Statistique pour la bio-informatique

Séance 1 - 1 Octobre 2003

Présentation et Rappels

1 Présentation du module

Ce cours réparti en 14 séances d'une heure trente a pour sujet la présentation de modèles statistiques utilisés dans le contexte de l'analyse de données en biologie moléculaire et en génétique. Elles s'inspirent notamment des excellents ouvrages de Ewens et Grant "Statistical Methods in Bioinformatics", springer NY, 2001, et Durrett "Probabilistic Models of DNA sequences", Springer NY, 2003, ainsi que de nombreuses notes disponibles sur internet (S. Tavaré, T. Speed ou R. Karp, pour ne citer que quelques uns). Elles s'adressent en particulier aux élèves de l'option bioinformatique de l'ENSIMAG, mais à toute personne, étudiant de troisième cycle ou chercheur, désirant de manière élémentaire se former en probabilités et statistique pour la biologie. Le format encore préliminaire de ces notes n'engage que leur auteur.

L'analyse de données biologiques et en particulier de données génétiques (génomes, protéomes, transcriptomes, etc) constitue une nouvelle branche de la science : la bio-informatique, une approche *in silico* venant compléter les approches traditionnelles en biologie. La bioinformatique recourt à la *méthode comparative*. Des logiciels de comparaison explorent les banques de données de gènes et de protéines déjà identifiés et annotés par des équipes scientifiques réparties dans le monde entier afin d'établir des rapprochements avec des séquences ou structures inconnues que l'on étudie et que l'on cherche à identifier. Les approches *statistiques* et la *modélisation probabiliste* sont extrêmement importantes. Des logiciels analysent les données pour tenter de dégager des règles présentant un caractère systématique ou général. L'objectif de la *démarche probabiliste* est la construction de modèles qui tentent d'extraire les propriétés communes des séquences ou des structures.

La première partie de ces notes discute des modèles d'évolution moléculaire. Cette discussion constitue un préalable indispensable aux calculs de scores de comparaison, à l'évaluation des distances évolutionnistes séparant des espèces, à la reconstruction d'arbres phylogénétiques sensés décrire l'évolution de ces mêmes espèces. La seconde partie décrit quelques principes sous-jacents à la théorie du logiciel BLAST, l'un des plus utilisés pour la recherche de séquences similaires dans les banques de données. Les idées mathématiques décrites sont au coeur du fonctionnement du logiciel. Elles illustrent sans doute de l'une des plus belle manière l'apport des probabilités dans ce domaine. La troisième partie est consacrée à certains aspects de la reconstruction d'arbres phylogénétiques, méthodes de distances et maximum de vraisemblance. Elle est complétée par une présentation de la théorie de la coalescence, une vision moderne de la génétique des populations. La dernière partie traîte des chaînes de Markov cachées, un puissant outil de modélisation dans le domaine de l'analyse des séquences. Nous présenterons les principaux algorithmes d'estimation dans ce domaine.

Les prérequis de ce cours concernent le calcul des probabilités, quelques notions d'estimation, de théorie des tests statistiques et des connaissances élémentaires concernant les processus de Markov. Une partie de ce matériel est couvert par le cours de probabilités de première année, une autre partie par les cours de processus aléatoires, et de statistique de deuxième année.

Le volume de 21 heures se décompose en deux bimestres de cours de 10H30.

- Bimestre 1 : 7 séances de 1H30
  - séance 1 : Introduction, rappels de bio, prérequis en proba/stat
  - séance 2 : Modèles pour l'évolution moléculaire, Wright-Fisher, fixation
  - séance 3 : Modèles de Jukes-Cantor et Kimura
  - séance 4 : Calcul de scores, matrice PAM
  - séance 5 : Loi des extrêmes, Blast
  - séance 6 : Maximum de vraisemblance, algorithme EM
  - séance 7 : Phylogénie moléculaire 1, distances
- Bimestre 2 : 7 séances de 1H30
  - séance 8 : Phylogénie moléculaire 2, maximum de vraisemblance
  - séance 9 : HMM1 : Exemples, modèles
  - séance 10 : HMM2 : Estimation dans les HMM
  - séance 11 : HMM3 : Estimation dans les HMM
  - séance 12 : Prédiction de structure 1 ou coalescence 1
  - séance 13 : Prédiction de structure 2 ou coalescence 2
  - séance 14 : recalage ou révision

Les thèmes présentés dans ces documents constituent une sélection de quelques aspects concernant l'application des probabilités en génétique et en bioinformatique correspondant au volume limité de cours et de travaux dirigés. D'autres thèmes importants ne seront présentés que de manière allusive par faute de temps. Ces thèmes pourront faire l'objet de projets personnels séparés. Voici une liste non-exhaustive de sujets d'approfondissement liés à ce cours.

