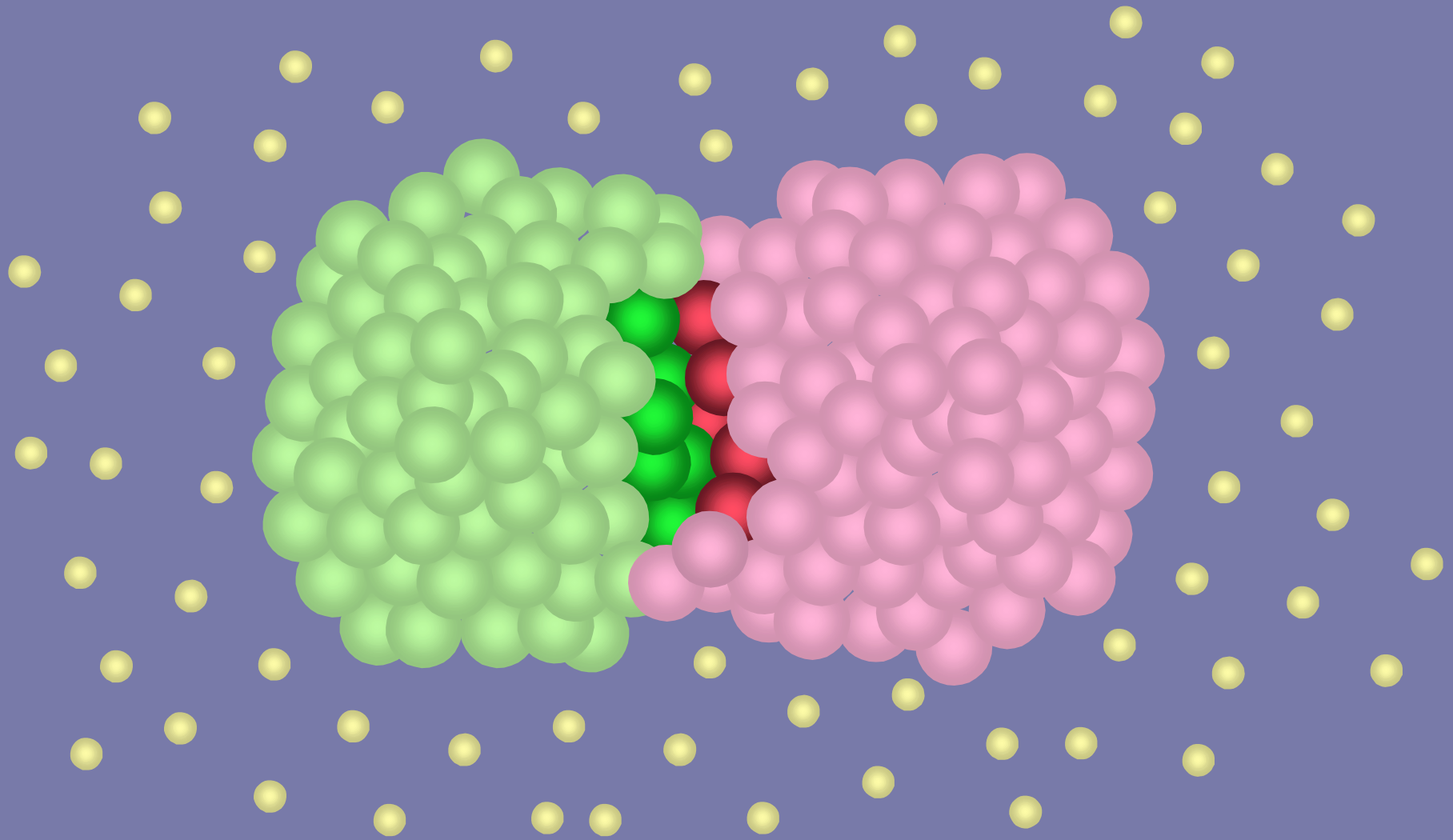


Le problème des volumes



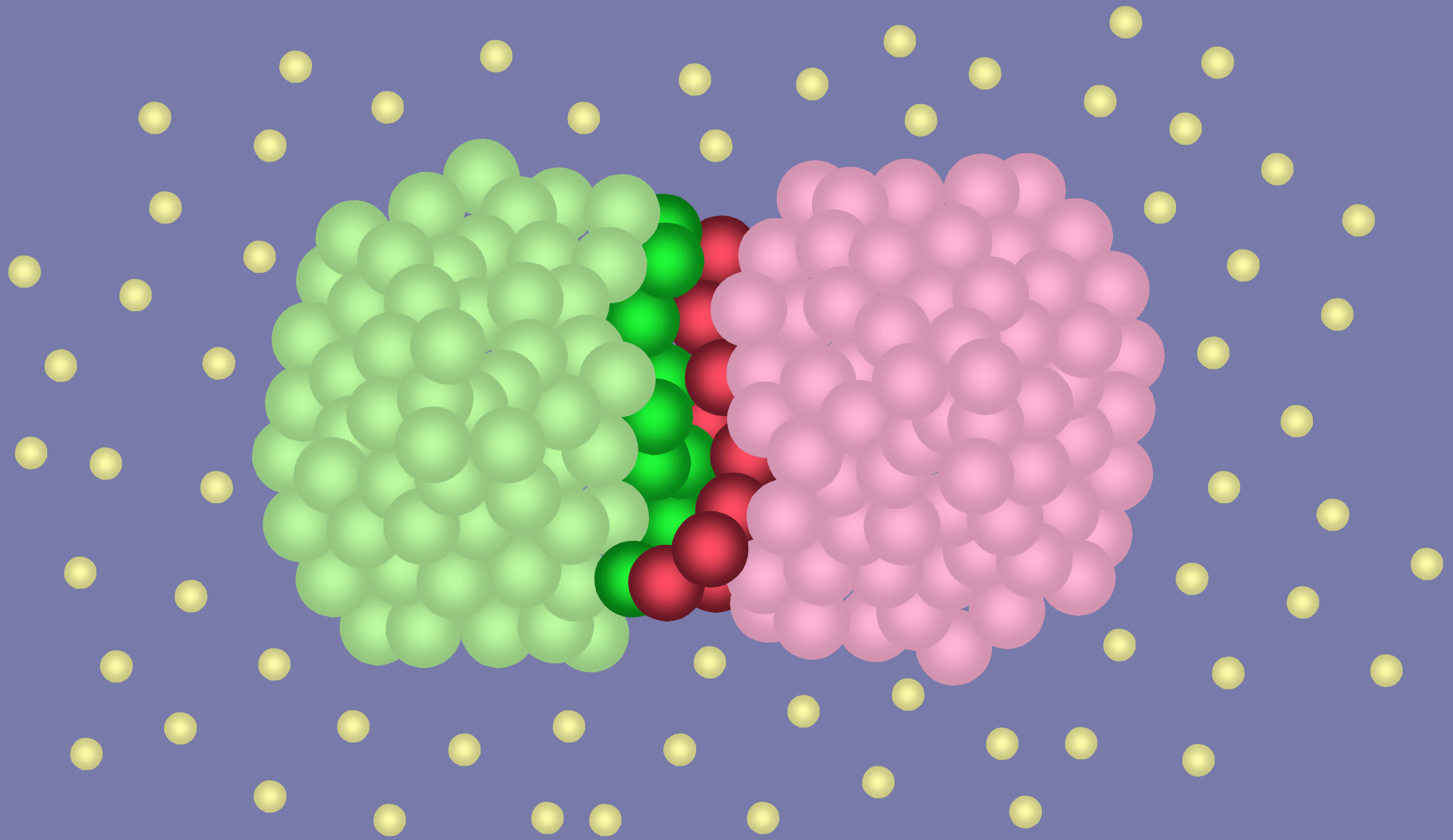
Le cœur et la couronne

On mesure les paramètres seulement sur le cœur de l'interface. Problème : beaucoup de valeur manquantes !



Le cœur et la couronne

On mesure les paramètres seulement sur le cœur de l'interface. Problème : beaucoup de valeur manquantes !



Le cœur et la couronne

On mesure les paramètres seulement sur le cœur de l'interface. Problème : beaucoup de valeur manquantes !

Intégration des résidus de la couronne. Test sur les cibles CAPRI.

Target	Rank core[a]	Rank ring[b]	Total[c]
22	174	170	272
23	46	4	386
24	50	80	100
25	84	39	701
26	261	99	1567
27	1008	1146	1490
28	532	411	1573
29	1037	149	2187
32	98	1	599

Partitionnement

Les données d'apprentissage sont très hétérogènes. On va essayer de faire des groupes plus homogènes.

