1

Modèles statistiques pour la génomique

Session organisée par Catherine Matias

La biologie moléculaire a connu un certain nombre de révolutions technologiques au cours des années passées qui ont conduit à une production massive de données dont il faut à présent tenter d'extraire de l'information. Ces données sont de types variés (séquences sur des alphabets finis, vecteurs de très grande dimension, images en dimension 2...) et les outils probabilistes et statistiques utilisés le sont encore davantage.

Cette session a pour but de montrer la diversité des problèmes statistiques et probabilistes posés par la biologie moléculaire. Pour une introduction générale à l'étude plus spécifique des séquences biologiques, nous renvoyons à (1).

Références :

[1] *Biological sequence analysis : probabilistic models of proteins and nucleic acids* (1998) Durbin, Eddy, Krogh, and Mitchison, Cambridge University Press.

Adresse de l'organisateur :

Catherine Matias
CNRS - Laboratoire Statistique et Génome
Tour Evry 2
523 place des Terrasses de l'Agora
91000 Evry, FRANCE
E-mail : matias@genopole.cnrs.fr
<http://stat.genopole.cnrs.fr/~cmatias>

Un modèle de substitution nucléotidique avec dépendance au voisinage

par **Jean Bérard**, Jean-Baptiste Gouéré et Didier Piau

Les modèles les plus simples de substitution nucléotidique font en général l'hypothèse que les substitutions affectent indépendamment les différents sites d'une séquence d'ADN. Il est néanmoins connu qu'une telle hypothèse n'est pas réaliste d'un point de vue biologique, comme en témoigne, par exemple, la sur-mutabilité des paires CpG. Des modèles incorporant une dépendance au voisinage des taux de substitution ont donc été introduits récemment dans la littérature biologique. Nous montrons qu'une vaste classe de tels modèles est soluble, au sens où les fréquences à l'équilibre des polynucléotides peuvent être obtenues comme solution de systèmes linéaires explicites et de taille finie. L'effet de la dépendance au voisinage sur la distribution d'équilibre peut ainsi être mis en évidence. Des résultats concernant la vitesse de convergence du modèle vers l'équilibre, ainsi que des propriétés d'indépendance spatiale de la distribution d'équilibre, sont également présentés.

Adresses :

Jean Bérard
Institut Camille Jordan
Université Claude Bernard - Lyon 1
50 avenue Tony-Garnier,
69366 Lyon Cedex 07 France
E-mail : `jean.berard@univ-lyon1.fr`
<http://math.univ-lyon1.fr/~jberard/>

Jean-Baptiste Gouéré
Laboratoire de Mathématiques, Applications et Physique Mathématique d'Orléans
Université d'Orléans
B.P. 6759
45067 Orléans Cedex 2, France
E-mail : `Jean-Baptiste.Gouere@univ-orleans.fr`
<http://www.univ-orleans.fr/mapmo/membres/gouere/>

Didier Piau
Institut Camille Jordan
Université Claude Bernard - Lyon 1
50 avenue Tony-Garnier,
69366 Lyon Cedex 07 France
E-mail : `Didier.Piaup@univ-lyon1.fr`
<http://www.lapcs.univ-lyon1.fr/~piau/>

Estimation des paramètres d'un modèle pair-Markov caché (pair-HMM).

par **Ana Arribas-Gil**, Elisabeth Gassiat et Catherine Matias

Les modèles pair-Markov caché ont émergé dans le contexte de l'analyse de sequences biologiques dans les années 90, lorsque l'on a commencé à traiter l'alignement de séquences dans un contexte évolutif. Ils permettent une approche d'estimation qui court-circuite le problème du choix des paramètres d'une fonction de score pour l'alignement. Cependant, leur fondement statistique n'a encore jamais été établi. Nous nous intéressons donc à la convergence des algorithmes d'estimation des paramètres dans un modèle pair-Markov caché, notamment a la proprieté de consistence des estimateurs bayésien et par maximum de vraisemblance. Nous prouvons l'existence des taux d'information (divergences). Sous des hypothèses supplémentaires, nous prouvons que ces taux sont strictement positifs en tout point différent du vrai paramètre (propriété de divergence). Ceci entraine la consistance des estimateurs dans les modèles où la propriété de divergence est vérifée. Des simulations illustrent des cas qui ne sont pas compris dans nos résultats.

Adresses :

Ana Arribas-Gil
Laboratoire de Mathématiques, Équipe Probabilités, Statistique et Modélisation
Université Paris 11- Orsay
Bât 425
91405 Orsay Cedex, France.
E-mail : ana.arribasgil@math.u-psud.fr
<http://www.math.u-psud.fr/>

Elisabeth Gassiat
Laboratoire de Mathématiques, Équipe Probabilités, Statistique et Modélisation
Université Paris 11- Orsay
Bât 425
91405 Orsay Cedex, France.
E-mail : elisabeth.gassiat@math.u-psud.fr
<http://www.math.u-psud.fr/~gassiat>

Catherine Matias
CNRS - Laboratoire Statistique et Genome
Tour Evry 2
523 place des Terrasses de l'Agora
91000 Evry
E-mail : matias@genopole.cnrs.fr
<http://stat.genopole.cnrs.fr/~cmatias>

