ENV-STAT-EXP

(Projet de Recherche 2004)

ACI IMPBIO

Développement d'un environnement dédié à l'analyse statistique des données d'expression

http://www.lsp.ups-tlse.fr/Biopuces/

CONTEXTE

Le Laboratoire de Statistique et Probabilités (U.P.S., INSA, C.N.R.S. – U.M.R. 5583), la Plate-forme Transcriptome-Biopuces du Génopôle de Toulouse Midi-Pyrénées (localisée à l'INSA de Toulouse) et le Département de Génétique Animale (INRA de Toulouse) collaborent, depuis environ trois ans, dans le traitement statistique des données d'expression produites par les biopuces.

Le présent projet correspond à leur souhait commun de développer cette collaboration et de progresser à la fois dans l'expertise du traitement statistique de ce type de données et dans la mise à la disposition des biologistes d'outils appropriés et conviviaux.

OBJECTIFS

Le projet comprend trois volets :

1. Réaliser, en langage statistique R, des outils logiciels réalisant les traitements statistiques des données d'expression et, parallèlement, créer un site *web* contenant ces outils ainsi qu'une documentation relative aux données traitées, aux programmes ayant permis leur traitement et aux méthodes statistiques utilisées, le tout en accès libre.
2. Approfondir le travail de recherche en statistique, pour adapter les méthodes existantes et développer de nouvelles méthodes.
3. Plus en amont, mener une réflexion conjointe entre statisticiens et biologistes sur la meilleure façon de produire les données d'expression, afin, d'une part, de minimiser la variabilité des mesures due à l'expérimentation, d'autre part, de disposer des outils statistiques les plus appropriés pour répondre aux différentes problématiques biologiques.

PROJET

- Proposer aux biologistes une base statistique théorique et pratique applicable à tout jeu de données transcriptomiques.
- Réagir sur des problèmes biologiques spécifiques non résolus par une analyse standard.

RESULTATS

• Création et maintenance d'un site web :
www.lsp.ups-tlse.fr/Biopuces
• Mise au point d'une méthodologie standard combinant analyse exploratoire et modélisation pour l'analyse préliminaire d'un jeu de données [1, 3, 4]
• Méthodologie pour le traitement de données cinétiques [2, 5]
• Réunion mensuelle d'un groupe de travail (~15 personnes) sur l'analyse de données d'expression autour des participants à l'ACI

CALENDRIER

2004/05	Analyse standard des jeux de données des partenaires biologistes.
2005/06	Traitement de problèmes spécifiques (cinétique, discrimination).
2006/07	Valorisation des travaux.

LABORATOIRES IMPLIQUES

• Laboratoire de Statistique et Probabilités - U.M.R. C.N.R.S. 5583
• Plate-forme Transcriptome-Biopuces, Génopôle Toulouse Midi-Pyrénées
• Département de Génétique Animale - INRA, Toulouse

MOTS CLES

Biopuces, transcriptome, analyse statistique, logiciel R, interaction biologiste-statisticien

PUBLICATIONS PRINCIPALES

•[1] A. Baccini, P. Besse, S. Déjean, P. Martin, C. Robert-Granié, M. San Cristobal - Stratégies pour l'analyse de données transcriptomiques, *Journal de la Société Française de Statistique, à paraître.*
•[2] A. Baccini, P. Besse, S. Déjean, P. Martin, C. Robert-Granié, M. San Cristobal - Étude de données cinétiques issues de biopuces, *XXXVII[èmes] Journées de Statistique, Pau, 6-10 juin 2005*
• [3] A. Baccini, P. Besse, S. Déjean, C. Robert-Granié, M. San Cristobal - Analyse statistique des données d'expression génomique, *support de cours de la formation INRA-Génopôle Toulouse Midi-Pyrénées, 2005, disponible en ligne* www.lsp.ups-tlse.fr/Biopuces
• [4] S. Déjean - Étude de cas : analyse transcriptomique du cancer pancréatique humain, *support de cours de la formation INRA-Génopôle Toulouse Midi-Pyrénées, 2005, disponible en ligne* www.lsp.ups-tlse.fr/Biopuces
• [5] S. Sokol - Étude de cas : Étude cinétique de fermentation *Saccharomyces Cerevisae* à 13°C et 25°C, *support de cours de la formation INRA-Génopôle Toulouse Midi-Pyrénées, 2005, disponible en ligne* www.lsp.ups-tlse.fr/Biopuces

FORÊTS ALÉATOIRES

Les données

L'expérience menée au sein du Département de Génétique Animale vise à étudier les différences d'expression d'environ un millier de gènes dans des cellules de *granulosa* de truie à trois stades de développement du follicule. Par le biais des réplications techniques et biologiques, nous disposons de 52 biopuces. Le facteur d'intérêt de l'étude est la taille des follicules : Petit, Moyen, Gros. La question biologique posée est la suivante :

Quels sont les gènes qui discriminent au mieux les trois types de follicules ?

Analyse statistique

Une analyse exploratoire combinant analyse en composantes principales et classification permet de voir que la discrimination entre, gros follicules d'une part, petits et moyens d'autre part est relativement facile. En mettant en œuvre une analyse factorielle discriminante, on peut constater par ailleurs que la discrimination des petits et des moyens follicules est possible. Cependant, cette méthode ne permet pas d'identifier des gènes spécifiques.

Les forêts aléatoires permettent de classer les gènes selon leur importance dans la discrimination des trois types de follicules. On peut ainsi extraire un nombre restreint de gènes qui assurent la discrimination en minimisant le taux d'erreur de classement des types.

Représentation simultanée des classifications des biopuces (en colonne) et des 150 gènes les plus importants selon le critère d'importance de Gini (en ligne). On constate que les lames se répartissent en 3 groupes correspondant à la taille des follicules. Apparaît également le fait que les gros follicules se distinguent plus nettement des autres types de follicules.