L’interactome correspond à une vue statique des interactions.

Obtenir la structure 3D d’une macromolécule

biologique:

๏Résonance Magnétique Nucléaire (RMN).

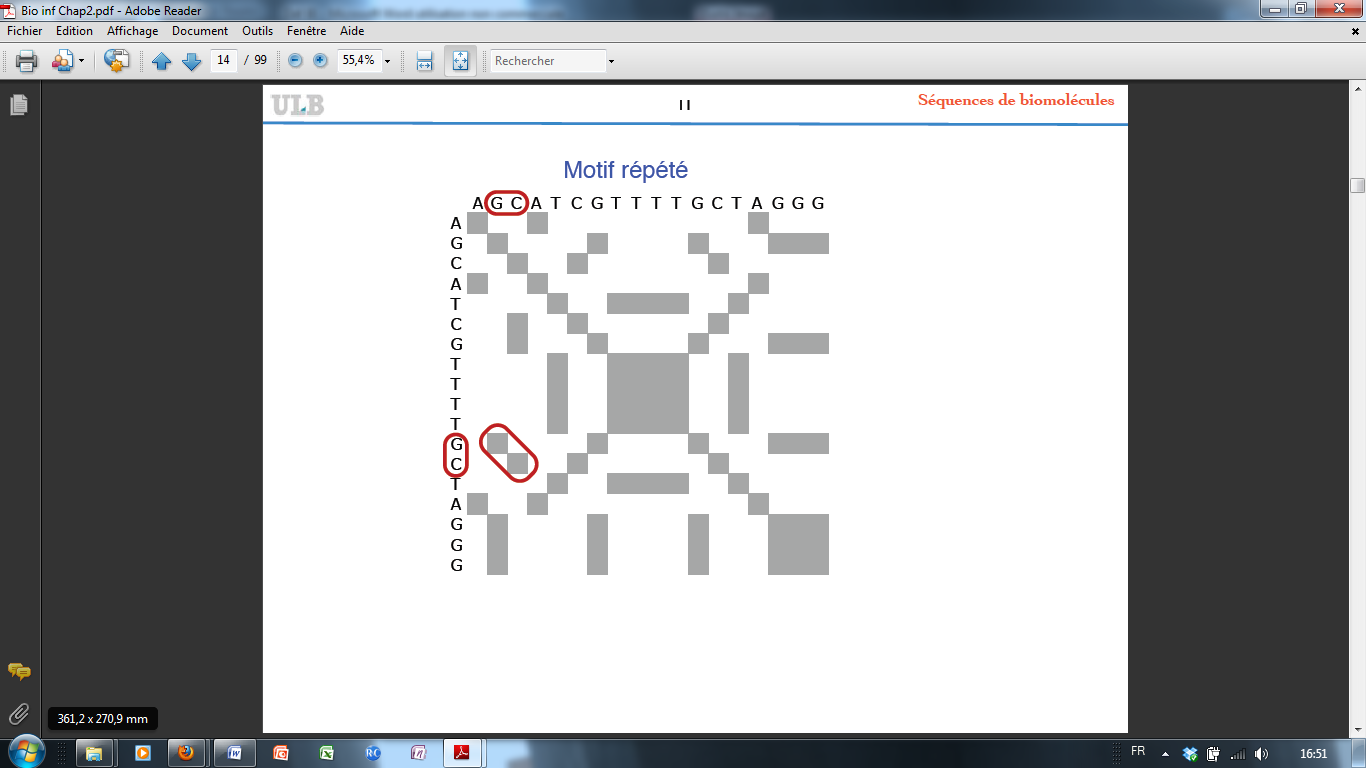
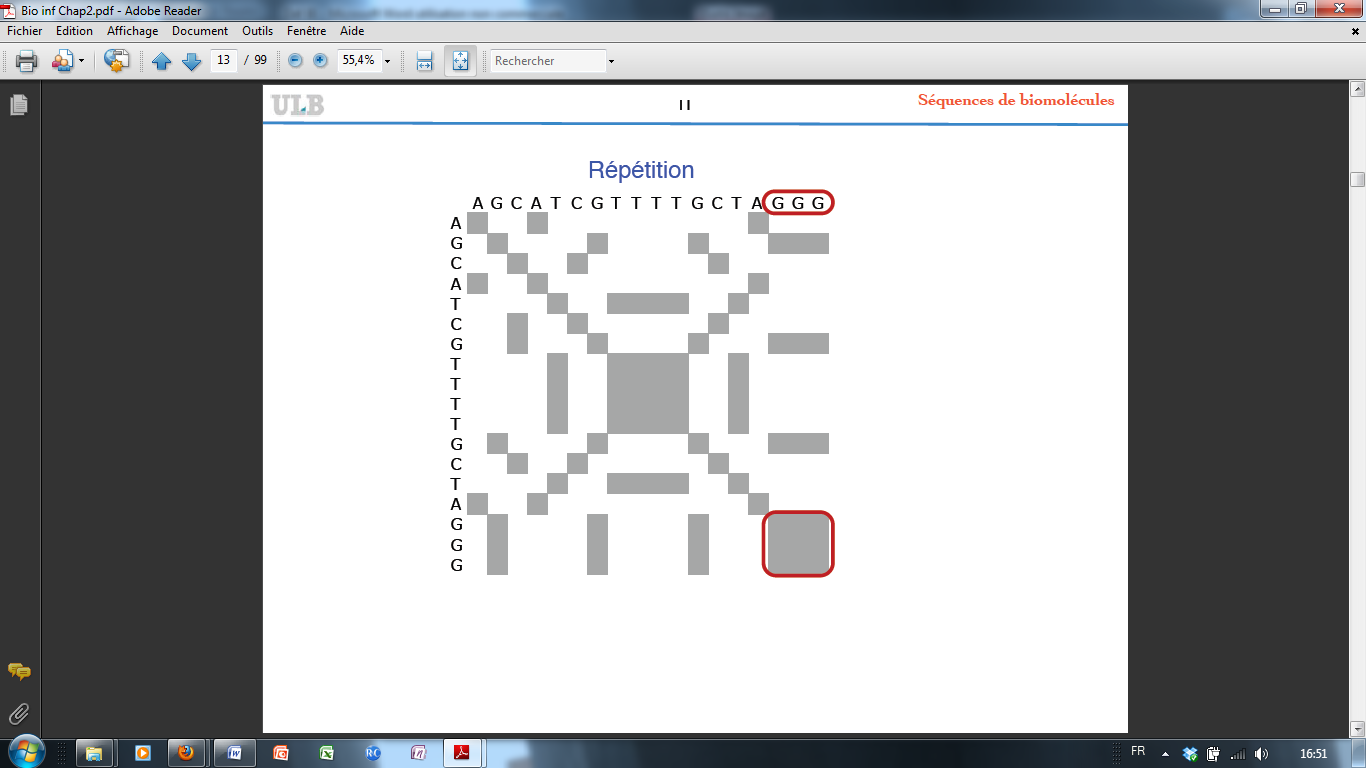
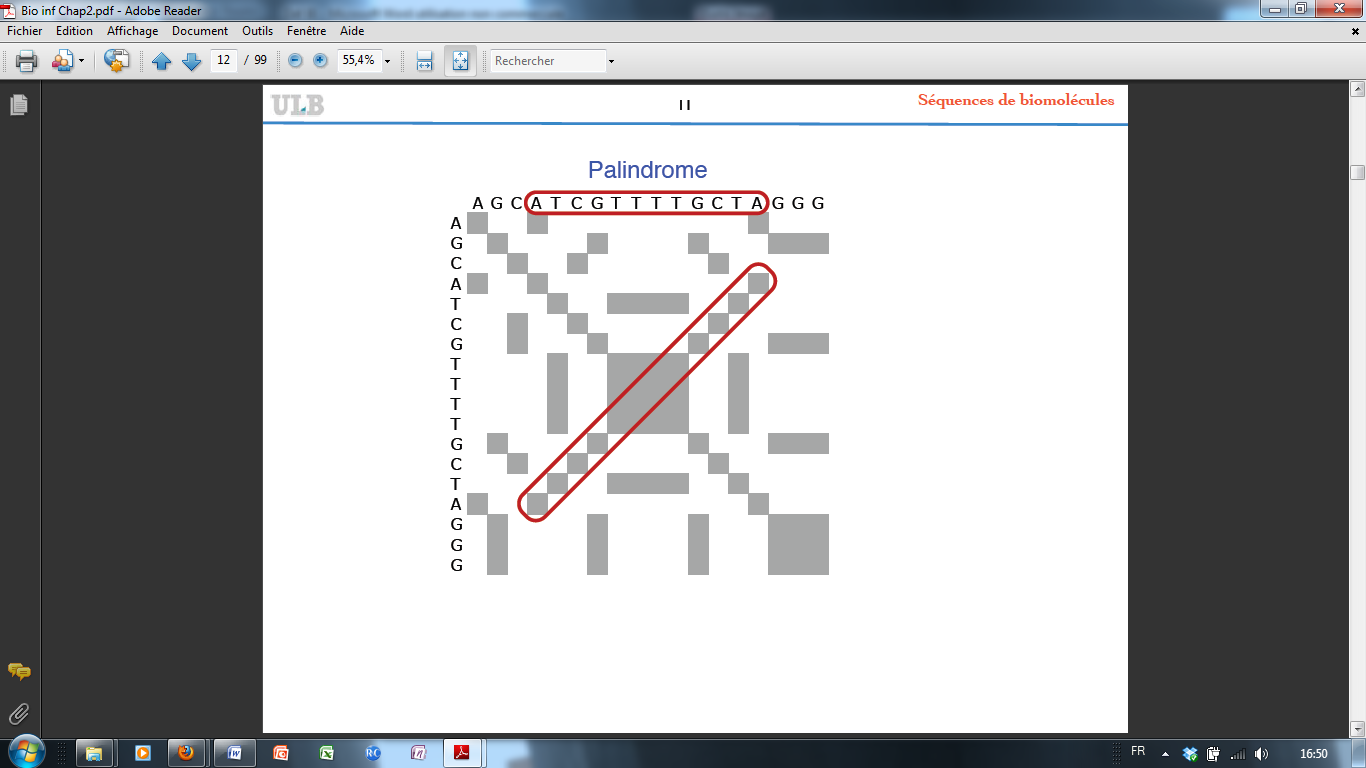
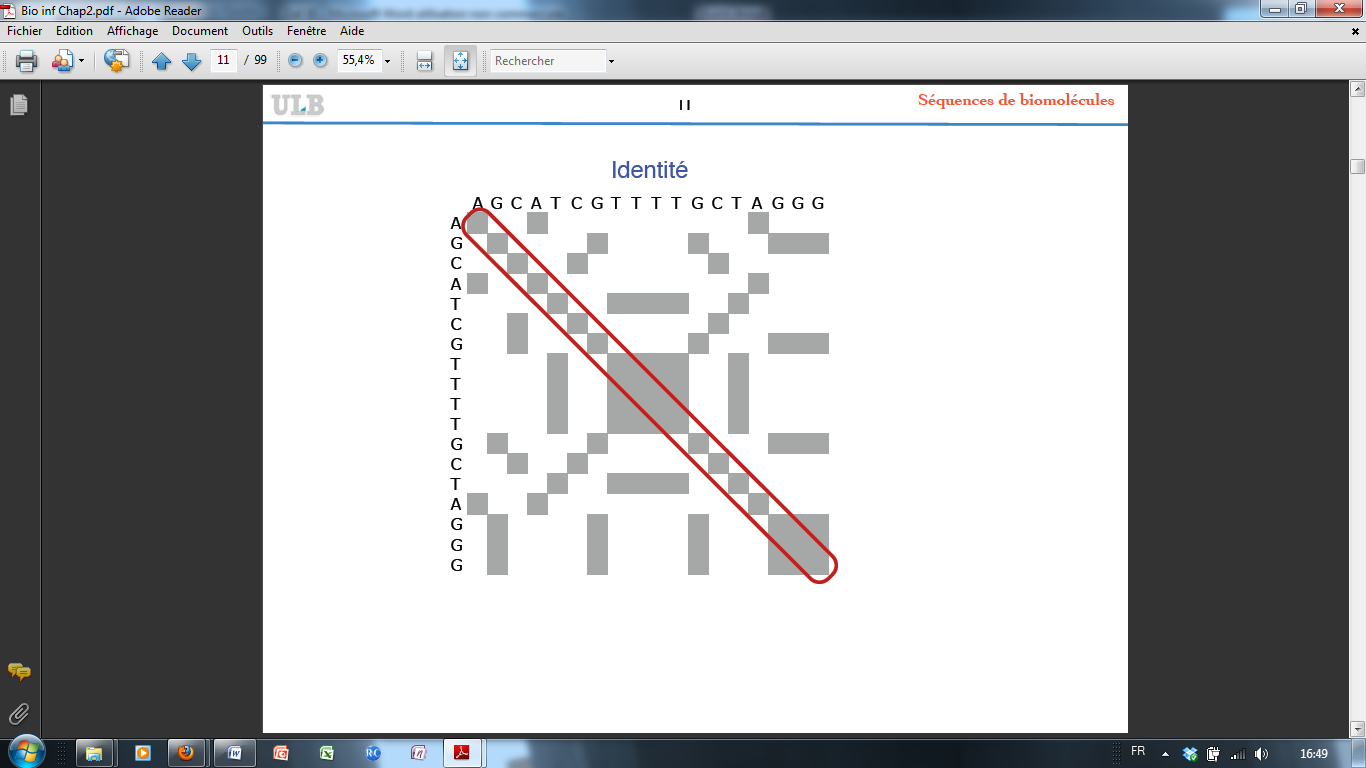
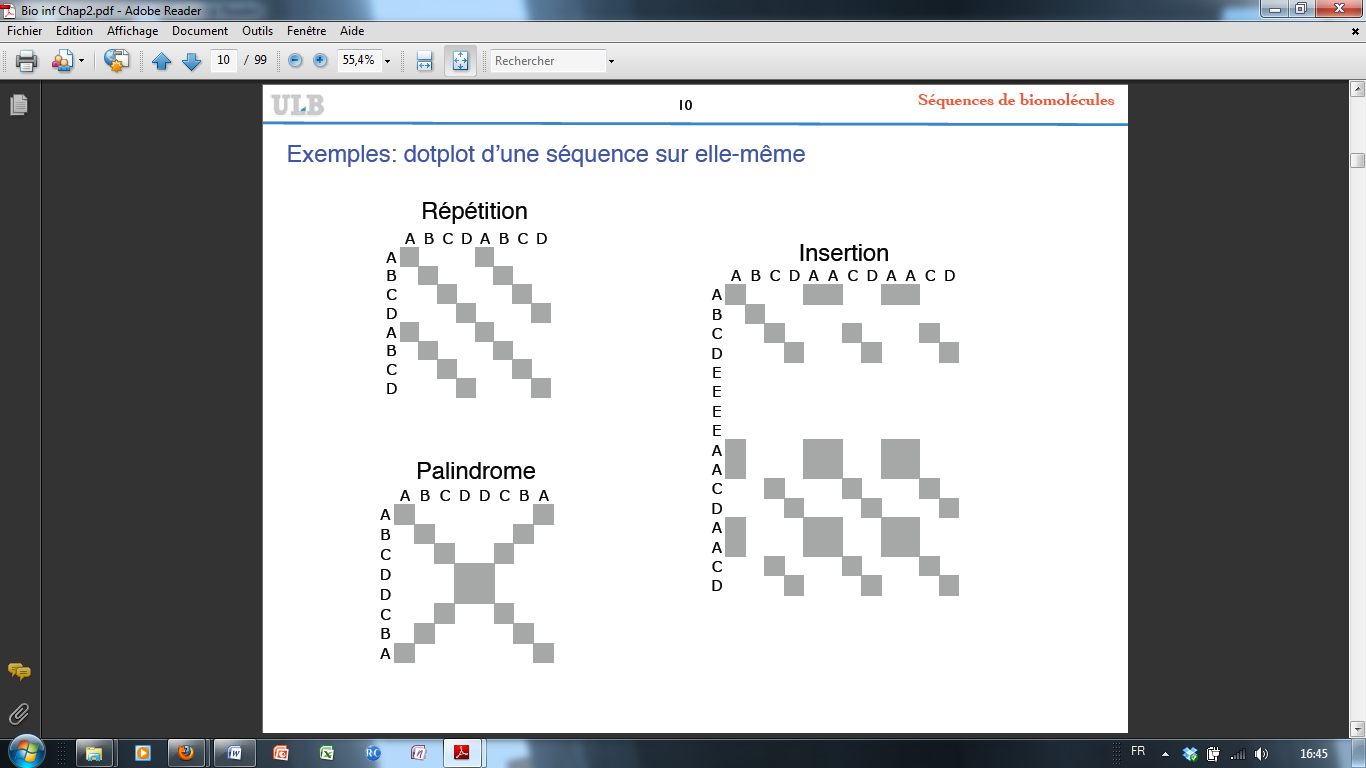
๏Cristallographie aux rayons X: nécessaire de cristalliser une protéine, puis étude de la diffraction des rayons X par le cristal. En déduit la position relative des atomes qui constituent la macromolécule.