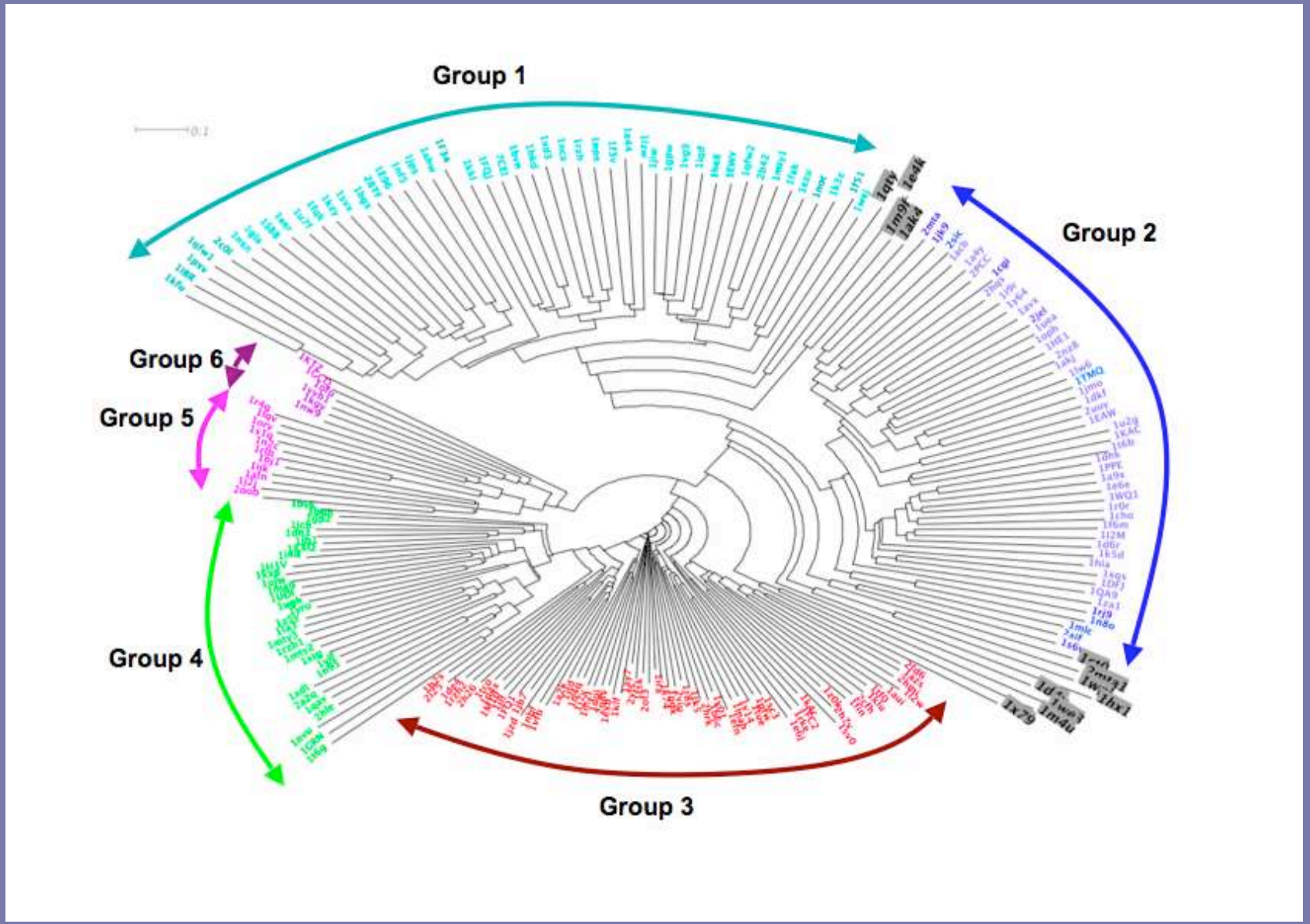
On utilise des paramètres qui mesurent l'entropie de la surface.

Clustering selon la distance (cosinus).

Pour chaque structure native, on agrège toutes les structures natives ayant un cosinus > 0,96.

Pour 211 structures natives, on construit 169 clusters ayant plus de 20 membres.

Partitionnement



Partitionnement

On apprend une fonction de score sur chaque cluster.

Une conformation donnée est évaluée seulement dans les clusters dont elle est proche (cosinus > seuil)

<rang> : rang moyen

<cos> : cosinus moyen

$C4 = \langle rang \rangle . [(1-seuil)/(1-\langle cos \rangle)]^4$

Partitionnement

Rang de la première conformation au moins acceptable

	Seuil					Jeu
Cible	0.99	0.98	0.975	0.97	global	***/**/*/Total
22	1	1	1	1	4	32 / 27 / 96 / 272
23	3	1	1	1	2	22 / 37 / 287 / 386
24	8	2	4	7	60	0 / 0 / 4 / 100
25	4	8	8	10	25	12 / 3 / 12 / 701
26	1	1	1	1	2	680 / 48 / 117 / 1567
27_1	4	2	2	2	4	61 / 81 / 183 / 1490
27_2	1	1	1	1	1	499 / 131 / 106 / 1490
28	5	9	4	8	5	23 / 56 / 180 / 1573
29	14	5	6	7	17	65 / 85 / 79 / 2187
32	4	2	2	1	4	1 / 11 / 177 / 599

Conclusion et perspectives

On classe une bonne solution dans le top 10 dans tous les cas !

La méthode peut encore être améliorée, on doit pouvoir atteindre le top 5.

Différentes pistes :
- Elargir le jeu d'apprentissage
- Améliorer le partitionnement
- Optimisation des fonctions de score : essayer d'autres types de fonctions, d'autres fonctions d'évaluation dans l'algorithme génétique
- Améliorer la combinaison des scores obtenus dans les différentes partitions

Le temps de calcul reste trop long (quelques heures à quelques jours pour un couple de protéines).

Deux pistes :
- Parallélisation
- Heuristique

Remerciements

Génomique Structurale de la Levure
Joël Janin
Julie Bernauer
Thomas Bourquard

Laboratoire de Recherche en Informatique, Orsay
Jérôme Azé
Christine Froidevaux

INRIA, Groupe Géométrica, Sophia Antipolis
Frédéric Cazals
Mariette Yvinec
Jean-Daniel Boissonat

Département Spectroscopie RMN, Université d'Utrecht
Alexandre Bonvin