

PROPOSITION DE SUJETS DE PROJET

• Analyse de séquences d'ADN et « marche ADN »

La fonction de l'ADN est de fournir les indications nécessaires à la fabrication des protéines dont l'organisme a besoin. L'instruction pour cette opération est inscrite dans la molécule d'ADN sous la forme d'une séquence de quatre bases azotées, notées A, T, C et G. Le code génétique est le système de correspondance qui permet à la cellule de fabriquer des protéines à partir de son ADN. Chaque acide aminé étant associé à un groupe de trois bases appelé codons. D'autre part, une grande quantité d'ADN est « non codant » : cet ADN, dont le rôle est encore mal connu, présente une corrélation à longue portée qui traduit son implication dans l'organisation globale de l'ADN au sein du noyau. L'idée c'est de traduire la séquence de lettres en séquence numérique et calculer fonction de corrélation et densité spectrale de puissance, pour mettre en évidence les principales propriétés des séquences codantes et non codantes. Dans ce sujet, on propose d'appliquer cette technique à des séquences réelles, en essayant de déterminer les propriétés statistiques de différentes séquences, et de les interpréter sur la base de leur fonction. Le sujet est ouvert à l'utilisation de techniques alternatives.

• Analyse de séquences d'ADN et « marche ADN »

La fonction de l'ADN est de fournir les indications nécessaires à la fabrication des protéines dont l'organisme a besoin. L'instruction pour cette opération est inscrite dans la molécule d'ADN sous la forme d'une séquence de quatre bases azotées, notées A, T, C et G. Le code génétique est le système de correspondance qui permet à la cellule de fabriquer des protéines à partir de son ADN. Chaque acide aminé étant associé à un groupe de trois bases appelé codons. D'autre part, une grande quantité d'ADN est « non codant » : cet ADN, dont le rôle est encore mal connu, présente une corrélation à longue portée qui traduit son implication dans l'organisation globale de l'ADN au sein du noyau. Au TP 1, nous allons traduire la séquence de lettres en séquence numérique et calculer fonction de corrélation et densité spectrale de puissance, pour mettre en évidence les principales propriétés des séquences codantes et non codantes. Dans ce sujet, on propose de retrouver les mêmes propriétés par une méthode alternative, la « marche ADN ».

La progression attendue du sujet est :

- construire une marche aléatoire à partir de la séquence d'ADN (« marche ADN »), en associant une vitesse +1 ou -1 selon un code choisi
- étudier les propriétés statistiques du phénomène diffusif ainsi obtenu et les comparer aux résultats du TP.
- Comme application de cette technique, on décomposera une (ou plusieurs) séquences codantes en trois sous-séquences correspondantes aux positions $3n$, $3n+1$ et $3n+2$ des codons et analyser leurs propriétés statistiques, pour mettre en évidence une corrélation à longue portée dans la troisième sous-séquence, pour laquelle le choix de la base est moins contraint à cause de la dégénérescence du code génétique.

• Modélisation statistique d'une séquences d'ADN

La fonction de l'ADN est de fournir les indications nécessaires à la fabrication des protéines dont l'organisme a besoin. L'instruction pour cette opération est inscrite dans la molécule d'ADN sous la forme d'une séquence de quatre bases azotées, notées A, T, C et G. Le code génétique est le système de correspondance qui permet à la cellule de fabriquer des protéines à partir de son ADN. Chaque acide aminé étant associé à un groupe de trois bases appelé codons. D'autre part, une grande quantité d'ADN est « non codant » : cet ADN, dont le rôle est encore mal connu, présente une corrélation à longue portée qui traduit son implication dans l'organisation globale de l'ADN au sein du noyau. Les propriétés de corrélation longue portée des séquences, plus ou moins importantes selon qu'il s'agisse de séquences codantes ou non codantes, peuvent être reproduites en combinant un générateur de séquences corrélées – qui génère une suite déterministe – et un bruit aléatoire qui vient partiellement «effacer» la séquence corrélée. Dans ce sujet, on propose donc de construire une séquence artificielle ayant les mêmes propriétés statistiques d'une séquence d'ADN donnée.

