Données de grande dimension en biologie

Président: Yutaka Yasui (University of Alberta)

KAREN A KOPCIUK, Alberta Health Services - Cancer Care

Méthodes d'estimation de la taille d'échantillon pour données métabolomiques de la résonance magnétique nucléaire (RMN)

Les méthodes typiques basées sur la projection pour l'analyse de données métabolomiques, telle la régression des moindres carrés partiels, n'ont pas de méthodes d'inférence statistique ou de méthodes de la taille d'échantillon correspondantes. Pour surmonter cette limite dans la planification de l'étude, nous avons étudié des approches d'estimation de la taille d'échantillon pour données de micropuces à haute dimension afin de tester leur pertinence avec des données métabolomiques RMN. Des études par simulation ont comparé trois méthodes selon différentes caractéristiques pertinentes. Les résultats de l'étude par simulation et une application aux données sur le cancer du pancréas ont suggéré deux méthodes à utiliser pour l'estimation de la taille d'échantillon pour données métabolomiques RMN. A l'avenir, nous étudierons ces méthodes avec des données métabolomiques par spectrométrie de masse.

TOBY KENNEY, Dalhousie University

Modèle généralisé à base de codons pour la substitution de nucléotide dans des séquences d'ADN codant pour des protéines

Nous proposons pour l'analyse phylogénétique un modèle généralisé à base de codons pour les séquences d'ADN codant pour des protéines. Ce modèle fournit un cadre unifié pour les modèles de codons (ou d'ADN ou d'acides aminés) existants. Il offre également une grande flexibilité pour le choix de la matrice de transition, permettant de facilement étendre les modèles existants pour incorporer un plus grand nombre de forces motrices de l'évolution moléculaire, les informations sur la structure et les propriétés des acides aminés. Nous offrons un progiciel appelé Codon Optimal Likelihood Discoverer (COLD) pour mettre en oeuvre les modèles généralisés à base de codons proposés. Nous démontrons avec la théorie habituelle de la vraisemblance le fonctionnement du cadre de notre modèle permettant la sélection de modèle.

HONG GU, Dalhousie University

Estimation de la sélection positive de Darwin à l'aide de modèles généralisés à base de codons

Le modèle généralisé à base de codons pour les séquences d'ADN codant pour des protéines offre une grande flexibilité pour le choix de la matrice de transition et améliore significativement la qualité de l'ajustement. Nous développons davantage l'approche du modèle mixte pour ce type de modèle en permettant les effets aléatoires pour un ou plusieurs paramètres sur des sites d'acides aminés. En permettant les effets aléatoires pour le paramètre estimant le taux de mutation non synonyme sur différents sites, les modèles généralisés à base de codons incluent la série de modèles M de Yang et al. (2000) à titre de cas particuliers. Nous démontrons l'estimation plus précise de la pression sélective hétérogène aux sites d'acides aminés, comparativement à l'estimation par la série de modèles M.

THIERRY CHEKOUO TEKOUGANG, Université de Montreal

Approche statistique des algorithmes populaires de grappes bi-dimensionnelles

Récemment, plusieurs algorithmes de recherche de grappes bi-dimensionnelles ont été proposés. Ces algorithmes ont pour but de déterminer une sous-matrice de la matrice de données dont les lignes exhibent un comportement similaire à travers les colonnes, et vice versa. Peu de ces algorithmes sont basés sur des modèles statistiques explicites. Ce travail propose des modèles statistiques sous-jacents aux algorithmes les plus populaires. Il montre que ces algorithmes peuvent être justifiés dans un cadre bayésien et peuvent être dérivés à travers des techniques computationnelles bayésiennes.

DENA GIVARI, University of Guelph

De l'application d'un modèle de mélange bayésien infini

En analyse de micropuces, les gènes sont exprimés à différents niveaux selon les conditions cellulaires. L'identification de gènes est supposée renseigner sur leurs fonctions biologiques lorsque leurs formes d'expression sont similaires dans des conditions démontrant des régimes variables de temps, de croissance, de topographie, d'histologie et de physiologie. L'objectif est de créer des grappes rassemblant les gènes dont les niveaux d'expression sont liés entre eux sous différentes conditions. Nous avons développé un algorithme récent basé sur un modèle de mélange bayésien infini, et nous permettons en particulier des matrices de covariance non isotropes afin de mieux refléter la réalité. Notre algorithme est présenté et illustré au moyen de données réelles et simulées.

VAHID PARTOVI NIA, McGill University
Classification double agglomérative bayésienne

Dans plusieurs domaines biologiques tels que la métabolomique, la protéomique et la génétique, une matrice de données dont les sujets correspondent aux lignes et les variables aux colonnes, est produite. Pour ce type de données, la classification simultanée des sujets et des variables, appelée classification double, est intéressante. Par exemple, avec des données d'expression génétique, la classification des sujets groupe les sujets ayant des similarités génétiques, tandis que la classification des gènes groupe les gènes ayant des fonctions similaires. Nous suggérons un algorithme de classification double complètement automatisé en utilisant un modèle bayésien dont il n'est pas nécessaire de connaître le nombre de partitions. De plus, la méthode agglomérative produit une représentation graphique des classes doubles à l'aide d'un dendrogramme.