Quelques problèmes de bioinformatique

Marc Bailly-Bechet

Université Claude Bernard Lyon I – France

Biologie & Modélisation 2009-2010 (saison 1)

Table des matières

Table des matières

Qu'est ce qu'un alignement ?

Un alignement de séquences biologiques x et y (ADN, ARN ou protéine) consiste, à partir d'une distance définie pour chaque couple (x_i, y_i), à trouver le positionnement relatif d'une séquence par rapport à l'autre pour minimiser la somme des distances sur tous les couples (i.e avec un décalage constant k).

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .****************    :  *.: :
```

Intuitivement on veut maximiser le nombre de lettres qui "correspondent" dans les deux séquences.

Deux versions du problème

Alignement global : on prend deux séquences et on cherche le meilleur alignement sur l'ensemble des deux séquences. Exemple : comparer deux séquences d'ADN théoriquement identiques, entre deux individus proches.

Alignement local : on cherche un très bon alignement, mais on accepte qu'il ne concerne qu'une toute petite sous-partie des deux séquences. Exemple : trouver les parties communes de l'ADN de l'homme et de la truite.

Difficultés des alignements

- Multi-alignements
- Problèmes algorithmiques : trouver les méthodes qui donneront les meilleurs alignements, le plus vite possible
- Problèmes statistiques : évaluer l'intérêt d'un alignement local. Si je trouve que l'homme et la truite partagent les séquences ATTGGTGAC, est-ce important ?

Biologiquement, à quoi ca sert ?

Les alignements de séquence sont très utilisés en biologie moderne. On considère que deux séquence proches ont un ancêtre commun récent – et partagent donc, en plus d'une similarité de séquence, une similarité au niveau, par exemple, de la fonction biologique.
On peut ainsi chercher à aligner les séquences génétiques de l'homme et de la souris, pour pouvoir ensuite faire des expériences de génétique chez la souris, et en tirer des conclusions chez l'homme.