Qualité d’une structure est évaluée par la résolution (plus elle est basse, meilleures est la résolution; typiquement: bonne résolution si elle est inférieure ou égale à 2,5 Å). Le facteur R indique dans quelle mesure les données récoltées correspondent au modèle de structure proposé (recherche d’un facteur R le plus bas possible).

Le “dotplot” est une représentation qui donne une vue d’ensemble des similarités entre 2 séquences. Il s’agit d’une matrice dont les colonnes correspondent aux acides aminés ou aux bases nucléiques de la première séquence et les lignes à ceux de la seconde séquence.

Principe général: les cases de la matrice sont remplies si une similarité entre les 2 séquences est détectée.

Informations qui peuvent en être déduites: vue d’ensemble rapide, zones où a des identités de séquence, insertions, palindromes,



En procédant de cette manière avec une séquence nucléique complète il faudra filtrer les résultats en choisissant une fenêtre (représentant un nombre d’acides aminés ). On compare chaque fois une fenêtre p.e 10aa à un endroit de la séquence complète. Puis on décale la fenêtre et on recompare. Quand cela correspond une ligne/case foncée est représentée. Donc au plus on élargira la fenêtre et cela correspondra tout de même, au plus on aura une plus grande similarité entre les deux séquences.

L’alignement de séquence consiste à identifier des motifs similaires/identiques dans des séquences et à les aligner.

Du dotplot à l’alignement: recherche d’un chemin dans la matrice du dotplot.

Mouvement diagonal: alignement; mouvement horizontal ou vertical: introduction d’un “trou”.

Alignement global : aligne les séquences complètes. Les algorithmes d’alignement global seront plutôt utilisés pour comparer des séquences homologues.

Alignement local : aligne des motifs locaux. Quand les 2 séquences n’ont que certains domaines/portions de séquence en commun ou quand on désire comparer une séquence avec les

séquences contenues dans une banque de données.

Alignement de paires: on compare deux séquences entre elles, et on identifie les modifications pour passer de l'une à l'autre.

Alignement multiple: on identifie les événements évolutifs qui distinguent un large groupe de séquences.

Matrices de score

Il faut avoir à disposition un système d’évaluation qui permet de prendre en compte la substitution d’un acide aminé/d’une base nucléique par un/une autre dans l’alignement, ainsi que les insertions et les délétions. Plus la fréquence de l’aa est basse (donc aa rare) plus son score sera élevé car la probabilité que cet aa ait été substitué là par hasard est faible.

Deux grands types de matrices permettant d’évaluer la similarité entre deux aa :

PAM : Percent Accepted Mutation. Matrice « passe partout » utilisée pour étudier des séquences complètes. Plus le chiffre derrière PAM est grand, plus on considère que les séquences ont divergées. Représentent les échanges possibles ou acceptables d’unacide aminé par un autre lors de l’évolution des protéines.Déduit la probabilité de remplacement d’un acide aminé par un autred’alignements de paires de séquences homologues. Une substitution entre

acides aminés équivalents devrait se produire plus fréquemment; le score dece type de changement devrait donc être élevé.

1) Alignements de séquences très homologues. Prend des séquences trèshomologues pour faire l’hypothèse qu’il n’y a pas eu de mutation multiple à unsite donné (ce qui biaiserait le calcul de probabilité).

2) Calcule une matrice de probabilité de remplacement d’un acide aminé parun autre durant 1 étape d’évolution; elle correspond à une substitutionacceptée pour 100 sites durant un temps d’évolution donné " 1PAM ((Construction d'une matrice de probabilité de mutation pour une distance d'évolution donnée))

3) Pour X mutations indépendantes: (1PAM)X.

4) Pour arriver aux matrices PAMX: divise la probabilité de mutation entre 2 acides aminés par le taux de mutation attendu de la fréquence relative des acides aminés, puis on en prend le logarithme. (PAMX=PAM100, PAM250, ...). Le score de l’alignement final peut être obtenu par somme des scores individuels des paires d’acides aminés alignés (Σ log Pi = log Π Pi )

BLOSUM : Block Substitution Matrix. Obtenue par une recherche des différences de séquences au sein de régions très conservées de familles protéiques. Matrice qui étudie seulement des bloques de séquences filtrées selon un pourcentage d’identité. Plus le chiffre de la blosum est faible, plus les séquence sont divergentes. Sélection d’un bloque de séquences qui partag un pourcentage de similarité de séquence (p.e. 62% pour blosum62)

1) Blocs de séquences collectés dans une base de données

2) Dérive une table de fréquences indiquant le nombre de paires d’acides aminés différents observés en association dans ces séquences conservées.

3) Normalise cette fréquence par la probabilité d’observer une paire d’acides aminés donnée et en prend le log.(Rapport entre Pab = probabilité d’observer les résidus a et b alignés dans les séquences homologues, et fa x fb = probabilité d’observer les résidus a et b en moyenne dans les protéines.)

Si la fréquence observée est inférieure à la fréquence attendue, l’élément de matrice est négatif.

4) Diverses matrices BLOSUM sont caractérisées par un seuil d’identité de séquence différent lors de la définition des régions conservées. BLOSUM62: pour définir la région comme conservée (définir un bloc), il faut 62% d’identité de séquence (! différence identité et similarité !).

les matrices BLOSUM sont un peu plus performantes, sont meilleures pour détecter des alignements

locaux;

La matrice BLOSUM62 est la meilleure pour détecter la majorité dessimilarités entre protéines;

La matrice BLOSUM45 conviendra mieux pour détecter des alignementslongs avec une faible identité de séquence.

Dans la plupart des alignements de séquences, il sera nécessaire d’introduire

des “trous”, pour prendre en compte les insertions/délétions. On prendra une pénalité d’initiation de “trou” et une pénalité d’extension de “trou” par résidu.

Alignement global

L’objectif est de trouver le meilleur alignement entre deux séquences complètes. De déterminer le pourcentage de position où les aa sont similaires. L’algorithme classique utilisé pour un alignement global est celui de Needleman-Wunsch.

1) On crée une matrice reprenant les 2 séquences et on la remplit à l’aide de la matrice de score.

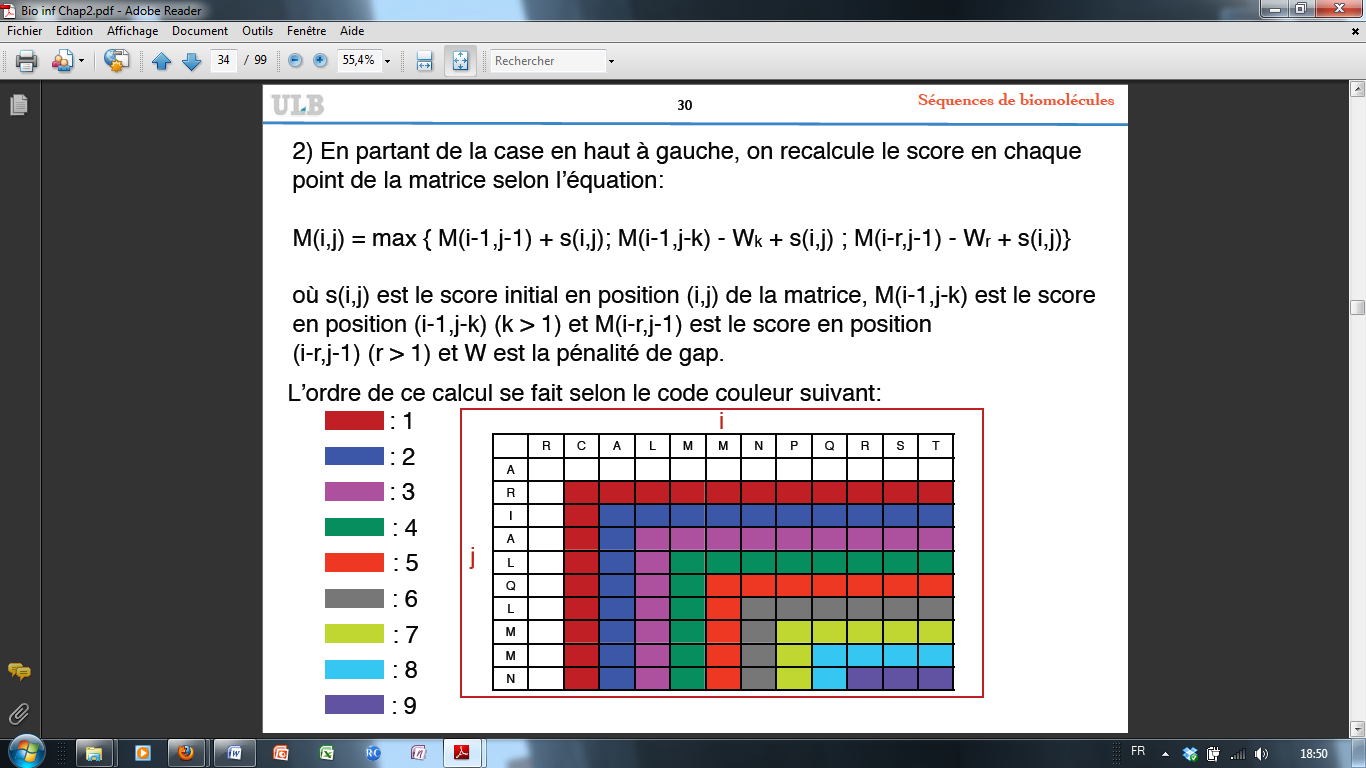
2) En partant de la case en haut à gauche, on recalcule le score en chaque point de la matrice selon l’équation:

M(i,j) = max { M(i-1,j-1) + s(i,j); M(i-1,j-k) - Wk + s(i,j) ; M(i-r,j-1) - Wr + s(i,j)}

où s(i,j) est le score initial en position (i,j) de la matrice, M(i-1,j-k) est le score en position (i-1,j-k) (k > 1) et M(i-r,j-1) est le score en position (i-r,j-1) (r > 1) et W est la pénalité de gap.

Au niveau des colonnes i, pour i-1 on recule d’une colonne vert la gauche, pour j-1 on monte de 1 ligne vers le haut.

L’ordre de ce calcul se fait selon le code couleur suivant:



L’étape suivante consiste à rechercher un chemin dans cette matrice, en partant du score maximum, et qui relie les scores les plus élevés. On recherche le chemin qui conduit au score total le plus élevé.

Déplacement diago alignement des 2 résidus

Déplacement vers le haut insertion d’un gap dans la séquence “horizontale”

Déplacement vers la gauche insertion d’un gap dans la séquence “verticale”

Construction de l’alignement : Commence à partir du score le plus élevé sur une ligne/colonne extérieure.Score total de l’alignement est cumulatif: somme le long du chemin. Le meilleur alignement a le score total le plus élevé.

Alignement local

L’objectif est de trouver des segments de séquences similaires.

L’algorithme classique utilisé pour un alignement local est celui de Smith-Waterman.

Le score en chaque élément de la matrice est calculé comme suit:

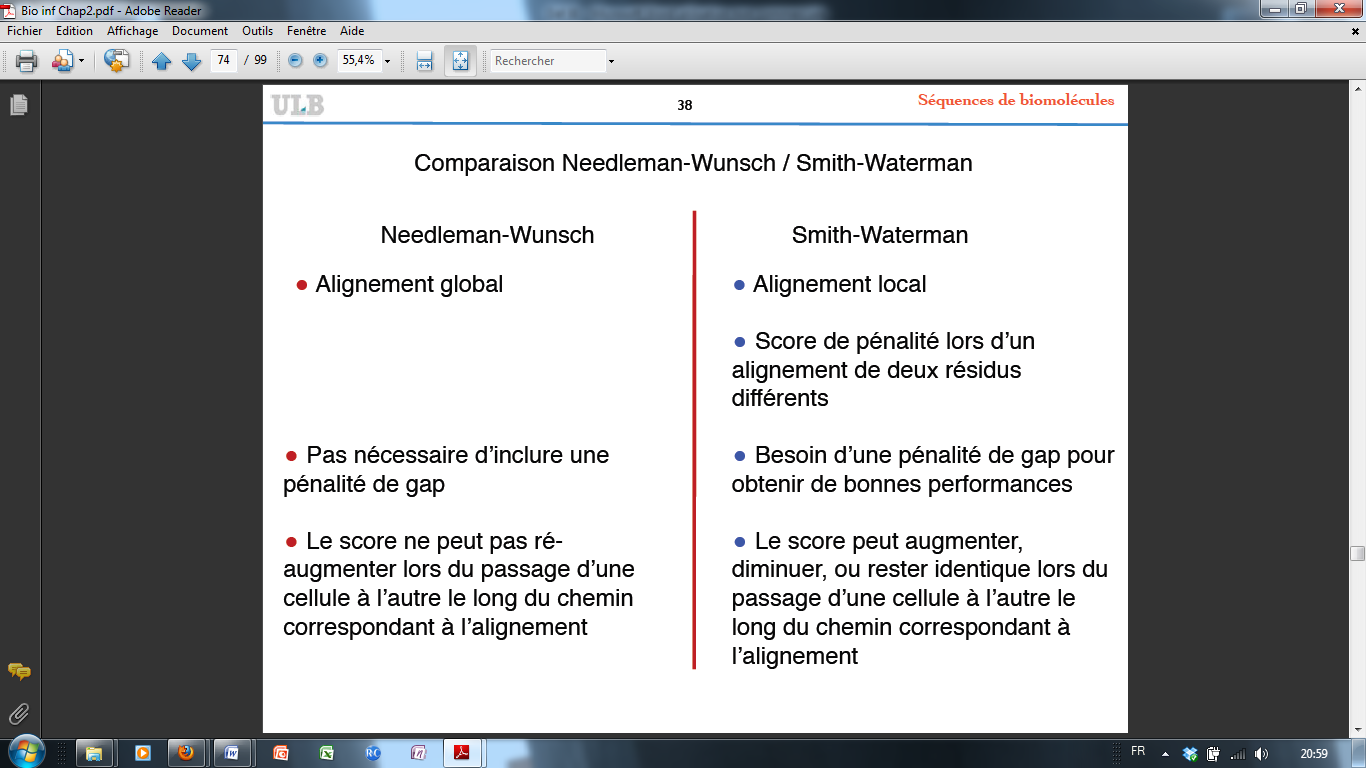
M(i,j) = max { M(i-1,j-1) + s(i,j); M(i,j-k) - Wk ; M(i-r,j) - Wr ; 0}

où s(i,j) est le score correspondant à l’alignement du résidu en position i avec celui en position j, M(i,j-k) est le score en position (i,j-k) (k > 0) et M(i-r,j) est le score en position (i-r,j) (r > 0) et W est la pénalité de gap (=1+0,3xk=valeur donnée pour la matrice) et k est l’extension de gap.

Dans ce cas-ci où s(i,j) = 1 si tes résidus sont les mêmes ou = -0,4 si tes résidus sont différents, Wk ici = 1,3 car=pénalité de gap=Wr (et k=1 pr calcul de la pénalité) selon moi toutes ces valeurs ont été choisies arbitrairement et sont seulement pour ce cas ci. Et sinon comme pour needleman pour le i-1,j-1 etc.

1) Démarre avec une matrice nulle; on place le score correspondant aux résidus identiques dans la 1ere ligne et la 1ere colonne.

2) On recalcul un score pour chaque case ac l’équation ci-dessus. Sachant les valeures pour : Résidus identiques (ici+1), Résidus différents (ici -0,4), Pénalité de gap (ici 1+0,3 x k)(k=extension de gap)



Alignement multiple

L’objectif est d’aligner plusieurs séquences en une même étape. L’alignement multiple permet de mettre en évidence des motifs d’acides aminés conservés.

Quelques exemples de l’utilité d’un alignement multiple:

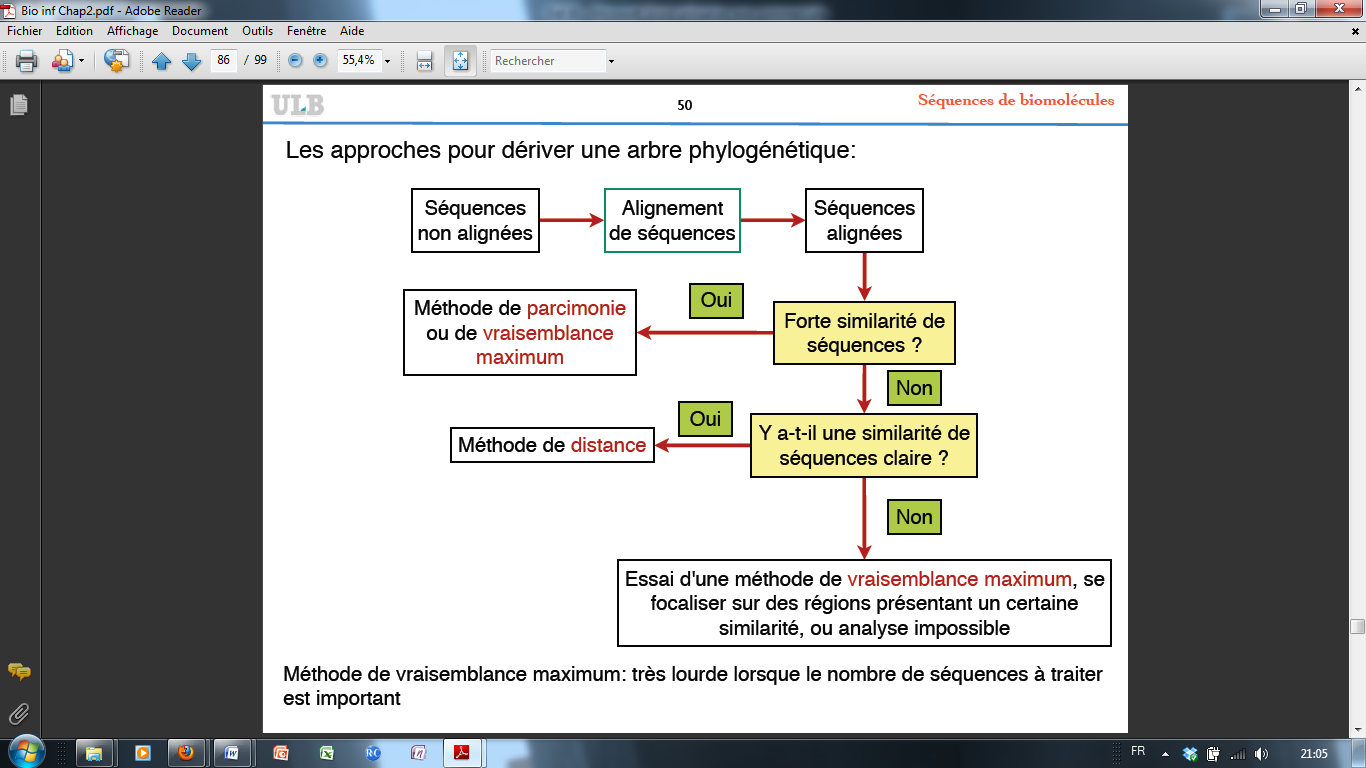
• Des relations de type “distances entre séquences” peuvent en êtredétectée et inférées.

• Les outils de prédiction de structure (voir chapitre 3) donnent de meilleursrésultats lorsqu’ils exploitent des alignement de séquence multiple.