Estimateurs par vraisemblance locale dans les modèles linéaires généralisés en indice simple et application à la classification supervisée de données issues des biopuces

par **Sophie Lambert-Lacroix** et Julie Peyre

Une des principales caractéristiques des données issues des biopuces réside dans le fait qu'on dispose d'un nombre de données n très faible devant un grand nombre de gènes p. Un des enjeux majeurs du traitement statistique de ces données est l'analyse discriminante à but décisionnel. D'un point de vue statistique, ce grand nombre de covariables devant un petit nombre d'observations rend l'analyse discriminante difficile. Une façon de contourner ce "fléau de la dimension" consiste à réduire cette dimension. Dans cette présentation, la classification supervisée est vue comme un problème de régression avec peu d'observations et beaucoup de covariables. Nous présentons une approche de réduction de dimension. Utilisant une approche semi-paramétrique, nous utilisons des estimateurs par vraisemblance locale dans les modèles linéaires généralisés en indice simple. Nous illustrons cette approche lorsqu'elle est appliquée à l'analyse de trois jeux de données réels : Colon, Leukemia et SRBCT.

Adresses :

Sophie Lambert-Lacroix
Laboratoire de Modélisation et calcul. IMAG
Université Joseph Fourier
51 rue des Mathematiques,
B. P. 53,
38041 Grenoble Cedex 9, France
E-mail : Sophie.Lambert@imag.fr
<http://www-lmc.imag.fr/lmc-sms/Sophie.Lambert/>

Julie Peyre
Laboratoire de Modélisation et calcul. IMAG
Université Joseph Fourier
51 rue des Mathematiques,
B. P. 53,
38041 Grenoble Cedex 9, France
E-mail : Julie.Peyre@imag.fr
<http://www-lmc.imag.fr/lmc-sms/Julie.Peyre/>

Compression pour la classification supervisée

par **Tristan Mary-Huard** et Stéphane Robin

On s'intéresse ici au traitement de la redondance des variables en classification supervisée. L'outil couramment employé pour traiter cette redondance est la sélection de variables : elle permet d'éliminer à la fois les variables non informatives pour la réponse, et les variables redondantes avec celles déjà sélectionnées. Ce traitement de la redondance ne semble pas toujours satisfaisant. Lorsque le nombre de variables est grand comparé au nombre d'observation, il entraîne une instabilité de la sélection, rendant les résultats difficiles à interpréter. Lorsque le niveau de bruit est élevé et que deux variables sont pertinentes et redondantes, n'en sélectionner qu'une peut sensiblement dégrader les performances du classificateur construit. Nous présentons un nouvel algorithme de compression de variables, basé sur la classification hiérarchique ascendante, et visant à regrouper les variables partageant la même information vis-à-vis de la réponse. Le principe de la méthode consiste à choisir une similarité entre variables pertinente pour la méthode de classification choisie. Une première application de cet algorithme à la méthode des k plus proches voisins donne des résultats encourageants, et facilite l'interprétation des résultats.

Adresses :

Tristan MARY-HUARD
UMR ENGREF/INA PG/INRA 518, Mathématiques et Informatique Appliquées
Département O.M.I.P., INA P-G
16, rue Claude Bernard
75231 Paris CEDEX 05, France
E-mail : `maryhuar@inapg.fr`
<`http://www.inapg.inra.fr/ens_rech/maths/index.html`>

Stéphane ROBIN
UMR ENGREF/INA PG/INRA 518, Mathématiques et Informatique Appliquées
Département O.M.I.P., INA P-G
16, rue Claude Bernard
75231 Paris CEDEX 05, France
E-mail : `robin@inapg.fr`
<`http://www.inapg.inra.fr/ens_rech/maths/index.html`>

Prise en compte de la dépendance pour estimer le nombre de composantes non nulles d'un vecteur gaussien

par Yannick Baraud, Sylvie Huet et **Fanny Villers**

Nous considérons un modèle gaussien, dans lequel les erreurs ne sont pas indépendantes. Le but est d'estimer le nombre de composantes non nulles de l'espérance d'un vecteur gaussien dans un tel modèle. Nous proposons un critère de choix de modèle basé sur la vraisemblance pénalisée, qui tienne compte de la dépendance du modèle.

Adresses :

Yannick BARAUD
Laboratoire J.A.Dieudonné
Université de Nice Sophia-Antipolis,
Parc Valrose
06108 Nice Cedex 2, FRANCE
E-mail : `baraud@math.unice.fr`
<`http://math1.unice.fr/~baraud/`>

Sylvie Huet
INRA - Unité Mathématiques et Informatique Appliquées
NRA - Domaine de Vilvert
78352 JOUY-EN-JOSAS Cedex, FRANCE
E-mail : Sylvie.Huet@jouy.inra.fr
<http://www.inra.fr/miaj/public/statenv/Equipe/sh.fr.html>

Fanny Villers
INRA - Unité Mathématiques et Informatique Appliquées
NRA - Domaine de Vilvert
78352 JOUY-EN-JOSAS Cedex, FRANCE
E-mail : Fanny.Villers@jouy.inra.fr
<http://www.inra.fr/miaj/>