Cela permet également d'étudier la génétique de certains organismes, et de découvrir des traces de leur évolution récente (i.e duplication de gènes ou de génomes).

Qu'est ce qu'un arbre phylogénétique

Un arbre phyolgénétique est une représentation de l'histoire évolutive des espèces. Il contient autant de feuilles que d'espèces étudiées, et est binaire : chaque division est un évènement de spéciation.
On n'a pas accès aux séquences génétiques des espèces ancestrales dans l'arbre : on doit donc l'inférer à partir des séquences génétiques actuelles.

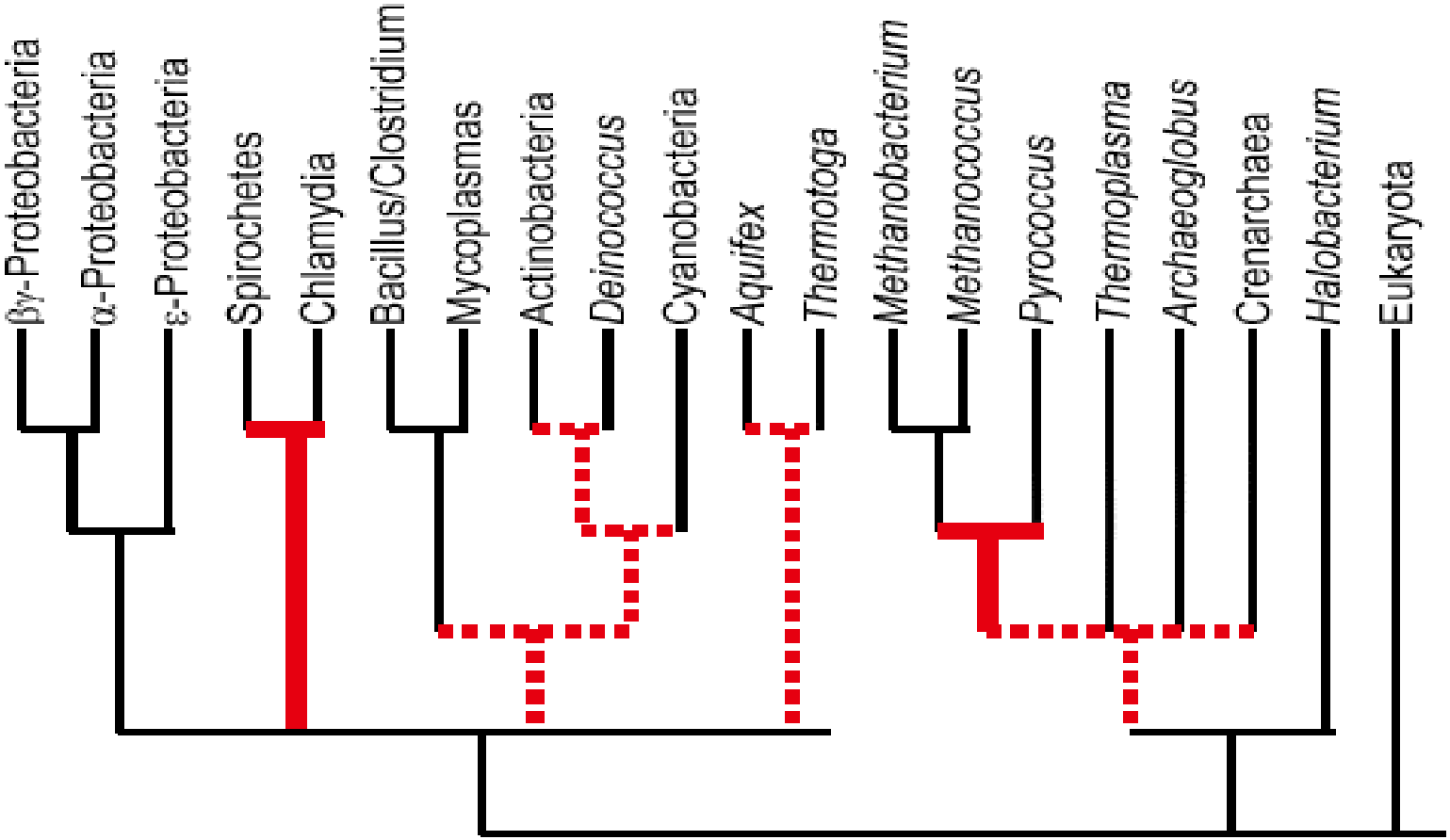
Théorie de l'évolution en deux phrases

Darwin :

"Les espèces évoluent de manière aléatoire, mais que seuls les changements bénéfiques sont conservés. On parle de sélection naturelle".

Et – pour information en ces heures de créationisme et "intelligent design" – scientifiquement, ca marche.

La phylogénie des bactéries



Techniques de reconstruction d'arbres

La procédure générale est :

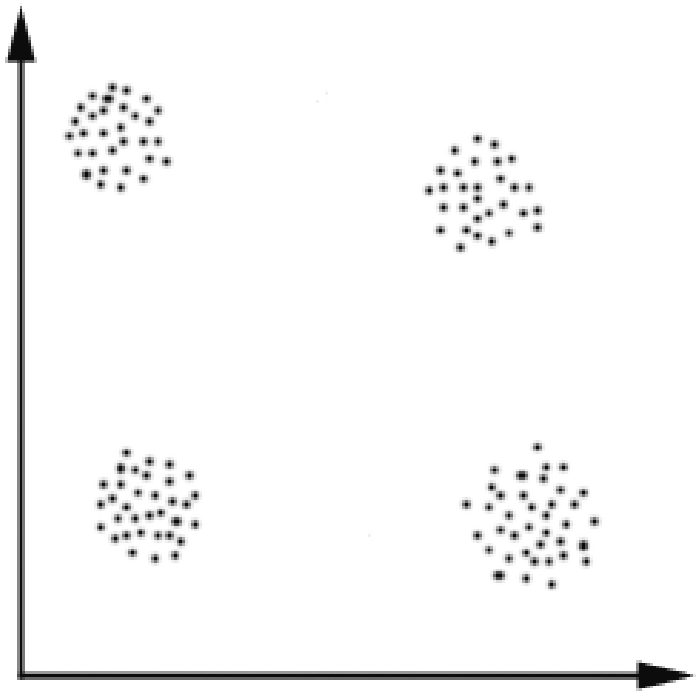
1. Récupérer les séquences génétiques (ADN) des espèces dont on veut reconstruire l'histoire
2. Calculer toutes les distances inter-séquences
3. Grouper les séquences sous forme d'arbre, de manière à ce que les séquences proches aient un ancêtre commun récent.

Les ancêtres, ca n'a pas d'importance...

- Les développements les plus récents sur notre compréhension du HIV sont dus à l'étude de la phylogénie des virus dans un même individu infecté (phylogénie populationelle)
- Avant cela, on avait déjà pu relier le HIV aux SIV grâce à des études phylogénétiques
- On peut même étudier des mécanismes fondamentaux de la génétique, comme la recombinaison à l'échelle des populations, par la phylogénie
- L'arbre de la vie est un ancien objectif, maintenant globalement atteint.