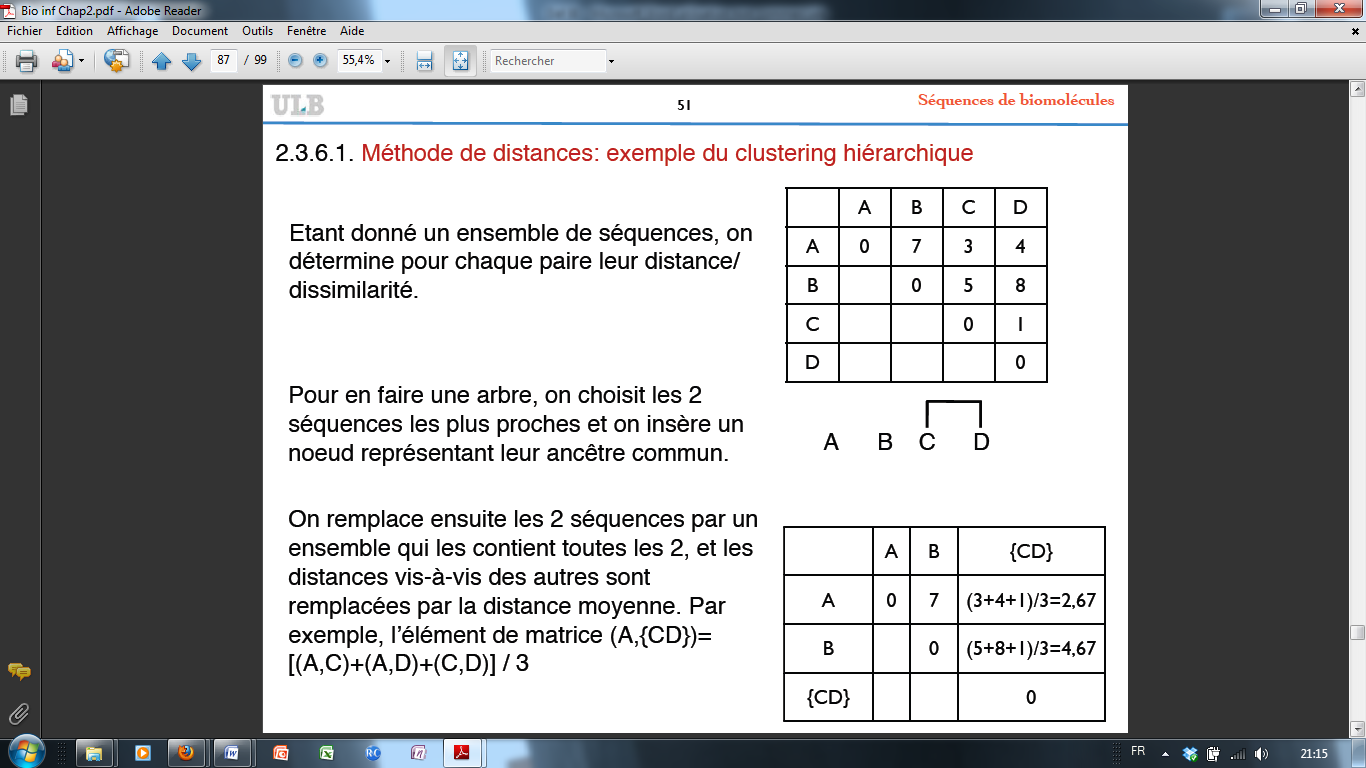
Le pourcentage d'identité de séquences est le pourcentage des positions alignées pour lesquelles l'acide aminé est identique dans les deux séquences.

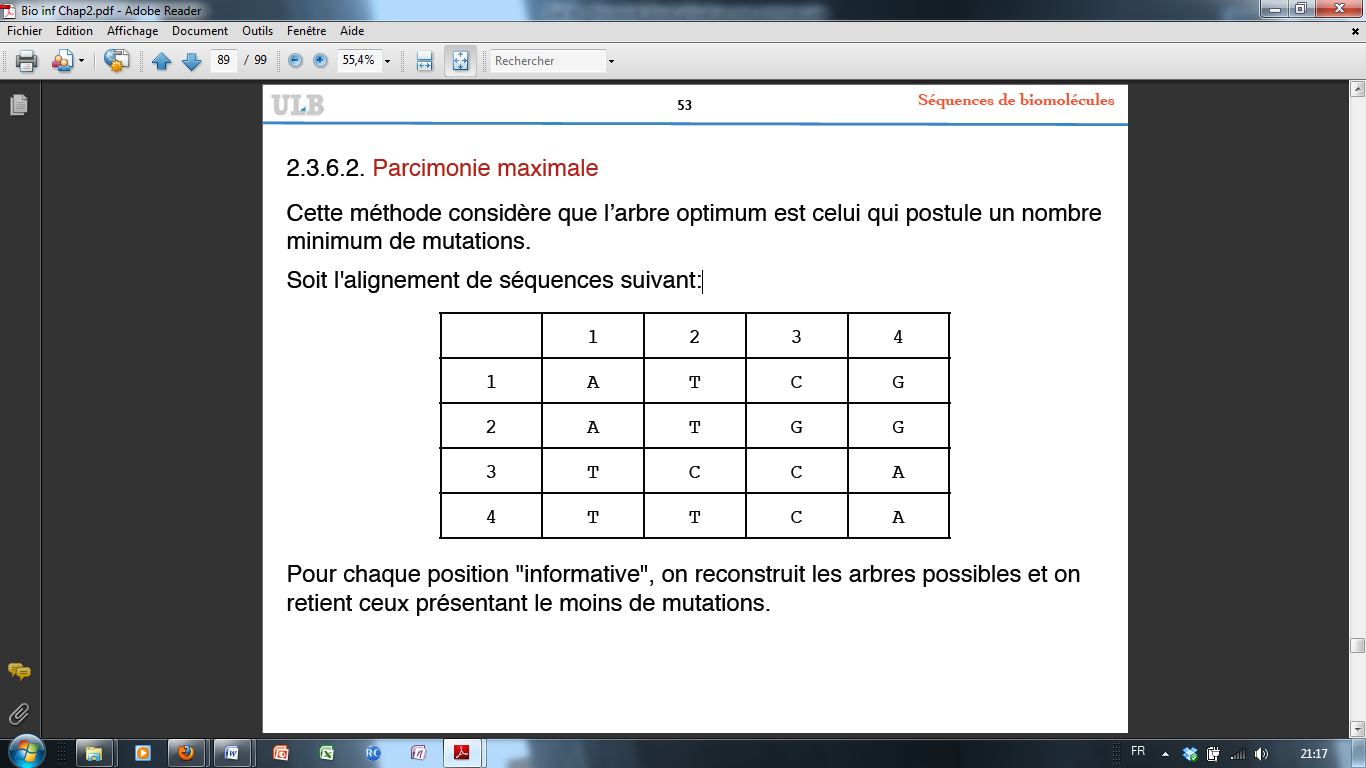
Le pourcentage de similarité de séquences est le pourcentage des positions alignées pour lesquelles l'acide aminé est similaire dans les deux séquences.

=> nécessité de définir un seuil de similarité (par exemple: si une matriceblosum62 a été utilisée pour générer l'alignement, on peut décider que deuxacides aminés sont similaires si le score blosum62 est positif).



Méthode de distances:

Exemple du clustering hiérarchique. Méthode de classification que l’on peut utiliser dans beaucoup de cas.Etant donné un ensemble de séquences, ondétermine pour chaque paire leur distance/dissimilarité.Pour en faire un arbre, on choisit les 2séquences les plus proches et on insère unnoeud représentant leur ancêtre commun.On remplace ensuite les 2 séquences par unensemble qui les contient toutes les 2, et lesdistances vis-à-vis des autres sontremplacées par la distance moyenne. Parexemple, la mesure de la distance entre A et [CD] =l’élément de matrice (A,{CD})=[(A,C)+(A,D)+(C,D)] / 3= (3+4+1) /3=2,67

Parcimonie maximale

Cette méthode considère que l’arbre optimum est celui qui postule un nombreminimum de mutations.Pour chaque position "informative", on reconstruit les arbres possibles et onretient ceux présentant le moins de mutations.

Seules les positions présentant des caractèresidentiques dans au moins 2 paires de séquences sontinformatives (voir colonne 2 et 3 par exemple: noninformatif), et permettent de faire la différenceentre les arbres.

Positions informatives: colonnes 1 et 4.

Vraisemblance maximale

Cette méthode assigne des probabilités aux mutations possible plutôt que deles compter. Les ancêtres sont ensuite reconstruits en chaque noeud, et unelongueur de branche est également assignée sur la base des probabilités desmutations considérées.

Pour toute topologie d’arbre, les vitesses de substitutions supposées sontmodifiées afin d’identifier les paramètres qui conduisent à la plus grandevraisemblance de produire les séquences observées. L’arbre optimal est celuiqui présente la plus grande vraisemblance de générer les données

observées.

Certaines méthodes prennent en compte l’horloge moléculaire (qui évite les biais dû au fait que les vitesses d’évolution de séquences sont différentes).

Recherche dans des bases de données :

Lorsqu’on veut savoir quelle séquence connue s’apparente à notre séquence X découverte. Utilisation d’un BLAST qui permet d’aligné une séquence donnée à toutes les séquences trouvées qui sont contenues dans une banque de données.

Evaluer la significativité statistique du résultat obtenu, d’un alignement

Le Z-score

zscore = S - < S >/ σS

Où S est le score, <S> est la moyenne des scores sur la population, et σS estl’écart type de la distribution.

Le z-score renormalise le score, il représente l’écart entre le score de l’alignement que l’on considère et la moyenne des scores obtenus pour des alignements sur la population, exprimé en unité d’écart type.

Plus le z-score est élevé, en valeur absolue, plus le score de la séquence estéloignée de la moyenne, en unité d’écart-type, plus la probabilité quel’alignement obtenu n’est pas dû au hasard est élevée. (si petit, proche de zéro et cela veut donc dire que le score trouvé n’est pas très éloigné du score moyen).

La “P-value”

Il s’agit de la probabilité que l’alignement n’est pas meilleur que l’aléatoire.Plus P est petit, plus l’alignement est significatif.

La “E-value”

Il s’agit du nombre d’alignements attendus pouvant présenter par chance unscore plus grand ou égal à celui obtenu avec l’alignement en question. Onrecherche donc une “E-value” la plus petite possible.Attention, cette valeur dépend de la taille de la base de données testée.

Structure de biomolécules

Alignement de structures protéiques

Aligner/superposer des structures sert à identifier les régions qui présententune structuration identiques et à classifier ces structures, à les regrouper endiverses catégories.

Lorsque les structures de protéines sont disponibles, l’alignement de structures,plutôt que l’alignement de séquences, est un outil qui permet la mise évidenced’une homologie. En effet, les structures tendent à moins diverger que lesséquences. Des protéines présentant une certaine similarité de séquenceprésentent des structures fort semblables. La détermination descorrespondances résidu-résidu par superposition de structure de protéines estune méthode puissante d’alignement de séquence.

De manière générale, au delà de 40% de similarité de séquence, les structuressont très semblables. Il y a des cas où la similarité de séquence est plus faiblealors que la similarité de structure est très élevée.

Une application de la classification de structure: prédiction de la fonction sur labase de la similarité de structure. Une grande partie des nouvelles structures ontleur site actif qui a été prédit par comparaison de structure.

Mesure de la similarité de structure

La mesure la plus couramment utilisée est le rmsd (root mean squaredeviation) après superposition.

Il s’agit de la distance moyenne minimale entre atomes correspondant après superposition des structures (racine carré pcq on calcule la distance entre deux points). Rmsd<3-4 A, structures relativement similaires MAIS cela dépend de la taille des protéines qu’on compare.

Homologie

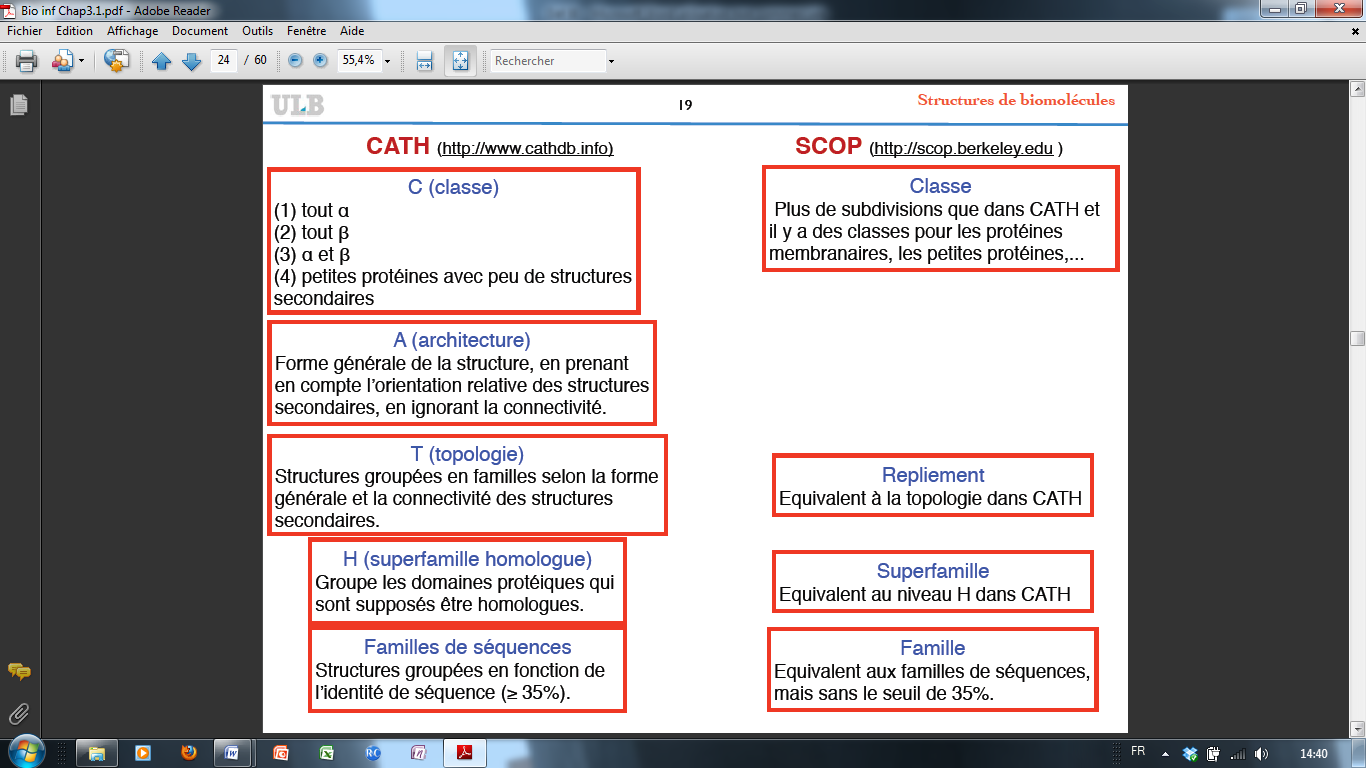
-Au delà d’un certain niveau de similarité structurale, même si la similarité deséquence est insignifiante.

-Conservation de caractéristiques structurales moins fréquentes: par exempleun tournant important dans la structure tel une unité βαβ gauche.

-Identité de séquence basse, mais significative, calculée après superpositionde structure.

- Présence des résidus clé au niveau de l’activité, même en absence desimilarité de séquence globale.

-Transitivité: A et B sont homologues, B et C aussi, alors A et C le sont.



Prédiction de structure de protéines

Le but est d’étudier ses propriétés, ses modifications caractéristiques physico-chimiques, et de pouvoir concevoir des molécules qui pourrait interagir avec la protéine (médoc).

Deux approches sont envisageables pour déterminer la structure 3D d’une protéine

- Techniques expérimentales (RMN, cristallographie RX)

- Prédiction in silico (bioinfo). Prob : grand nombre de conformations possibles, la structure 3D correspond en général à une des confo le minimum d’E libre, le temps nécessaire pour qu’une protéine adopte sa structure 3D=1ms à 1s.

Effet hydrophobe: le fait que certaines chaînes latérales des acides aminés sont hydrophobe est la source d’un des principe qui gouverne la structure des protéines. Cet effet hydrophobe est dû à l’eau, qui est un solvant assez structuré (ponts hydrogènes) et au fait que d’y placer un acide aminé apolaire empêche la formation de ponts H solvant-soluté, et renforce la structuration de l’eau. Ceci provoque une diminution de son entropie, et est défavorable pour l’énergie libre du système (ΔG=ΔH-TΔS). Pour limiter cela, les molécules apolaires (hydrophobes) ont tendance à se grouper, afin de minimiser la surface en contact avec le solvant.

Fonctions d’énergie : évalue, pour chaque méthode de préditcion de structure, la compatibilité entre la séquence dont on veut prédire la structure et diverse conformations envisageables (peut aussi utilisé une fonction de score au lieu d’une fonction d’énergie). Deux grandes classes :

-Potentiels semi-empirirques

-Potentiels dérivés d’une base de données contenant des protéines de séquences et de structures connues.

Potentiels semi-empiriques

Ces potentiels correspondent à une expression analytique qui décrit les diverses interactions inter-atomiques. Leur forme générale est: E=Eliaison + Eangle + Etorsion + Enon-liés + Eautre

P.e. Eliaison : Cette équation modélise, décrit, la variation de l’énergie en fonction de la distance par rapport à la distance “optimale”.

Obtention des paramètres d’une fonction d’énergie semi-empirique: soit par calcul de mécanique quantique, soit par optimisation empirique de manière à ce que les calculs correspondent aux mesures expérimentales.

=> certains champs de force semi-empiriques sont plus adaptés à l’étude de certaines classes de (bio)molécules.

Potentiels dérivés d’une base de données

Principe: dériver des paramètres énergétiques / des scores au départ de fréquences d’observation d’associations entre des éléments de séquence et de structure au sein d’une base de données de protéines dont on connaît les séquences et les structures.

Fréquence d’association entre des paires d’acides aminés et des domaines de distance.Fréquence d’association entre des acides aminés et des domaines d’angles de torsion de la chaîne principale. On relie ces fréquences d’observation à un paramètre énergétique au travers de la relation de Boltzmann.

Prédiction de la structure secondaire de protéines

Prédire à partir de la séquence, quelles portions se structureront en héliceα, brinβ en fonction de la « préférence » des aa pour certaines structures.

Il existe 3 grandes cathégories de prédiction de structures :

1) Méthodes physico-chimiques/empiriques

Elles dérivent des règles sur la base d’une analyse de structures 3D connues en mettant en relation certaines propriétés physico-chimiques des acides aminés et la structure secondaire.

2) Méthodes statistiques

Ces méthodes se basent sur la recherche de relations statistiques entre séquence et structure secondaire adoptée dans une base de données de protéines de structures connues.

p.e. GOR :Calcule la propension d’un acide aminé à adopter une conformation, sachantcelle de ses voisins. Prend en compte : information qu’un résidu porte à propos de sa structure secondaire, à propos de la structure secondaire d’un autrerésidu, quelle que soit la nature du résidu, à propos de la structure secondaire d’un autrerésidu, sachant qu’il est de nature sj: P(ci|si,sj) (pas dans la version originale).

3)Méthodes d’intelligence artificielle

Par exemple, apprentissage par machine (machine learning), réseaux de neurones artificiels. Le principe: l’agorithme apprend à reconnaîtres des motifs complexes dans une base de données d’apprentissage (ici: apprend à reconnaître des associations séquence-structure secondaire dans une base de données de protéines de structures connues).

-Des unités reçoivent des données pondérées en entrée et renvoient des signaux de sortie

-entraînement du réseau pour reconnaître des motifs de structure secondaireet trouver des poids qui optimisent les sorties avec les entrées;

- organise le réseau en couches.

Phase d’apprentissage: une entrée est fournie; les valeurs introduites sont combinées et multipliées par des fonctions (s ety dans l’exemple); les poids et biais (w et b dans l’exemple) sont initialisés aléatoirement; w et b sont ensuite optimisés pour que l’output soit aussi proche quepossible de la vraie solution.

Phase de prédiction:Les valeurs des paramètres obtenus à l’étape précédente sont utilisés pour

prédire la structure secondaire d’une séquence donnée.

Pour évaluer la fiabilité on peut tester sur des séquences-protéines connues (mais attention celles-ci doivent être différentes de celles utilisées pour établir la méthode).

Prédiction de la structure tertiaire de protéines

L’objectif est donc de trouver la structure qu’adopte une séquence protéique donnée.

Le problème principal est qu’il existe un trèsgrand nombre de conformations possibles,mais la protéine en adoptera, en général,une. Celle-ci correspond au minimumd’énergie libre.

La fonction d’énergie ou la fonction descore servira à discriminer les diverses conformations possibles. Il existe 3 méthodes. Le choix d’une méthode se fera en fonction des informations disponibles à proposde la protéine cible:

Si celle-ci présente des similarités de séquence élevées avec des protéines dont lastructure est connue, on utilisera préférentiellement la modélisation comparative(méthode qui conduit aux meilleurs résultats).

Si ce n’est pas le cas, la reconnaissance de repliement sera testée.

Enfin, la prédiction ab initio est très clairement l’approche la plus difficile. Mais elleprésente l’avantage de pouvoir prédire la structure de protéines dont le type derepliement n’est pas encore connu et qui ne présentent pas de similarité deséquence avec des protéines de structure connue.

1) Modélisation comparative : Cette méthode se base sur le fait que l’espace des conformations

possibles est plus petit que l’espace des séquences.En effet, en consultants des bases de données de domainesstructuraux, par exemple, on remarque que des protéines quiprésentent des séquences similaires adoptent des structures similaires.

Recherche dans une base de données de protéines dont la structure estconnue des protéines dont la séquence est similaire à celle de la cible. Ellesconstitueront des patrons (templates). On construit la structure de la cible (proté inconnue) sur la base de celle des patrons.L’alignement de séquence doit être bon pour avoir une bonne qualité de modèle final.

Deux conditions pour construire un bon modèle:

1) la similarité entre la séquence de la cible et celle de la structure doit être

suffisante;moins de 40% d’identité de séquence, l’alignement n’esten général pas optimal. L’alignement devient réellement difficile s’il y a moinsde 30% d’identité de séquence.

2) l’alignement entre ces séquences doit être suffisamment correct.

Deux grandes méthodes de modélisation comparative:

-Assemblage de corps rigides: consiste à assembler un petit nombre decorps rigides, qui correspondent aux régions conservées dans l'alignement deséquences. Se base sur la subdivision des structures en régions de cœur conservées, de boucles variables et en chaînes latérales.

-Modélisation par satisfaction de contraintes spatiales: génère descontraintes sur la structure de la cible, en utilisant l’alignement avec lesstructures des patrons. On suppose ainsi que les distances et angles entreacides aminés alignés sont identiques. Puis le modèle est dérivé en minimisantles violations à toutes les contraintes.

Modélisation des boucles optimisé en travaillant avec des boucles déjà connues, risque d’erreur plus importante dans ces régions là.

Placement des chaines latérales via une fonction de score qui permet de distinguer les diverses solutions possibles.

Evaluation de la qualité du modèle

Il existe diverses méthodes d’évaluation de la qualité d’un modèle

- sur la base de la stéréochimie (longueurs de liaisons, valeurs d’angles,planarité des cycles, chiralité, ...)

-par des méthodes de profils 3D ou des potentiels statistiques: évaluation del’environnement de chaque résidu en comparaison de celui observé dans desstructures RX de bonne qualité.

2) Reconnaissance de repliement

Si aucune protéine de séquence similaire à la cible n’est trouvée, on peutprocéder par reconnaissance de repliement.On évite une recherche dans l’espace des conformations complet enconsidérant que la séquence adopte un repliement connu.

La séquence cible va être enfilée sur des structures appartenant à une librairie de repliements connus càd qu’on peut « mélanger » des morceaux de structure provenant de proté différentes en insérant des gaps (ms pas trop non plus !).

L’énergie de toutes les associationsséquence-structure est évaluée à l’aided’une fonction d’énergie.

**!!! 3)**Modélisation ab initio

Cette approche consiste à effectuer une recherche dans l’espaceconformationnel de la protéine. Comme dans les approches précédentes, on se base sur l’hypothèse que toute l’informationnécessaire au repliement de la protéine est encodé dans la séquence.

Espace conformationnel=ensemble des conformations que peut adopter une protéine. Hypersurface multidimensionnelle obtenue en portant sur un graphe l’énergie libre en fonction des conformations adoptées par le système (spécifiées par les coordonnées des atomes).

Lorsqu’elle se replie, la protéine “se balade” sur cette hypersurface vers sa structure repliée, dépliée ou mal repliée.

Pour un système composé de N atomes, l’énergie est fonction de 3Ncoordonnées cartésiennes => pas possible de visualiser la surface complète.

De manière générale, on représente une section de cette hypersurface le long d’une coordonnée de réaction choisie judicieusement (exemple:angles de torsion de la chaîne principale, nombre de contacts natifs,compacité de la structure,....)

Dans le cas d’une protéine le principe est d’explorer l’espace conformationnel de la protéine et de rechercher le minimum d’une fonction de score. En général, cette fonction de score correspond l’énergie d’interaction inter-atomique(ou inter-résidu) de la protéine.