- Application des chaînes de Markov cachées en bio-informatique.
- Méthodes de Monte Carlo par chaînes de Markov : principes, algorithmes, appli-

cations à l'estimation de taux de mutation, aux alignements multiples (d'aprés Tavaré, d'aprés Liu)
- Approximations poissoniennes et méthode de Chen-Stein. Application au calcul de p-values (d'aprés Waterman).
- Réarrangement du génome (d'aprés Durrett)
- Blast dans le détail (d'aprés Karlin-Dembo)
- Génétique et statistique : interactions historiques, Galton, Fisher et les autres
- Méthodes du maximum de vraisemblance dans le calcul des phylogénies (D'aprés Hulsenbeck et Bollack).
- Prédiction de la structure 3D, réseaux neuronaux, algorithmes génétiques

2 Rappels de Biologie

Avant de se lancer dans la discussion de modèles pour l'évolution de molécules d'ADN, on doit garder à l'esprit un certain nombre de propriétés biologiques des objets en question[1]. Le message important est que *les séquences d'ADN ne sont pas de longs mots pris au hasard dans un alphabet de quatre lettres.* La chimie et la biologie (l'évolution) jouent un rôle majeur dans l'organisation de ces séquences.

L'information héréditaire contenue dans chaque organisme vivant et transmise de génération en génération est véhiculée par les molécules d'ADN (acide désoxyribonucléique), constituées de 2 chaînes complémentaires enroulées l'une dans l'autre en une double hélice. Chaque chaîne peut être vue comme une suite de nucléotides appelés *adénine* (a), *cytosine* (c), *guanine* (g) et *thymine* (t). Dans la chaîne complémentaire, a est toujours apparié avec t et c avec g. les bases sont dites *complémentaires*. La liaison $a = t$ est constituée de deux liens d'hydrogène. Elle est plus fragile que la liaison $c \equiv g$ qui en comporte trois. Par ailleurs, on dit que a et g sont des *purines* tandis que c et t sont des *pyrimidines*.

La molécule d'ADN est orientée. Son squelette est fabriquée de sucres (désoxyribose) liés par des phosphates. Sur un brin, il y a un radical phosphate à une extrémité (l'extrémité 5') et un hydroxyle (OH) à l'autre extrémité (extrémité 3'). La situation est inversée sur le brin complémentaire et la convention veut que l'on écrive les séquences en partant de 5' vers 3'. Cela correspond à la direction de transcription de l'ADN en ARN.

Par complémentarité, il y autant de a que de t dans la molécule d'ADN, et autant

[1] Un lecteur ayant déjà des connaissances élémentaires en biologie peut sauter cette introduction dont l'objectif est de préciser quelques définitions

de c que de g. En revanche, les fréquences observées sur chaque brin varient d'une espèce à l'autre. Chez le colibacille (*Escherichia coli*), on peut considérer que les 4 fréquences sont égales (à 0.25). Dans le génome de la levure (un brin pris dans chaque chromosome), il y a 12 millions de nucléotides. On estime que les fréquences respectives des quatre nucléotides sont a : 0.309, t : 0.308, c : 0.192, g : 0.191.

Le génome humain comporte environ trois milliards de nucléotides, dont plus de 95% ne sont apparemment pas directement liés à la synthèse de protéines. Le rôle de cette grande fraction, appelée *ADN non codant* est en fait mal compris de nos jours. Dans les 5% restant, on trouve entre 40000 et 60000 gènes codant pour des protéines.

Les gènes sont transcrits en acide ribonucléique (ARN), aussi appelé ARN messager, qui pour les organismes eukaryotes sort du noyau de la cellule afin d'être lu par les ribosomes et traduits en protéines dans le cytoplasme. Dans l'ARN, on ne trouve qu'un brin. Le nucléotide t est remplacé par le nucléotide *u uracyle*.