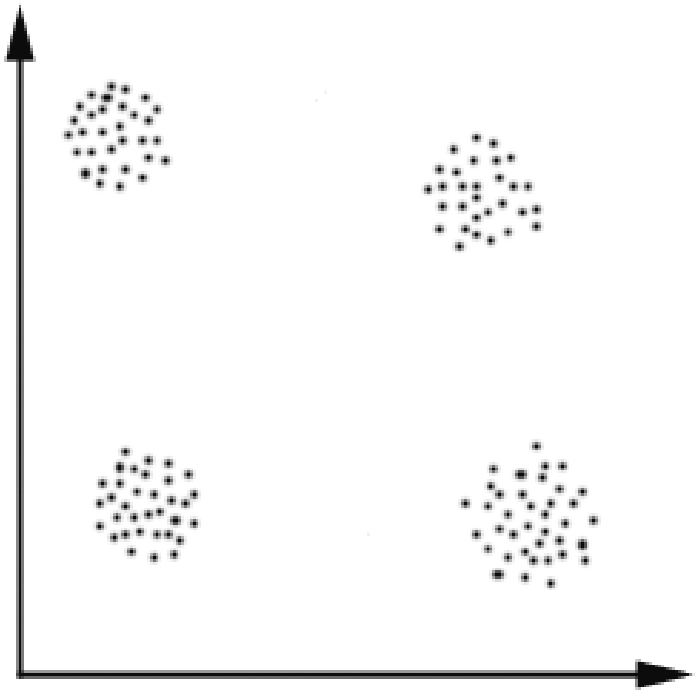
Généralisation : problèmes de classification

Le problème énoncé précédemment est un problème de classification. Un exemple plus général, en images :



Généralisation : problèmes de classification

Le problème énoncé précédemment est un problème de classification. Un exemple plus général, en images :



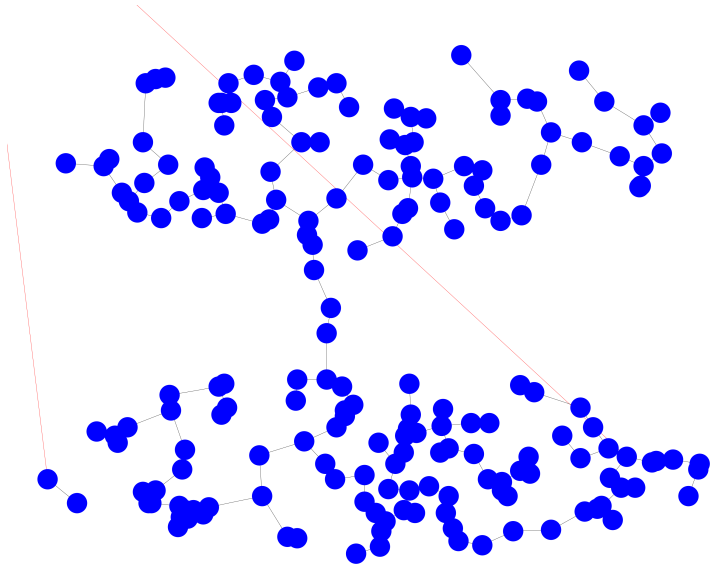
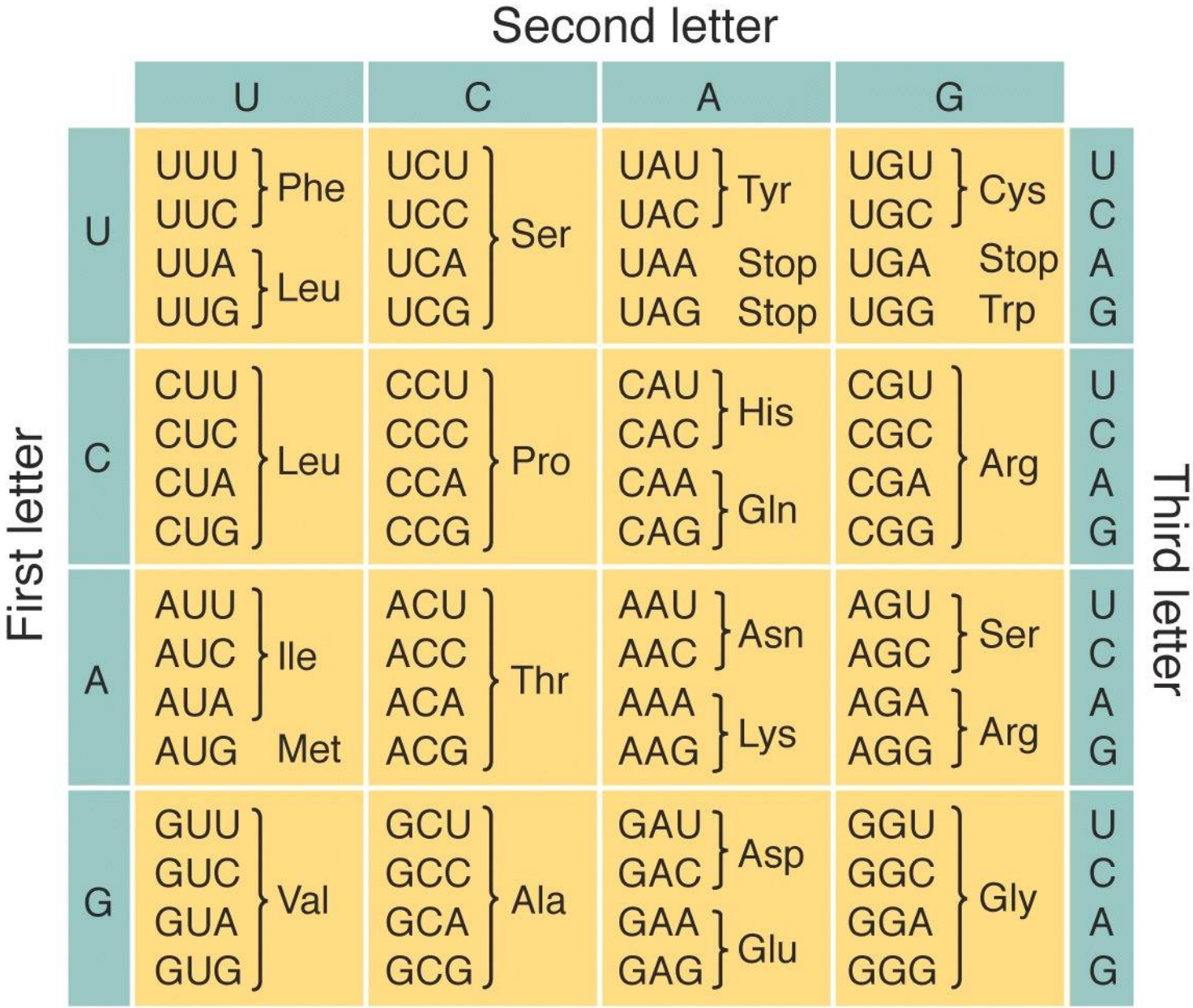


Table des matières

Une pro. . . quoi ?

Une protéine est une chaîne linéaire d'acides aminés (aa). On en trouve 20 (en réalité 23) chez l'ensemble des êtres vivants, toujours les mêmes. Les protéines représentent la plus grande partie des molécules du vivant, et ont des rôles fonctionnels très variés, aussi bien mécaniques qu'enzymatiques, de transport. . .
Les gènes sont les séquences d'ADN qui codent pour les protéines, par le code génétique. Ce code est déterministe, dégéneré et universel.

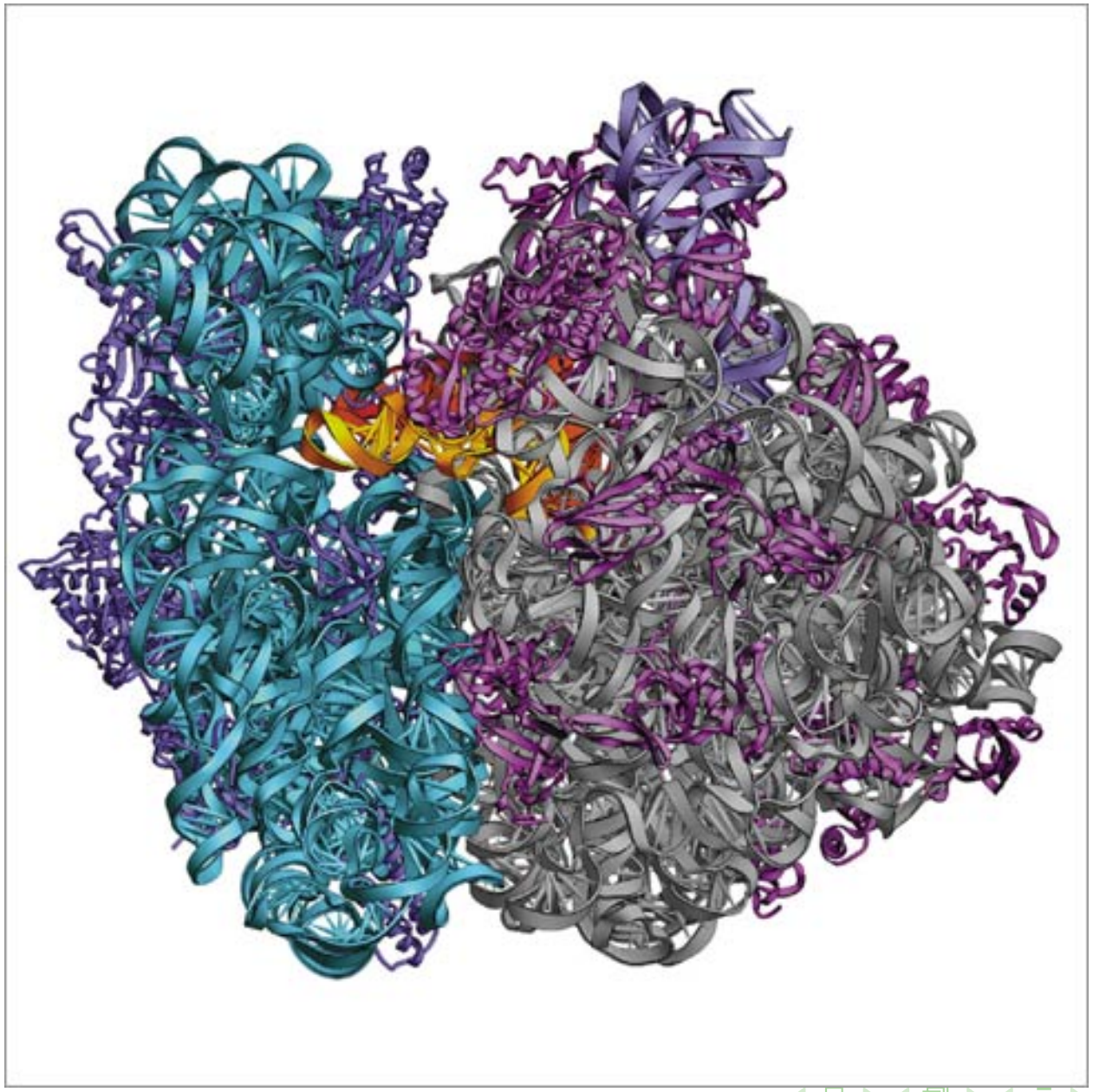
Le code génétique



Comment prédire informatiquement la fonction d'une protéine ?