Utilisation d’une méthode qui va explorer cette surface le plus intelligemment possible (puisque impossible de tout tester)

Pour cette approche il est nécessaire de définir :

-le modèle utilisé pr représenter la protéine (chaîne latérale, ou juste chaine principale,..) ;

-une discrétisation de l’espace de conformation (ne pas prendre en compte toutes les valeurs mais seulement certaines p.e. que les nbres entiers) ;

-une fonction de score (associer un score à chaque conformation) ;

-une méthode de recherche dans l’espace des conformations (monte carlos, colonies de fourmis…)

Ex : utilisation d’un algorithme d’optimisation par colonie de fourmis

-Représentation des chaînes latérales par le centre géométrique des atomes de la chaîne latérale.

-Discrétisation des angles dièdres de la chaîne principale.

-Fonction de score: fonction d’énergie permettant d’évaluer l’énergie d’interaction entre les acides aminés.

L’algorithme se base sur l’observation que les fourmis trouvent toujours le chemin le plus court entre leur nid et une source de nourriture.

Principe: une fourmi dépose sur son chemin des indicateurs perceptibles par ses congénères: les phéromones. Lorsqu’une fourmi doit choisir un chemin, elle prend celui qui présente la concentration en phéromones la plus élevée. Conséquence: suite aux divers allers-retours nid-source de nourriture, le chemin le plus dense en phéromones est le plus court.

Application pour la prédiction de structure protéique :

-Mise en place d’une colonie de fourmis artificielles. Chaque fourmi construit une conformation pour la protéine de structure inconnue.

-Dépôt de phéromones pour les meilleures solutions. Les conformations sont évaluées à l’aide d’une fonction d’énergie.

- Création d’une nouvelle génération de fourmis artificielles, qui construisent de nouvelles structures en prenant en compte les phéromones déposées.

Ex. utilisation du recuit simulé par Monte Carlo

Recherche dans l’espace des conformations: recuit simulé par Monte Carlo.

- Modification aléatoire de la conformation d’un acide aminé. On peut commencer à haute température pour permettre au système de se ballader un peu n’importe où, en espérant qu’il trouve le minimum global. À ce moment on diminue la température afin de limiter les mouvements et que le système reste dans la région d’énergie minimale globale.

- Evaluation du score de la nouvelle conformation

- Accepte la nouvelle conformation selon la condition:ε ≤ exp (-ΔE / kT)

où ε est un nombre aléatoire compris entre 0 et 1, k est la constante de Boltzmann, T est la température, ΔE est la différence de score entre la conformation générée et la dernière conformation acceptée.

On décide si on accepte ou on refuse cette nouvelle conformation trouvée. On accepte d’office si l’E diminue. Si l’E augmente on peut accepter en attribuant une probabilité. La probabilité sera d’autant plus grande que la différence d’E entre les deux conformations (ancienne et nouvelle) est faible. (À basse température on accepte plus difficilement les mouvements qui augmentent fort l’E). Si on accepte que les découvertes de conformation où l’E a diminuée, on finit par trouver une conformation qui correspond à un minimum local duquel on ne peut s’échapper, mais il faut un minimum global. Pour éviter cela on accepte certains mouvement où l’E augmente mais moyennant cette équation.

E

-T est diminuée graduellement

Structure des protéines membranaires

Prédiction de segments transmembranaires et des orientations dans la membrane.

Helix bundles: prédiction et topologie

Les segments transmembranaires sont formés de long segments principalement

apolaires qui se replient en hélices α transmembranaires.

Graphe d’hydrophobicité (hydrophobicity plots)

Rapporte l’hydrophobicité en fonction de la séquence.

Les hélices transmembranaires sont d’abord identifiées au travers d’un graphemettant en évidence l’hydrophobicité locale des séquences d’acides aminés enfonction de la position.

=> choix d’une échelle d’hydrophobicité : Il en existe un grand nombre. On utilise une combinaison de diverses échelles.

Si négatif c’est plutôt région relativement hydrophile (inversement : + hydrophobe)

Aucune méthode basée sur une analyse de l’hydrophobicité n’est fiable à 100%pour identifier le nombre correct de segments transmembranaires. La plupart desméthodes n’est pas très efficace pour prédire la fin exacte des hélices.

TopPred:

-construction dʼun graphe dʼhydrophobicité;

-identification de segments transmembranaires “certains”, qui présentent unehydrophobicité plus grande quʼune valeur seuil supérieure C1;

-identification de segments transmembranaires potentiels, qui présentent unehydrophobicité plus grande quʼune valeur seuil inférieure C2;

-construction de toutes les topologies possibles incluant les régions “certaines” eten incluant ou excluant les segments potentiels;

-pour chaque topologie, calcul des différences entre le nombre de Arg+lys àlʼextérieur et à lʼintérieur de la cellule. Exclut les boucles C-ter de plus de 60 résidus,mais inclut les boucles N-ter quelle que soit leur longueur;

- choisit la topologie avec la différence la plus grande;

- sʼil y a beaucoup de boucles de longueur supérieure à 60 résidus, leurcomposition en acides aminés peut indiquer leur localisation intra- ou extracellulaire.

β-barrels: prédiction et topologie

La structure 3D de trois porines sont connues et permettent la prédiction des segments transmembranaires dʼautres porines sur la base dʼun alignement deséquences.

Une méthode plus générale se base sur le fait que dans les brinstransmembranaires les 2ème résidus sont hydrophobes et sont face au lipide; lapartie interne peut être hydrophobe ou hydrophile. De plus, chaque brin est encadrépar des acides aminés aromatiques.

Généralement, il y a un nombre impair de brins β; ceux-ci sont antiparallèles.Lʼangle est de plus ou moins 45° par rapport à lʼaxe de la membrane.

**Prédiction de structure dʼARN**

LʼARN présente une structure 3D qui lui confèredes fonctions complexes.

A la différence de lʼADN, lʼARN est simple brin => possibilité de se replier sur lui-mêmeet création de régions en double hélice séparée par des boucles.

Le ribose à la place dudésoxyribose modifie les propriétés de lachaîne principale de lʼARN et permet unevariété plus large de structures.A lʼappariement des bases classique (A-U) et (G-C), il peut aussi y avoir un appariement moins stable (G-U) ainsi que des appariement autres que Watson-Crick.

Distinction d’une structure primaire (séquence) et d’une structure secondaire (appariement des bases)

Pour la prédiction de structure secondaire d'ARN, je pense pas qu'il faille tout retenir, juste savoir que tu peux faire un alignement de séquences ou travailler avec des méthodes basées sur un calcul d'énergie et que ça, avant, on maximisait le nombre d'appariements mais ils ont remarqué que certaines structures déstabilisaient le shmilblic...

Et pour sub-optimal, je pense qu'ils font ça pour favoriser leurs résultats pcq s'ils font une structure optimale qui propose une structure similaire mais qui ne respectent pas certaines règles (énergie, angles etc), les résultats ne seront pas bons, malgré la similarité des structures alors qu'avec une structure sub-optimale, la similarité de structures est moins bonne mais les règles d'énergie etc sont respectées... Enfin je l'ai compris comme ça, je sais pas si c'est correct et je ne sais pas si ça t'aide ^^

Prédiction de la structure secondaire

Alignement de séquences

Cette approche se base sur une analyse phylogénétique de familles deséquences dʼARN. Elle a pour objectif de révéler les pressions sélectives qui ontmaintenu des relations entre paires de bases. Elle procède par une analyse de lacovariance de paires de bases.

Cette approche se place dans lʼhypothèse que les bases sont en appariementWatson-Crick. La méthode requiert également une identité de séquence de lʼordrede 60 à 80% au sein de la famille.

Si le nombre de séquences est suffisamment grand, cette méthode peut être trèsefficace.

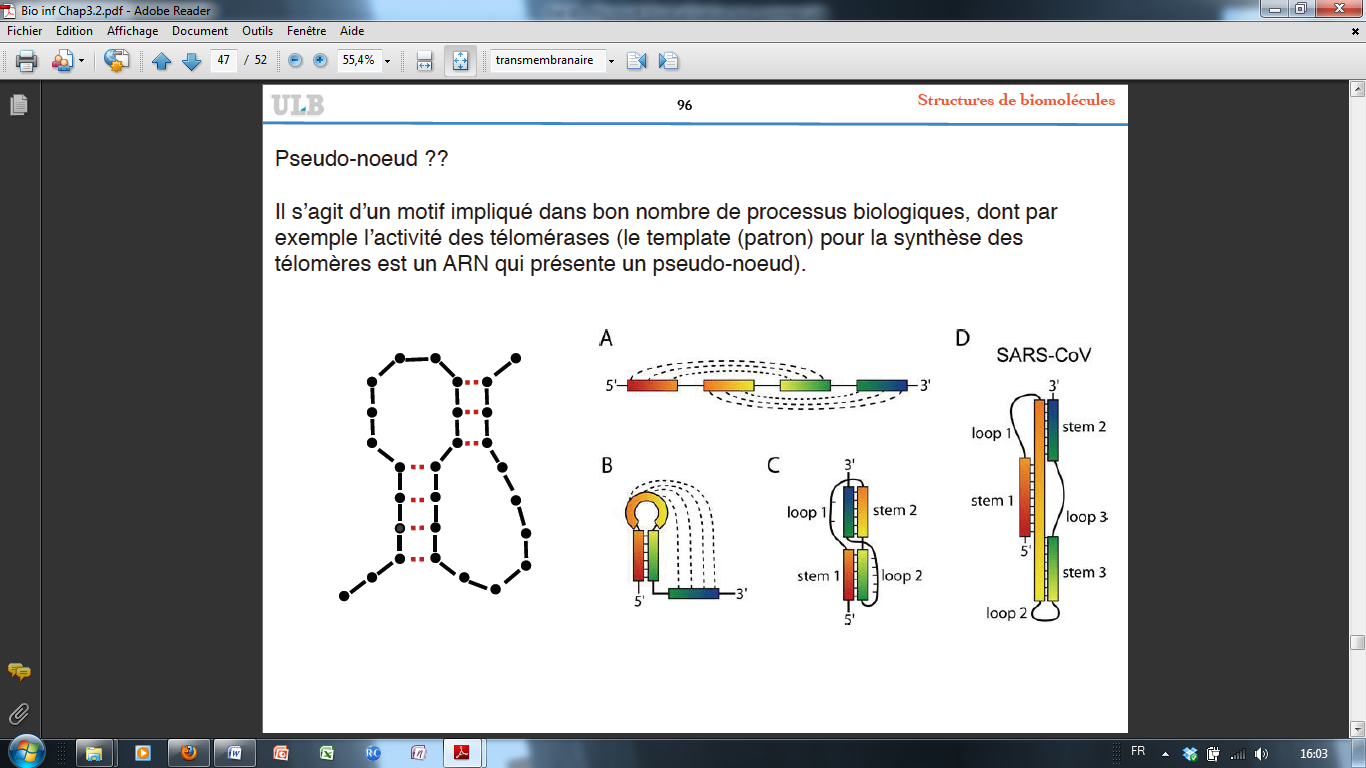
Méthodes basées sur un calcul dʼénergie

Lʼobjectif est de déterminer la conformation (les appariements de bases) quiconduit à lʼénergie libre minimale. Deux approches :

1) Maximiser le nombre dʼappariements.Ceci peut être raffiné en prenant en compte lʼénergie de déstabilisation, positive,de certaines structures particulières (hairpin, bulge, boucle intérieure, bouclesmultiples)

2) Minimiser lʼénergie de la structure = approche plus sophistiquée. On pourrait ainsi définir une énergie pour les appariements CG/GC, AU/UA et GU/UG. Et également ajouter des termes de déstabilisation pour certainesstructures (en fonction du nombre de bases non appariées, des pénalités d’asymétrie, du nombre d’hélices connectées, des énergies stabilisantes)

Problèmesde lʼapproche:

- la structure dʼénergie la plus basse nʼest pas garantie dʼêtre la native(thermodynamique <=> cinétique): pour de longues séquences (> 400 bases), il ya de grandes chances dʼerreur, avec tout de même des éléments structurauxcorrects (mais lesquels??);

-les programmes classiques ne prennent pas en compte les pseudo-noeuds,

quisont pourtant une caractéristique importante;

- si des séquences correctement alignées sont à disposition, des approches paralignement de séquences sont bien plus performantes; mais lorsquʼil y a peu deséquences et que leur alignement est compliqué, des méthodes plussophistiquées sont nécessaires;

- certains ARN se replient au contact de protéines ou dʼautres ARN, ce qui peutjouer un rôle important.

Solutions proposées pour aller au-delà de ces désavantages:

-les algorithmes calculent également les structures sub-optimales. Mfold parexemple propose des structures sub-optimales, en évitant que les structuressoient trop similaires. RNAsubopt fournit les structures sub-optimales jusquʼà uncertain seuil énergétique. RNAshapes: subdivise lʼespace des conformations enfamilles, caractérisées par un représentant.

- les pseudo-noeuds: des algorithmes ont été développé afin de les traiter(pknots par exemple). Se base sur des algorithmes de graphe ou sur desalgorithmes génétiques.

-lʼalignement de structures, sans passer par un alignement de séquencespermet également dʼaider la prédiction (RNAcast, RNAforster)

- lorsque les sites dʼinteractions ARN-autres molécules sont connus, cescontraintes peuvent être introduites dans lʼalgorithme (RNAhybrid).

Prédiction de la structure tertiaire d’ARN

Une des approches repose sur lʼanalyse de motifs récurrents au sein de structures connues dʼARN.Ces motifs résultent dʼappariements de Watson-Crick, mais égalementdʼappariement non-Watson-Crick (voir slides précédents).Cette approche recherche parmi une séquence dʼARN ces motifs structuraux.

La prédiction de structure secondaire peut être utilisée en complément.

Une autre approche consiste en la satisfaction de contraintes. Sur la base desséquences, dʼéventuelles données expérimentales, de prédiction de la structuresecondaire (et donc dʼappariement de bases), des contraintes sont dérivées. Puisune structure qui satisfait ces contraintes est construite.

Analyse et modélisation de systèmes biochimiques

Interactions entre biomolécules: analyse de lʼinteractome

Interactome: ensemble des interactions entre biomolécules, qui se produisent ausein dʼune cellule.

A lʼheure actuelle, il y a déjà de nombreuses données sur les interactions dans lesquelles lesdiverses biomolécules sont impliquées.

Permet une meilleure compréhension de lʼorganisme. Possibilité dʼinférer certainesrelations au départ dʼautres relations.

Principe : au départ dʼun graphe indiquant les interactions entre individus, nous déduisons denouvelles relations, sur la base de certaines règles.

Différentes méthodes permettent d’obtenir les données d’interactions :

* Littérature scientifique : il existe des techniques de bioinfo qui recueillent de manière automatisée des informations venant d’articles de la littérature.
* Expérience à grande échelle :

-protéine-protéine méthode de double hybride et purification de complexes : protéine appat(dont on veut déterminer les interactions) ac dom de fixation au prom et proté proie ac DA. Expression du gène si interaction.

-protéine-ADN : ChIP on chip : identifie quelle proté peut venir se fixer sur une zone précise de notre ADN

* Prédictions bioinformatiques : au fil de l’évolution, les gènes codants pour 2 proté qui interagissent entre elles peuvent se retrouver fusionnés dans d’autres organismes. Techniques utilisées dans les bases de données d’interaction proté-proté, proté-ADN

-“Pierre de Rosette”: deux protéines A et B qui sont fusionnées dans lʼorganisme 1sont susceptibles dʼinteragir dans lʼorganisme 2

-deux protéines dont les orthologues interagissent ont une grande probabilitédʼégalement interagir. (orthologue: similarité de deux gènes chez deux espèces différentes, et présentantune position équivalente dans le génome.)

MAIS les bases de données des interactions représentent cela sous forme statique, alors que pour la plupart des protéines ces interactions sont dynamiques (p.e. une protéine qui a 5 partenaires, n’interagit qu’avec 3 lors de la phase 1 de vie de la cellule, et qu’avec 2 lors de la phase 2 de vie de la cell)

Au départ de ces données on peut caractériser l’interactome

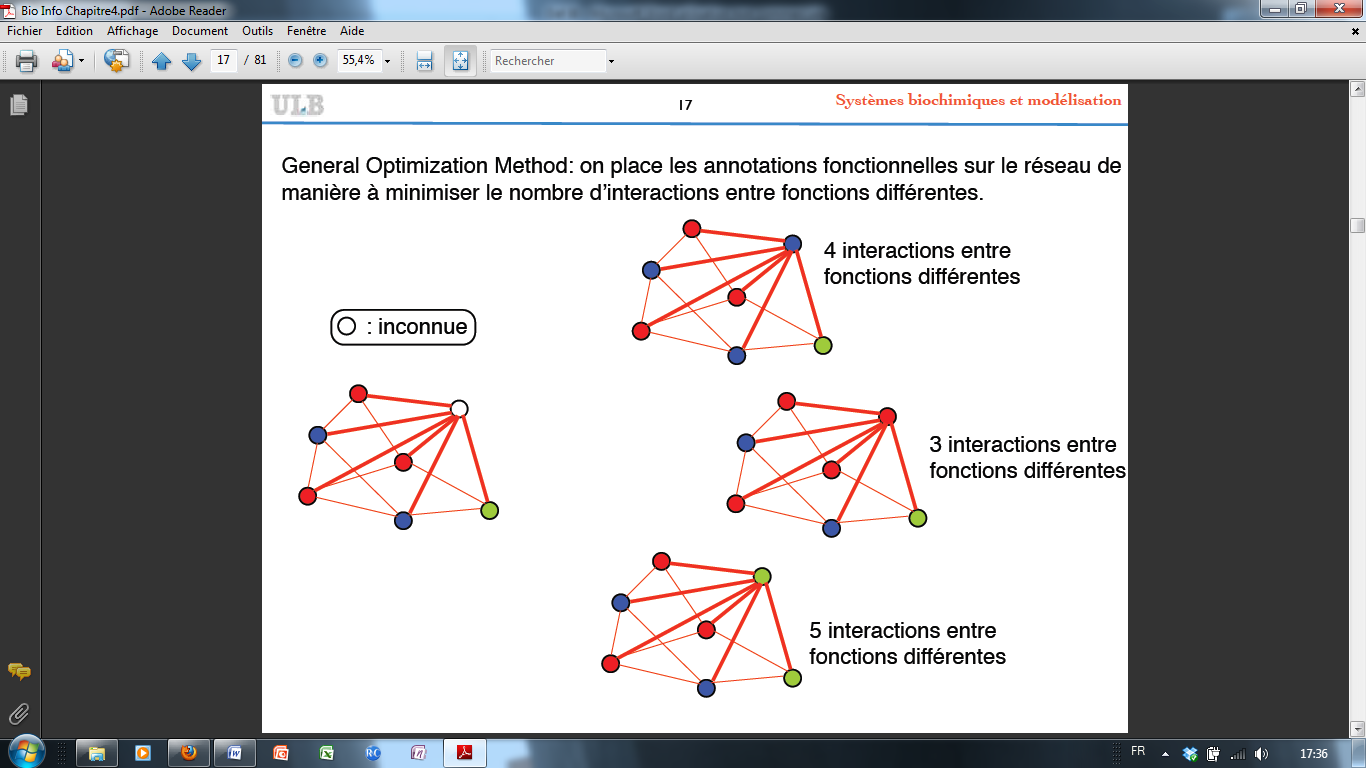
-Nombre de voisins en chaque point. Plus une protéine a de voisins, plus elle est essentielle. Calcul de « zones denses » = coefficient qui mesure la densité de connexion entre les voisins=

(nbre de connexions pr 1 proté donnée)/(nbre max de connexions rencontrées dans ensemble du graph)

Utilisation de lʼinteractome pour faire des prédictions: fonctions de protéines,appartenance à des voies de signalisation, ... On peut utiliser les réseauxdʼinteraction afin dʼinférer des fonctions.

On extrait des règles au départ dʼobservations, on suppose que ces règles sontuniverselles, on les applique pour prédire des fonctions chez des protéinesinconnues. Pas oublier une estimation du taux de réussite et de faussesprédictions.

Observation: les protéines de même fonction ont tendance à être en interactiondirecte les unes avec les autres => on peut déduire la fonction à partir de lafonction des voisines. Pour une protéine inconnue, on peut recenser sespartenaires et déterminer la fonction la plus fréquente parmi eux, et attribuer lafonction de la protéine inconnue sur cette base.

Exemple : General Optimization Method: on place les annotations fonctionnelles sur le réseau demanière à minimiser le nombre dʼinteractions entre fonctions différentes.On considère chaque cas : si la protéine avait la même fonction qu’une des protéine (p.e la bleue) combien d’interaction avec les mêmes protéines (les autres bleues) ferait elle ? On veut maximiser les inetractions entre fonctions les mêmes (ici si notre proté est rouge elle fait le max d’interaction avec ses semblables donc on considère que sa fonction est la même que les rouges).

Analyse de puces à ADN (microarray)

Le processus de base des puces à ADN est lʼhybridation d’ADN marqué différemment. Deux brins dʼADN vontsʼhybrider sʼils sont complémentaires. Lʼobjectif dʼune étude par puce à ADN est dedéterminer les gènes qui sont transcrits à un instant donné de la vie dʼune cellule.Ces puces à ADN génèrent un très grand nombre de données à traiter. Un risque:que cette masse de données ne génère plus de questions que de réponses. A moyen

terme, ces puces à ADN sont sans doute le plus prometteur dans le domaine dudiagnostique et du pronostic. A plus long terme, lʼinterprétation de ces puces à ADNest susceptible de mener à des découvertes intéressantes.

Il faut savoir aussi qu’il existe d’autres techniques comme le RNA séq qui permettent d’obtenir ces informations, ou d’autres info.

Lʼanalyse des données de microarrays comprend les 3 étapes suivantes:

- lʼanalyse de lʼimage, qui a pour objectif dʼextraire une intensité pour chaque spot;

-la normalisation des données: pour la comparaison de profils dʼexpression degènes sous diverses conditions, il faut bien sûr quʼils soient comparables. Desmicroarray relevés dans des conditions identiques doivent conduire à une réponseidentique.

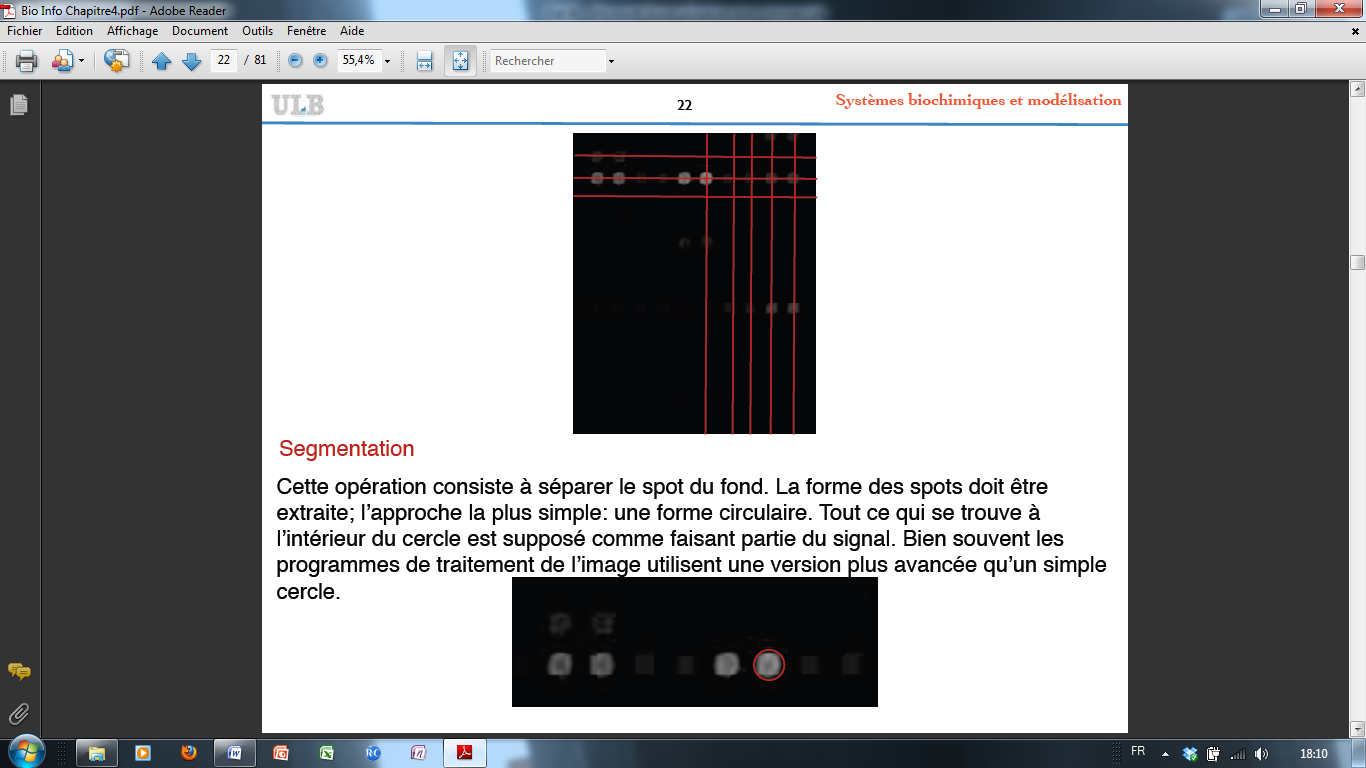
-la classification en groupes de gènes présentant des profils dʼexpression similaires.