La progression attendue du sujet est :

- comprendre le mode de fonctionnement d'une suite déterministe et l'utiliser pour déterminer une séquence de +1 et -1 corrélée à longue portée
- « mélanger » cette séquence avec un bruit aléatoire pour doser le poids de la corrélation
- estimer les propriétés de la séquence ainsi obtenue par les mêmes méthodes utilisées au TP1 (ou d'autres!) et les comparer à celle d'une séquence d'ADN, afin d'ajuster les paramètres du modèle et de mieux interpréter les propriétés de l'ADN.

- **Modèle dynamique de l'ADN et solitons**

La transcription implique l'ouverture locale de la double hélice, afin de lire la séquence de bases qui y est contenue. Cette opération est donc un véritable phénomène dynamique, où des déformations de la structure se produisent et se propagent le long de la molécule. Plusieurs modèles dynamiques de l'ADN ont été proposés pour décrire la structure mécanique de l'ADN et ont mis en évidence la possibilité que des ondes de déformation se propagent le long de la molécule sans déformation : des ondes de ce type sont appelées des «solitons». La dynamique interne de la molécule d'ADN peut être modélisée par des modèles dynamiques qui se placent au niveau des paires de bases. Chaque base est décrite comme une masse, qui interagit avec ses voisines sur le même filament et avec la base à laquelle elle est appariée. Dans les modèles les plus simples on peut alors décrire l'ADN avec un seul degré de liberté par paire de bases. Les équations du mouvement s'écrivent à partir du potentiel d'interaction $W_n = K (y_n - y_{n-1}) + V(y_n)$, où $V(y_n)$ peut être par exemple un potentiel de Morse ou une fonction sinus. Ces deux modèles sont en particulier caractérisés par la présence de solutions (analytiques) de type soliton, solutions localisées et stables qui se propagent à vitesse donnée. *La progression attendue du sujet est :*

- écrire les équation du mouvement du modèle
- étudier la forme analytique des solutions de type solitons que ces équations possèdent pour en comprendre l'allure
- intégrer numériquement les équation du mouvement avec les conditions initiales qui correspondent à la solution soliton cherchée et visualiser le comportement du système dans le temps pour vérifier la stabilité de la solution obtenue
- s'interroger sur la pertinence biologique d'une telle solution
- éventuellement, générer plusieurs solitons à la fois et observer leur interaction lors d'une collision.

- **Modèle thermodynamique de l'ADN**

Une augmentation de la température provoque en effet la rupture des liaisons hydrogène qui relient les deux brins d'ADN et donc leur séparation. Ce phénomène est un véritable changement de phases : il apparaît à une température bien déterminée et présente des analogie importantes avec la vaporisation d'un liquide. Il peut donc être décrit en termes physiques grâce à une modélisation appropriée de la molécule. Le modèle de Peyrard-Bishop est particulièrement adapté à ce problème, puisque il considère comme degré de liberté fondamentale la distance entre deux bases dans une paire. Chaque base est décrite comme une masse, qui interagit avec ses voisines sur le même filament et avec la base à laquelle elle est appariée. Les équations du mouvement s'écrivent alors à partir du potentiel d'interaction $W_n = K (y_n - y_{n-1}) + V(y_n)$, où $V(y_n)$ est un potentiel de Morse. On peut alors simuler le comportement dynamique de l'ADN à des différentes températures :

soit par des simulations dans l'ensemble microcanonique : énergie fixée, donnée par le choix des conditions initiales (vitesses) ;

soit dans l'ensemble canonique : en utilisant l'équation de Langevin ou bien en ajoutant quelques degré de liberté externes qui agissent comme un thermostat (de Nosé Hoover).

Le but du projet est de reproduire la dénaturation, c'est-à-dire la séparation des deux brins lorsque la température dépasse une valeur seuil, dans le cadre du modèle de Peyrard Bishop avec une (ou plusieurs) de ces méthodes.

La progression attendue du sujet est :

- écrire les équation du mouvement du modèle
- étudier comment ces équations doivent être modifiées si on souhaite faire des simulations à température constante
- intégrer numériquement les équation du mouvement avec les conditions initiales appropriées à décrire le système pour des différents valeurs de la température
- reconstruire la courbe de dénaturation et l'interpréter
- étudier comment l'introduction d'un couplage anharmonique entre les paires des bases modifie le comportement du système
- on pourra aussi s'intéresser au rôle des conditions initiales et à leur influence sur la métastabilité des états surchauffés et sous refroidis dans les simulations canoniques.

- **Etirer l'ADN**

Les expériences de manipulation mécanique de molécules uniques d'ADN par pinces magnétiques ou optiques permettent «d'attraper» les deux extrémités d'une molécule d'ADN et d'exercer une force de traction à une des extrémités. On peut ainsi mesurer l'élasticité de la molécule d'ADN, qui ne dépend pas tant de sa structure moléculaire propre, mais plutôt du fait que la chaîne d'ADN est soumise à l'agitation thermique et tend donc à se mettre sous forme d'une pelote désordonnée plutôt qu'à rester allongée. Cette propriété des polymères en solution se traduit par une forme d'élasticité (une force qui s'oppose à l'allongement) qualifiée d'«entropique».

Le but du TP est de modéliser le comportement de l'ADN sous traction. L'ADN est modélisé comme un polymère simple, ou chaîne librement jointe (FJC), c'est-à-dire comme une suite de bâtonnets dont l'orientation est aléatoire. Un coût énergétique est associé à la force extérieure qui tend à aligner les bâtonnets. Pour déterminer la valeur moyenne de la longueur bout à bout pour différentes forces, il faut alors explorer toutes les configurations possibles pour chaque force.

La progression attendue du sujet est :

- comprendre quels paramètres sont nécessaires pour déterminer une configuration de la chaîne librement jointe, selon qu'on veuille modéliser le système en 2D ou 3D
- mettre en place le programme de simulation Monte Carlo permettant de déterminer la distribution des longueurs bout à bout de la chaîne et donc la longueur moyenne, à force donnée
- tracer l'allongement de l'ADN en fonction de la force et comparer aux données expérimentales et aux formules analytiques
- étudier l'influence des différents paramètres sur la réponse observée
- éventuellement, on pourra comparer les résultats obtenus à 2D et à 3D

• Flashing ratchet et moteurs moléculaires

L'avancée d'un moteur moléculaire peut être modélisée comme un mouvement brownien rectifié. Le moteur possède deux états, un faiblement lié au filament sur lequel il se déplace, l'autre fortement lié. Dans le premier état son interaction avec le filament est pratiquement uniforme (énergie constante), alors qu'il est soumis à un potentiel périodique et asymétrique dans l'état fortement lié. La source d'énergie externe permet la transition à l'état fortement lié avec un taux d'excitation donné, suivi par une désexcitation vers l'état plat. Le moteur est aussi soumis à l'agitation thermique et donc il diffuse, librement dans l'état faiblement lié : il se déplace donc avec probabilité égale vers la droite et vers la gauche, mais dès que le potentiel asymétrique est allumé à nouveau, les probabilités de piégeage dans les puits $n+1$ et $n-1$ sont différentes, ceci résultant en un mouvement global dans une direction. La simulation de ce phénomène peut être obtenue simplement et comparée aux formules théoriques.

La progression attendue du sujet est :

- comprendre le modèle et définir les paramètres d'intérêt, choisir une manière de décrire les fluctuations d'excitation/désexcitation (de manière régulière ou aléatoire, avec quels taux)
- implémenter le modèle en utilisant un algorithme de Métropolis simple pour décrire la diffusion de la protéine (voir aussi TP2)
- simuler la progression d'un moteur moléculaire et étudier l'influence des différents paramètres du modèle sur son efficacité

• Diffusion d'une protéine le long de l'ADN : modèle à deux états

Un des mécanismes clés de la régulation de l'expression génétique est la cinétique d'association d'une protéine avec sa séquence cible sur l'ADN. Des mécanismes de « diffusion facilitée » sont observés, où la diffusion unidimensionnelle de la protéine le long de la double hélice s'alterne à sa diffusion libre 3D. Si une interaction de type électrostatique est sans doute responsable de l'attraction ADN-protéine, il faut remarquer aussi que la reconnaissance de la séquence cible implique un système de lecture de la séquence, typiquement associé à des liaisons hydrogènes entre protéine et ADN. La diffusion unidimensionnelle en est alors influencée par le paysage énergétique associé à ces interactions, et on observe un comportement « subdiffusif ». La prise en compte de l'interaction électrostatique semble toutefois modifier ce scénario car elle prévoit une répulsion à courte portée qui effacerait la rugosité du paysage énergétique presque partout, à l'exception des positions qui correspondent aux plus grands nombres de liaisons hydrogène. La question est alors d'explorer les propriétés diffusives de la protéine dans le cas du modèle effectif « à deux états » qui en résulte.