- Par recherche de signaux peptides
- Par similarité de séquence
- Par similarité de structure (quand vous l'avez)
- Par des méthodes d'apprentissage (machine learning, neural networks, HMM)
- Par prédiction de sa structure

Structure protéique



Prédiction de la structure

- Il est généralement accepté par les biologistes que c'est la structure tridimensionelle d'une protéine qui détermine sa fonction. Or cette structure ne peut être révélée expérimentalement que par des moyens très coûteux et très longs : on dispose actuellement de seulement quelques dizaines de milliers de structures.
- La prédiction de la structure tridimensionelle à partir de la séquence d'acides aminés est un problème ouvert.
- Le problème reste la taille des instances : on n'arrive à calculer le repliement de protéines que si elles font une trentaine d'acides aminés (longueur moyenne chez les bactéries : 300 acides aminés).

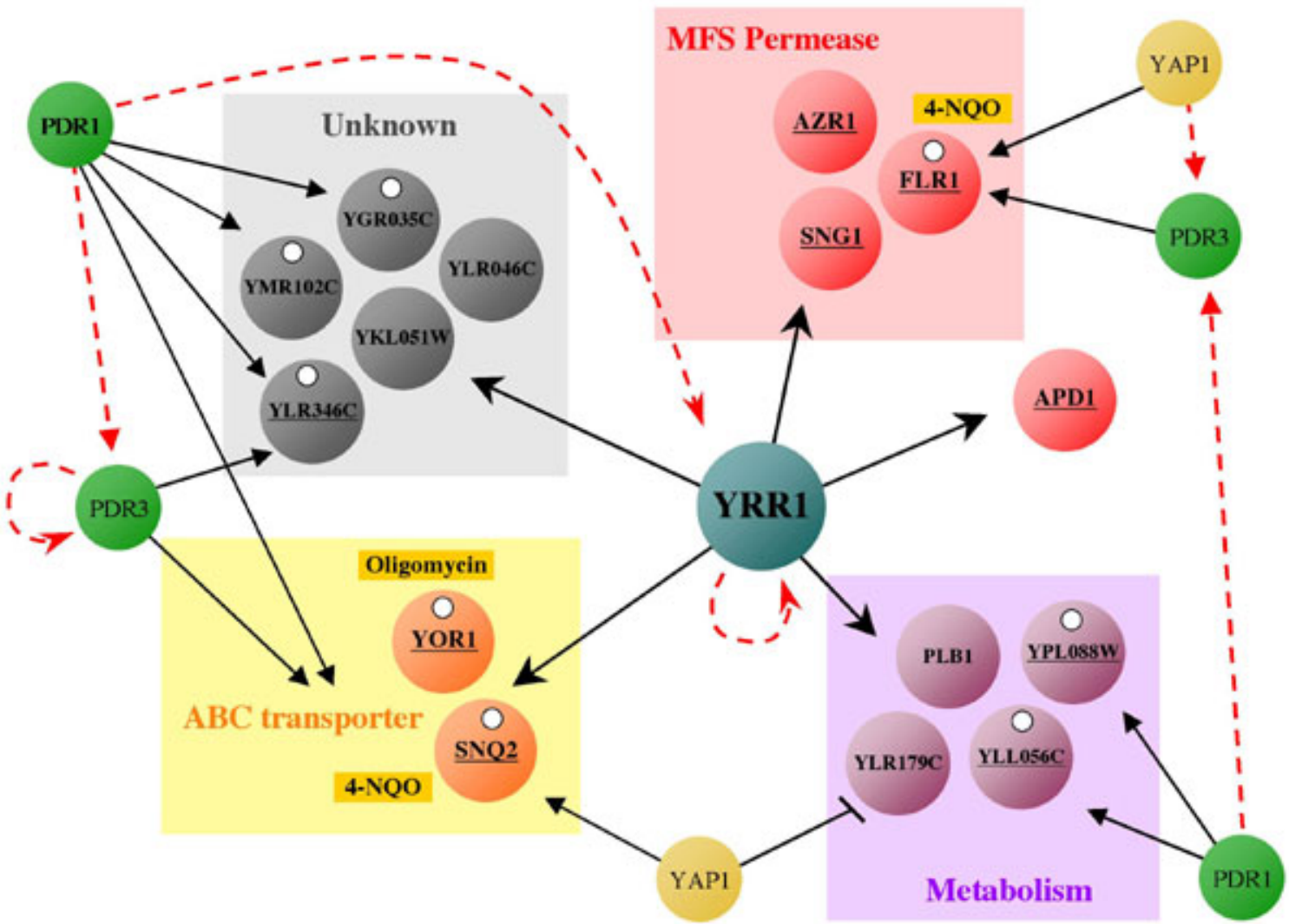
Table des matières

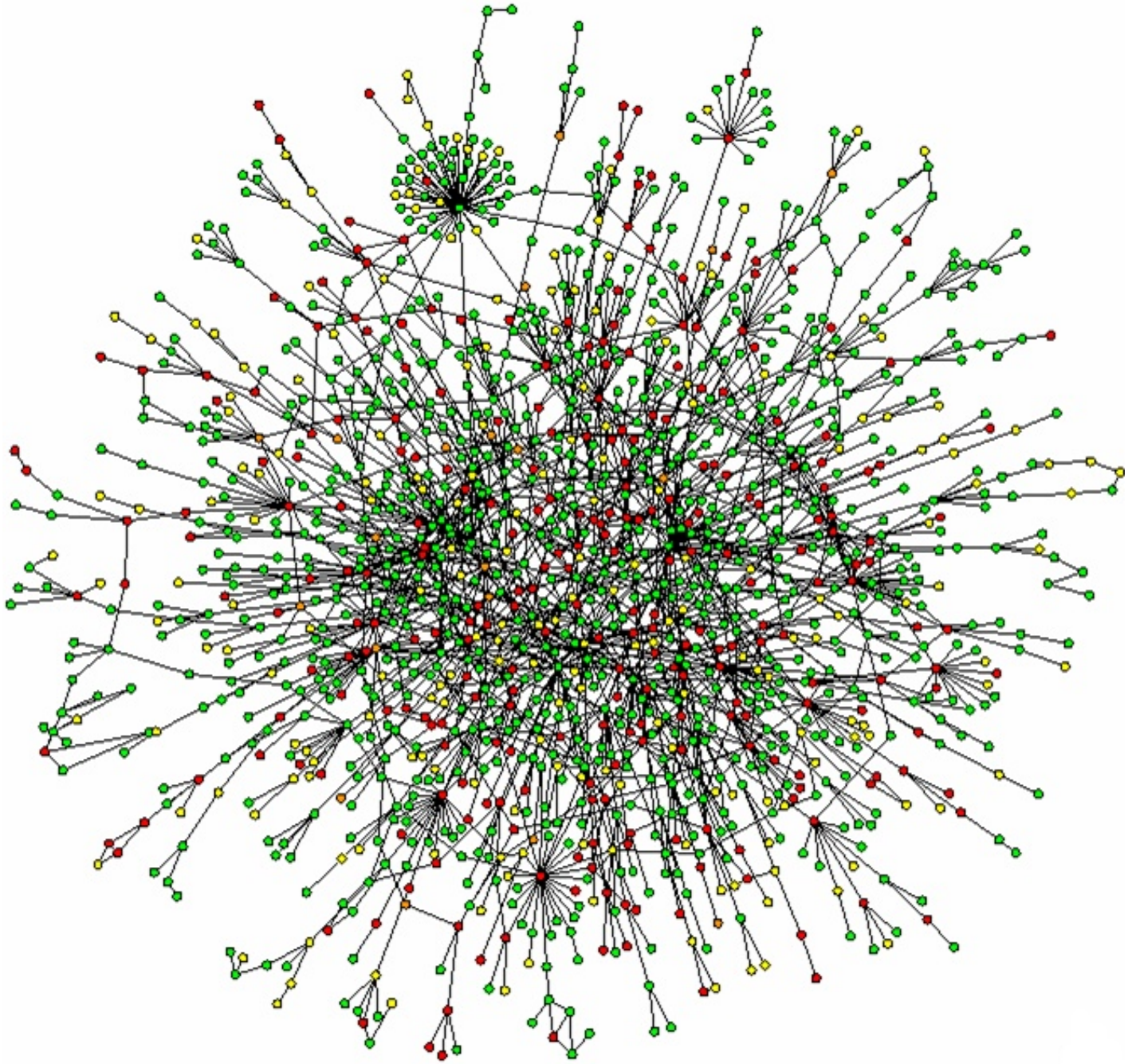
Qu'est ce qu'un réseau biologique ?

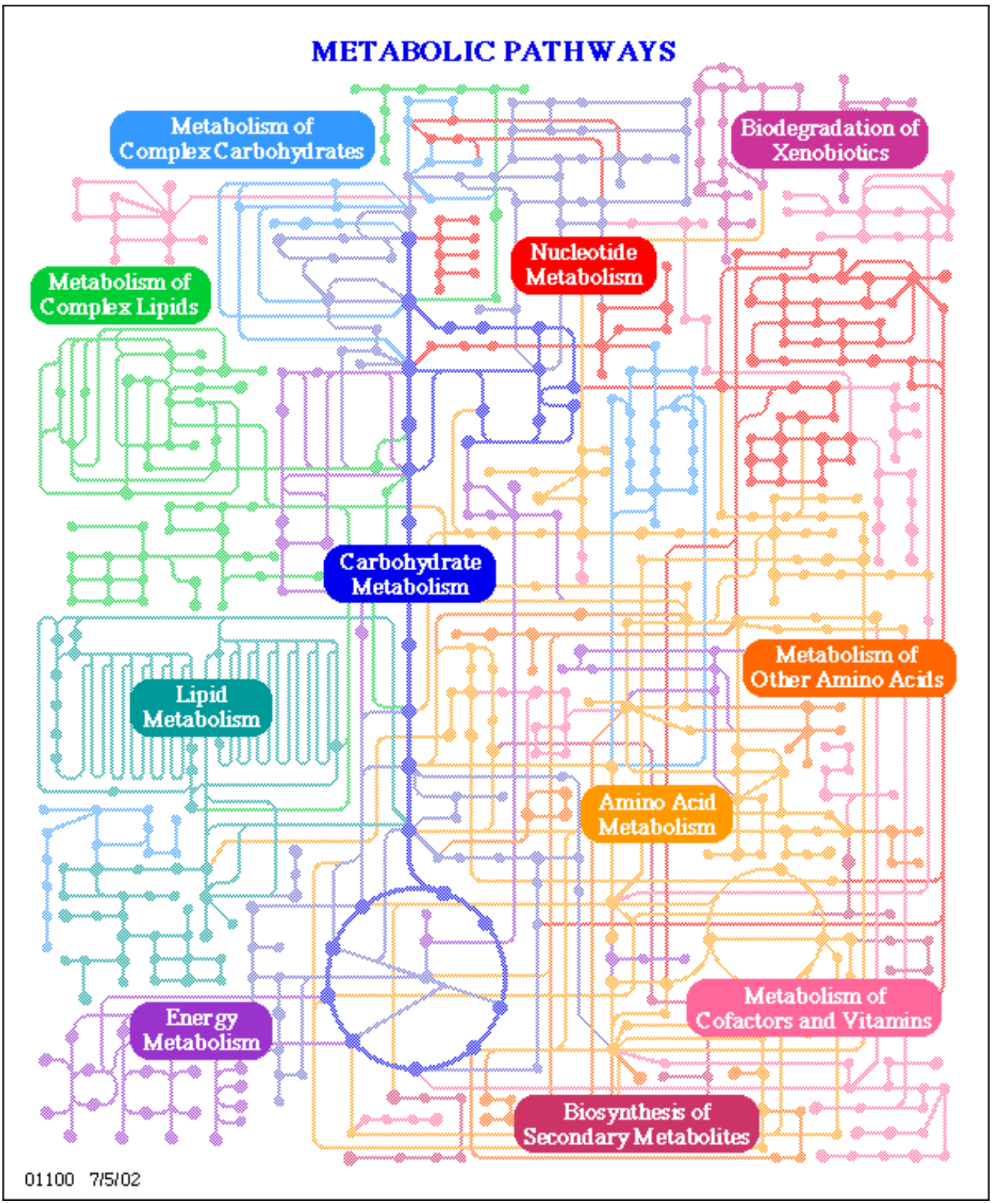
Un réseau biologique est une représentation de la circulation d'un certain type d'information dans la cellule. Il en existe plusieurs types :

- Réseau génétique ou de régulation : le gène A régule l'expression du gène B
- Réseau d'interaction protéine-protéine : La protéine A interragit physiquement avec la protéine B
- Réseau de signalisation : la protéine A transmet un signal informatif à la protéine B
- Réseau métabolique : l'ensemble des réactions chimiques dans une cellule

Quelques exemples






METABOLIC PATHWAYS
Metabolism of Complex Carbohydrates
Biodegradation of Xenobiotics
Nucleotide Metabolism
Metabolism of Complex Lipids
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Metabolism of Cofactors and Vitamins
Energy Metabolism
Biosynthesis of Secondary Metabolites
01100 7/5/02

Problématique

L'idée de l'inférence est d'utiliser des données accessibles en très grande quantité, ainsi qu'un modèle, pour reconstruire la structure du réseau biologique considéré.
Les données accessibles à grande échelle sont les interactions protéine-protéine et le niveau d'expression de l'ensemble des gènes dans des conditions contrôlées.
Cela s'applique essentiellement au réseau de signalisation et de régulation.

Inférence de réseau de régulation

L'idée clef de tous les modèles est que si deux gènes sont exprimés de manière similaire, alors ils ont un régulateur commun (ou se régulent l'un l'autre). Les problèmes sont :

- Le bruit dans les données expérimentales
- La causalité (A régule B ou l'inverse ? Besoin d'autres données)
- La taille de l'espace des réseaux à explorer : combien pouvez-vous écrire de réseaux différents avec 3 gènes simplement [1] ?

1. On rappelle que l'on cherche à appliquer ce type d'approche à des réseaux de l'ordre de plusieurs milliers de gènes

Intérêt biologique

La connaissance globale du réseau de régulation d'un organisme permet d'appréhender comment des modifications génétiques précises peuvent influencer globalement la survie d'une cellule : il y a donc une vocation thérapeutique.

De plus, l'étude des réseaux inférés permet de comprendre comment les systèmes vivants sont organisés, et éventuellement de mettre à jour des propriétés fondamentales.

Analyse de robustesse

On cherche à savoir comment le réseau dans son ensemble réagit à la destruction d'un noeud, correspondant à la mutation inactivante du gène correspondant.
Il a ainsi pu être montré que les réseaux de régulation sont très robustes à des destructions aléatoires de gènes. Ces études montrent comment un concept développé en informatique et télécom a permis de poser une question biologique pertinente.

Recherche de motifs

L'un des axes de recherche sur les réseaux est de trouver les motifs récurrents qui les composent. Cette recherche nécéssite à la fois des algorithmes puissants pour explorer le réseau et des modèles biologiques pertinents pour déterminer si leur présence est significative.
Les motifs de régulation peuvent avoir des rôles fonctionnels : c'est la théorie de la modularité, qui dit que le même type de fonctionnement été copié et réutilisé de nombreuses fois au cours de l'évolution.

Réseaux booléens

On peut étudier la dynamique de réseaux de régulation par des techniques issues des systèmes dynamiques : un affecte à chaque gène une variable d'état, et on définit une loi de changement d'état pour tous les gènes en fonction de leurs entrées, à chaque pas de temps.
On peut ainsi étudier les cycles limites et/ou les états stables d'un réseau de régulation.

Réseaux métaboliques

Dans les réseaux métaboliques, le type d'étude le plus pratiqué consiste à prédire les conséquences de la modification de la capacité d'une enzyme en un point du réseau, sur le fonctionnement global.
On étudie également l'espace des solutions des paramètres enzymatiques du réseau, et l'impact de maladies sur la position dans cet espace.
Finalement on peut cherche à identifier les paramètres qui permettent de contrôler un output donné, par exemple la biomasse produite.