1) Analyse de lʼimage

Lʼanalyse de lʼimage a pour but dʼextraire une intensité pour chaque spot, etégalement dʼeffectuer une séparation entre les différentes spots.

Découpage sous forme de grille : Lʼobjectif est ici de soit placer une grille, soit placer les spots dans des cases, lesmieux délimitées possibles.

Segmentation : Cette opération consiste à séparer le spot du fond. La forme des spots doit être

extraite; lʼapproche la plus simple: une forme circulaire. Tout ce qui se trouve àlʼintérieur du cercle est supposé comme faisant partie du signal. Bien souvent lesprogrammes de traitement de lʼimage utilisent une version plus avancée quʼun simplecercle.

Extraction de lʼintensité

Lorsque le spot a été séparé du fond environnant, une intensité doit être extraite. Lesmesures typiques sont lʼintensité moyenne ou médiane de tous les pixels du spot.

Correction du fond

Sur certaines images, un léger signal est vu dans lʼaire située entre les spots. Cʼest lesignal de fond, qui peut être soustrait de lʼintensité du spot, afin dʼavoir une estimationplus fiable du signal biologique au départ du spot. Cette approche peut se discuter, etdivers algorithmes choisiront diverses approches.

2) Normalisation des données

Les microarray sont généralement utilisés pour comparer des profils dʼexpression degènes dans diverses conditions, ou à divers instants de la vie cellulaire. Le fait detravailler par comparaison permet dʼéviter la plupart des biais et des limitations quiaffectent les mesures absolues. Il y a toutefois quelques exceptions. Lʼune dʼentreelles: il faut que ce quʼon compare soit comparable. Et les quantités dʼéchantillonfournies sur une plaque microarray doit être comparable. La normalisation a pour but

de distinguer les variations biologiques et expérimentales des variationssystématiques. Ces dernières proviennent des: - différences dans les rendements de marquage par les fluorophores;Car par exemple les différents fluorophores ne marquent pas de la même façon (temps de demi vie différent p.e.), ce qui pourrait donner une observation de populations différentes alors qu’en réalité elle est la même et que seule le rendement des fluorophores est différent !

-les différences de demi-vie des fluorophores;

- problèmes dus à lʼanalyse dʼimage

Représentation des données: les intensités mesurées sont généralement faibles. Travailler avec le log2 de ces intensités facilite leur exploitation et lʼutilisation dʼoutilsstatistiques.

Attention lors de la normalisation: important de ne pas laminer toutes les variations etde ne pas faire apparaître des faux positifs.

Hypothèse normalisation: majorité des gènes ont un niveau dʼexpression invariantentre 2 conditions (représentées car les canaux rouge et vert), et lorsquʼil y a desvariations positives ou négatives, elles se compensent en moyenne.



MvA plotindique clairement les biais systématiques (M(y) en fonction de A (x))

Variation mesurée:M=log2(puce1)-log2(puce2)

Intensité moyenne:A=[log2(puce1)+log2(puce2)]/2

Avant normalisation

Après normalisation : Les zones danses représentent qu’il n’y a pas de différences entre les niveaux d’expressions des gènes du patient sain et du malade, tandis que les points isolés représentent les sous-gènes dont l’expression est différente des sains.Au dessus : représente l’ensemble des gènes surexprimés, et en dessous : représente l’ensemble des gènes sous exprimés.

3) Méthodes de classification

Lʼobjectif de ces méthodes de classification est de regrouper dans une même classedes gènes qui présentent des propriétés similaires, au niveau de leur expression.

Il existe diverses approches pour réaliser cette classification. Nous aborderons ici laclustering hiérarchique et le k-means. Chacune de ces approches classera lesgènes en fonction dʼune distance entre eux.

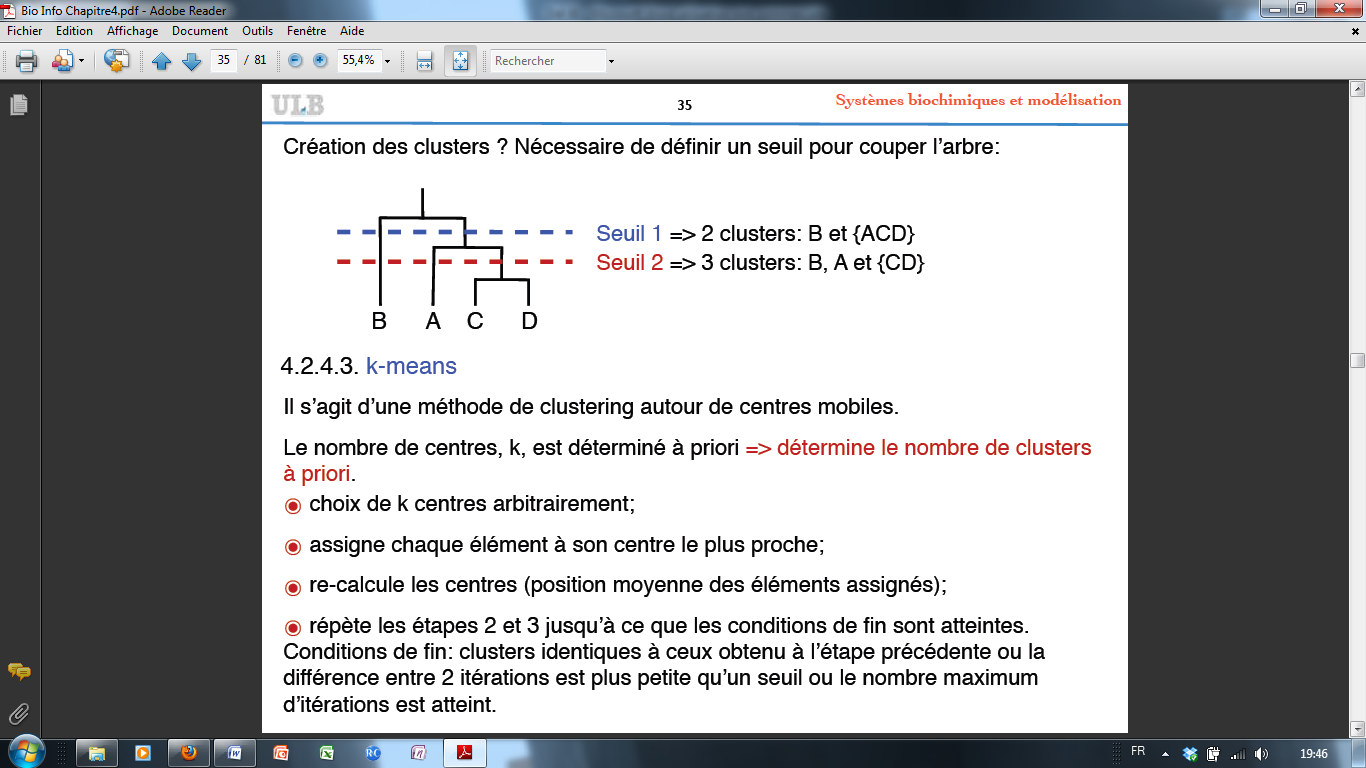
Choix dʼune métrique : selon la méthode utilisée pour évaluer la distance entre 3gènes, on obtient des résultats/classifictaions différentes.

La plus connue est la distance euclidienne qui mesure la distance entre deux profils. Entre deux points a et b dans un espace à N dimensions. Inclut un poids, w, dépendant de la dimension considérée.

La distance Manhattan est la somme pondérée de la différence absolue dans chaque dimension

Notons aussi le coefficient de corrélation de Pearson

Clustering hiérarchique

Lʼobjectif de la méthode est de regrouper de manièrehiérarchique des entités, en groupant en premier celles qui sont les plus proches,telle que défini par la métrique choisie. Cette méthode conduit à des arbres où les noeuds représentent les clusters(=groupes) et les branches les distances entre eux.

Pour créer ensuite les cluster à partir de l’arbre, il est nécessaire de définir un seuil pour couper l’arbre, afin de finalement regrouper les gènes en catégorie. P.e. seuil 1 :2 clusters B et (ACD) ; seuil2 : 3clusters : B, A et (CD)

K-means

Il sʼagit dʼune méthode de clustering autour de centres mobiles.

Le nombre de centres, k, est déterminé à priori => détermine le nombre de clustersà priori.

-choix de k centres arbitrairement;

- assigne chaque élément à son centre le plus proche;

- re-calcule les centres (position moyenne des éléments assignés);

- répète les étapes 2 et 3 jusquʼà ce que les conditions de fin sont atteintes.

Conditions de fin: clusters identiques à ceux obtenu à lʼétape précédente (donc itération qui reste stable, où il n’y plus de mélange entreles catégories) ou ladifférence entre 2 itérations est plus petite quʼun seuil ou le nombre maximumdʼitérations est atteint.

Evaluation des résultats du clustering

- consensus: en utilisant plusieurs méthodes et en comparant les résultats, en vuede dériver un consensus;

- analyse de la robustesse: teste lʼalgorithme plusieurs fois au départ de conditionsinitiales différentes. Que se passes-t-il si on assigne des paramètres initiauxdifférents ? Obtient-on le même clustering si on exclut certains gènes ?

- Utilisation de données biologiques: est-ce que le résultat du clustering est enaccord avec des annotations fonctionnelles ?

Les puces à ADN en tant que classificateurs moléculaires (diagnostique)

Une des applications prometteuses des microarray concerne le domaine médical,au travers de la classification de groupes de patients en fonction du profildʼexpression de gènes.

En effet, certains phénotypes particuliers peuvent être caractérisés par un profildʼexpression, extrait dʼun microarray (diagnostique). Lʼutilisation dʼun microarraypeut aussi se faire dans un objectif de pronostic, en vue de prédire lʼévolutionprobable dʼun patient. Enfin, ces microarray peuvent être utilisés pour prédire queltraitement est susceptible de convenir le mieux à un patient.

Choix des gènes

Par validation croisée, parmis 1000 gènes on en trouve 300 dans un groupe de patients qui permettent de dire appartenance au groupe 1.

Validation

-éviter la surparamétrisation: utiliser moins de paramètres que le nombre

dʼindividus utilisés pour construire le modèle;

-valider la méthode sur un groupe distinct de celui utilisé pour construire le

modèle. Parmis un autre groupe de patients on vérifie que ces 300 gènes permettent bien cette classification.

qqun a compris dans la modélisation des systèmes biochimiques, les différentes approches (discrète, continue, booléenne)??? et quand on utilise laquelle?

D'après ce que j'ai compris: pour la booléenne, tu n'as que 2 valeurs pour ta variable: 0 ou 1 (silence ou activité du gène par exemple) mais pour moi, ce n'est vraiment pas précis pcq si tu travailles en synchrone, tu réactualises chaque fois ton résultat et un seul état découle de ton état précédent donc pas exemple, tu ne peux pas étudier les gènes étudiés dans la différenciation car tu n'auras qu'un état alors que la différenciation peut donner plusieurs états...

Pour la discrète, c'est un peu pareil que booléenne, sauf que tu as plus d'états intermédiaires entre 0 et 1

Pour la continue, tu as tous les intermédiaires entre 0 et 1 donc tu verras plus de changements d'activité que dans les autres approches donc c'est plus précis pour étudier des régulations

Je sais pas si ça t'aide :s

Qqun sait expliquer avec des mots français ce que signifie: "discrétisation de l'espace des conformations" ? (pour prédiction de structure tertiaire des protéines - modélisation ab initio)

ca veut dire que tu prend pas en compte toutes les valeurs mais juste certaines (par exemple que les nombres entiers 1,2,3 mais pas les décimales...)discrétisation= numérisation

Donc ça signifie par exemple, pour l'optimisation par colonie de fourmis où il met "discrétisation des angles dièdres de la chaîne principale" que tu prends que les valeurs d'angles égales à (j'invente) 0°, 30°, 45°, 90° etc et pas tous les intermédiaires?