La progression attendue du sujet est :

- construire un modèle simple de diffusion 1D sur un paysage énergétique aléatoire, mais en remplaçant tous les niveaux d'énergie supérieurs à une valeur seuil avec une valeur constante
- étudier l'effet d'une barrière énergétique à passer pour entrer ou sortir des régions « favorables » vers les régions « défavorables »
- comparer les propriétés de diffusion du modèle à celles du modèle initial en faisant varier les paramètres du modèle (seuil d'énergie, énergie par liaison, valeur de l'énergie constante, barrière d'énergie).

• Diffusion d'une protéine sur un profil d'énergie 2D

Une approche alternative au problème discuté dans le sujet précédent, concernant les effets de ralentissement dus au profil d'énergie désordonné vu par la protéine qui se déplace le long de l'ADN, est de prendre en compte un degré de liberté additionnel : la distance

ADN-protéine. Si en plus on considère la contribution électrostatique à l'interaction, on se rend compte que le profil d'énergie vu par la protéine n'est pas une fonction seulement de la position z le long de l'ADN mais aussi de la distance L . On peut donc essayer de modéliser la diffusion de la protéine sur un profil d'énergie bidimensionnel, fonction de ces deux paramètres.

La progression attendue du sujet est :

- construire un modèle simple de diffusion 2D sur un paysage énergétique donné, basé sur l'interaction électrostatique, l'effet des ions et les interactions par liaison hydrogène
- étudier la dynamique de la protéine sur ce profil et en déterminer les caractéristiques principales : dans quelle région la protéine passe le plus de son temps, avec quelle probabilité elle s'éloigne de l'ADN, etc.
- comparer les propriétés de diffusion du modèle à celles du modèle initial et aux propriétés connues en faisant varier les paramètres du modèle.

• Diffusion d'une protéine en 3D et 1D le long de l'ADN

Un des mécanismes clés de la régulation de l'expression génétique est la cinétique d'association d'une protéine avec sa séquence cible sur l'ADN. Des mécanismes de «diffusion facilitée» sont observés, où la diffusion unidimensionnelle de la protéine le long de la double hélice s'alterne à sa diffusion libre 3D. Sous certaines hypothèses, on peut montrer en effet qu'il existe un choix optimal des temps de diffusion 1D et 3D qui rend minimal le temps de recherche. Le but de ce projet est de simuler la recherche d'une séquence cible sur l'ADN de la part d'une protéine en combinant des phases de lecture, où la protéine diffuse le long de l'ADN, et de phases de déplacement rapide, où elle diffuse 3D pour retomber sur l'ADN dans une nouvelle position. La diffusion 1D est « normale » et la durée t_{1D} des excursions 1D est distribuée selon une distribution exponentielle. Les sauts 3D sont des diffusions rapides dans l'espace 3D de durée t_{3D} , et on assume que la durée moyenne $\langle t_{3D} \rangle$ de ces excursions 3D soit fixée. Après un saut 3D, la position où la protéine arrive sur l'ADN est supposée complètement aléatoire. On peut alors montrer que le temps total de recherche de la cible T est minimisé (optimisé) si le temps moyen d'une recherche 1D est égale au temps moyen d'un saut 3D.

La progression attendue du sujet est :

- comprendre le modèle et décider si oui ou non on veut modéliser explicitement les sauts 3D ou utiliser directement leur propriétés dans le calcul du temps de recherche T
- implémenter le modèle en utilisant un algorithme de Métropolis simple pour décrire la diffusion de la protéine
- étudier l'allure du temps total de recherche T pour des différents choix de la distribution des durées moyennes t_{1D} de diffusion 1D et le comparer au résultat analytique simple décrit plus haut
- des formules plus complexes de la condition de minimisation de T peuvent être obtenues et comparées aussi au résultats de la simulation
- on pourra aussi comparer le comportement du modèle dans le cas où les excursions 3D sont modélisées explicitement ou pas.

• Interaction électrostatique ADN-protéine

Les protéines en solution sont presque toujours chargées positivement, et sont donc sensibles au champ électrostatique associé à l'ADN qui, lui, est très fortement chargé négativement. L'étude de cette interaction implique la résolution d'un problème d'électrostatique dans une solution, et donc en présence d'eau et d'ions. C'est un type de problème rencontré souvent dans le domaine de la physique des colloïdes (particules chargés en solution), bien que la taille des objets qui interagissent ici soit de quelques ordres de grandeurs inférieure, et qu'il existe deux types de charges en interaction. Pour déterminer cette interaction, dans une approche minimale, on modélise la protéine et l'ADN par deux plans parallèles. Il faut alors résoudre l'équation non linéaire de Poisson Boltzmann entre deux plans pour le potentiel électrostatique Φ (équation du type $\Phi'' = A \sinh(B\Phi)$) avec conditions au bord (définissant les charges des deux plans qui correspondent aux dérivées de Φ). On en déduit la pression par une formule, puis l'énergie libre du système en intégrant. La résolution de cette équation montre un effet inattendu : l'existence d'une répulsion à courte portée, d'origine osmotique, entre les deux corps de charge opposée.

La progression attendue du sujet est :

- Comprendre comment résoudre une équation différentielle avec conditions au bord sous scilab
- écrire l'équation de Poisson Boltzmann avec les bonnes constantes puis l'intégrer numériquement pour des valeurs réalistes des paramètres physiques obtenir le potentiel électrostatique, puis en déduire la force d'interaction
- étudier comment on peut écrire la pression qui s'exerce sur les deux surface et la calculer numériquement
- faire varier la concentration en sel, la charge de la protéine, et faire une étude du comportement de la pression
- éventuellement, on pourra intégrer au modèle une correction qui prend en compte de manière approchée l'effet de la courbure des deux surfaces qui se font face dans le système réel (approximation de Derjaguin) pour comparer le cas de protéines de

forme différente

- l'effet combiné de l'interaction électrostatique avec une interaction directe par liaisons hydrogène pourra aussi être abordé

- **Chromatine et pinces magnétiques**

L'organisation de l'ADN dans le noyau cellulaire est très particulière, car l'ADN y est associé à un grand nombre de protéines, appelées histones, qui se chargent d'organiser sa longue molécule dans une structure repliée et complexe appelé chromatine. Les protéines histones se regroupent par huit et forment un petit assemblage globulaire autour duquel l'ADN s'enroule en un tour trois quart (en hélice gauche) : on appelle cette bobine d'ADN un nucléosome. Les nucléosomes se succèdent le long de l'ADN en formant une structure en « collier de perles » qui se replie ensuite sur elle-même pour former un espèce de ressort complexe et dense, de 30 nm de diamètre, nommé fibre de chromatine, où les nucléosomes sont empilés les uns sur les autres. Le but de ce projet est de comprendre et modéliser le comportement de la fibre de chromatine sous traction et torsion observé dans les expériences de molécule unique par pinces magnétiques. Dans le TP3 on s'intéressera à la cinétique du passage du nucléosome « standard » à son état « renversé », le reversome, observé pour un couple appliqué très élevé. L'état « standard » du nucléosome est en réalité composé de trois états différents, caractérisés par un arrangement différent des ADN entrant et sortant du nucléosome, et qui coexistent à couple relativement faible : la probabilité de trouver le nucléosome dans un de ces trois états est donnée par la statistique de Boltzmann. On peut donc décrire, par un modèle statistique, l'évolution des populations des trois états, plus l'état reversome, en fonction du couple appliqué.

La progression attendue du sujet est :

- reprendre le TP3 et le modifier pour y ajouter l'existence des trois états du nucléosome « standard » et l'équilibre statistique entre ces états
- améliorer la description de la longueur z de la fibre pour obtenir un meilleur fit des courbes torsion-extension
- étudier la réponse de la fibre dans la région de faible couple, en la comparant aux données
- étudier l'influence des différents paramètres du modèle sur le fit des données

+ [Projets sur le haut débit]