Depuis la bactérie jusqu'à des organismes plus complexes, toutes les protéines sont bâties à partir de 20 acides aminés. Les 20 acides aminés sont constitués de triplets de nucléotides adjacents dans l'ARN, appelés codons. Sur les 64 codons possibles, 61 codent pour l'un des 20 acides aminés. Le code est dit *redondant*. La table de correspondance *codons-acides aminés* s'appelle le code génétique. Les acides aminés se distinguent par des lettres majuscules de d'alphabet (A :Ala :*Alanine*, R :Arg :*Arginine*, N,D, etc). Trois codons signifient `stop` et indiquent la fin de la transcription (*uaa*, *uga*, *uag*). De même, dans l'ADN cette fois, un triplet donne le signal de début de transcription : il s'agit de *atg* (qui n'est pas un acide aminé).

La redondance du code génétique (61 triplets pour 20 acides aminés) joue un rôle important dans la structure de l'ADN, en particulier vis à vis de l'évolution de cette molécule. En effet toutes les mutations n'ont pas le même effet sur le sens d'un gène. Certaines mutations d'un nucléotide ne changent pas la nature de l'acide aminé codé alors que le codon lui-même peut changer. De telles mutations sont appelées *substitutions synonymes* tandis que les autres sont qualifiées de *non-synonymes*. Par exemple,une mutation entre a et g peut préserver la signification `stop` d'un codon ($uaa \to uga$).

Dans l'ADN, les substitutions qui préservent le type (purine a, g ou pyrimidine c, t) sont appelés des *transitions*. Les autres sont des *transversions*. Dans certaines parties de l'ADN, les transitions peuvent être de 10 à 20 fois plus fréquentes que les transversions.

Un gène peut être compris comme une suite de nucléotides spécifiant la séquence d'acides aminés d'une protéine. Toutefois, les règles gouvernant les gènes sont abominablement plus complexes que le code génétique. Pour les bactéries, le génome est compact et les gènes se succèdent dans l'ADN. Pour les organismes plus évolués, les gènes sont constitués de sections transcrites ou codantes (*exons*) séparées par des sections non-codantes (les *introns*). Les gènes sont constitués de quelques milliers de paires de bases (Mega bp) et peuvent comporter des dizaines d'introns. Le rôle des introns permet d'assurer la multi-fonctionnalité d'un gène par le phénomène d'épissage alternatif (choix des exons exprimés). De plus, les gènes peuvent interagir et s'exprimer simultanément par le biais de réseaux d'interactions géniques très complexes et encore généralement méconnus. Les régions adjacentes ou flanquantes non-transcrites sont aussi importantes, car elles contiennent les messages de transcription des gènes.

Certaines espèces sont haploïdes c'est à dire qu'elles ne contiennent qu'une seule copie des gènes (bactéries par exemple). La plupart sont diploïdes (2 copies) à l'exception des gènes portés par les chromosomes sexuels. L'évolution (et la construction) des gènes est elle-même très complexe et ne se limite pas aux phénomènes de substitution. Certaines théories impliquent l'interaction avec les espèces parasites (transferts horizontaux), l'existence de l'ADN pour lui-même (ADN égoiste), des sauts à l'intérieur de la molécule etc.

3 Banques de données

La compilation, l'organisation des données et la création de bases de données est l'une des vocations premières de l'utilisation de moyens de calcul en Biologie. Les moyens informatiques sont naturellement utilisés pour le stockage ou la gestion des données de séquences d'ADN ou de protéines mais également pour l'interprétation de ces données.

Les grandes banques de séquences généralistes telles que Genbank ou l'EMBL sont des projets internationaux et constituent des leaders dans le domaine. Elles sont maintenant devenues indispensables à la communauté scientifique car elles regroupent des données et des résultats essentiels dont certains ne sont plus reproduits dans la littérature scientifique. Leur principale mission est de rendre publiques les séquences qui ont été déterminées. L'un des intérêts de ces banques est la masse de séquences qu'elles contiennent. On y trouve également une bibliographie et une expertise biologique directement liées aux séquences traitées. Pour que l'utilisateur puisse s'y repérer, toutes ces informations sont mises à la disposition de la collectivité scientifique selon une organisation en rubriques ou en champs.

En parallèle, pour les protéines, deux banques principales ont été créées. La première se nomme la Protein Identification Ressource (PIR-NBRF). La deuxième, Swissprot a été constituée à l'Université de Genève à partir de 1986 et regroupe entre autres des séquences annotées de la PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL.