Les tests multiples en Bioinformatique

Yohann.Foucher@univ-nantes.fr

Equipe d'Accueil 4275 "Biostatistique, recherche clinique et mesures subjectives en santé", Université de Nantes

Master 2 - Bioinformatique, 23 Novembre 2011

UNIVERSITÉ DE NANTES

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

itun institut transplantation urologie néphrologie INSERM - UMR 643

Plan

Plan

Problématique

- Les nouveaux outils d'analyse à haut débit permettent la mesure simultanée de milliers caractéristiques d'un individu

  ex : Puces ADN : elles peuvent mesurer plusieurs milliers de gènes.

- Objectif : identifier des caractéristiques différemment exprimées entre plusieurs groupes.

  ex : identifier les gènes différemment exprimés entre les malades et les non-malades pour identifier des potentielles cibles thérapeutiques.

- Méthode : calcul des probabilités critiques (test de Student si deux groupes)

Problématique

- Supposons que le nombre de gènes étudiés est égal à k
- Deux groupes 1 et 2 de tailles N_1 et N_2
- On observe les expressions du gène j :
  - $x_{j11}, x_{j12}, ..., x_{j1N_1}$ dans le groupe 1 $\rightarrow$ moy. : $\bar{x}_{j1}$ et e.t. s_{j1}
  - $x_{j21}, x_{j22}, ..., x_{j2N_2}$ dans le groupe 2 $\rightarrow$ moy. : $\bar{x}_{j2}$ et e.t. s_{j2}
- Soit t_j la statistique de test de Student pour le gène j

$$t_j = \frac{\bar{x}_{j1} - \bar{x}_{j2}}{\sqrt{\frac{s_{j1}^2}{N_1} + \frac{s_{j2}^2}{N_2}}}$$

- Soit p_j la probabilité critique associée à t_j.

Problème : Nombre variables >>> Nombre de sujets

1. Comparaisons multiples et augmentation du nombre d'erreurs de 1ère espèce.
2. Impossibilité de connaitre la distribution des statistiques de test sous l'hypothèse nulle.

Plan

Multiple Test Procedure (MTP)

- M.J. van der Laan, S. Dudoit, K.S. Pollard (2004), Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives, Statistical Applications in Genetics and Molecular Biology, 3(1).
- S. Dudoit and M.J. van der Laan. Multiple Testing Procedures and Applications to Genomics. Springer Series in Statistics. Springer, New York, 2008.
- Deux grands axes de correction du problème de dimension :
  1. Estimation non-paramétrique des distributions sous H_0 des statistiques de test.
  2. Pénalisation des probabilités critiques pour la prise en compte des comparaisons multiples.

Distributions sous H_0

- Permutation des libellés des colonnes représentant les groupes.

$$\underbrace{1, 1, 1, 1, 1, ..., 1}_{N_1}, \underbrace{2, 2, 2, 2, 2, ..., 2}_{N_2}$$

$$\Downarrow$$

$$\underbrace{1, 2, 1, 2, 2, 1, 1, 1, 2, 2, 1, 1, 2, ..., 1}_{N_1+N_2}$$

- Les expressions des gènes sont alors indépendantes des groupes.

Distribution sous H_0

- On réalise B itérations.
- Pour chaque itération b ($b = 1, 2, ..., B$) :
  - Permutation des colonnes.
  - Calcul des statistiques de test pour chaque gène :

$$t_1^{(b)}, t_2^{(b)}, ..., t_k^{(b)}$$

- Calcul des probabilités critiques (non corrigées) :

$$p_j^* = \frac{\sum_{b=1}^{B} I(|t_j^{(b)}| \geq |t_j|}{B}$$

où $I(a)$ est égale à 1 si la condition a est respectée et 0 sinon.

Correction des probabilités critiques

- La méthode de Bonferroni est la plus connue.
- Pour le gène j, soit $\tilde{p}_j$ la probabilité corrigée associée à p_j^*.

$$\tilde{p}_j = min(kp_j^*, 1)$$

Exemple :

ex : $p_{52}^* = 0.0023$ et $k = 2000$ gènes.

$\rightarrow$ $\tilde{p}_{52} = min(2000 * 0.0023, 1) = min(4.6, 1) = 1.$

Problème pour le 52ème gène :

1. Méthode très conservative.
2. Faible puissance due au non rejet quasi-systématique de H_0.

Correction des probabilités critiques

- Procédure de Holm moins conservative
- Idée : en ordonnant les probabilités critiques, on peut réaliser les tests de la plus faible probabilité à la plus importante. A chaque fois qu'un test est significatif, le test suivant est réalisé parmi le nombre de gènes restant à tester (mois 1 gène à chaque étape).
- Posons $p_{r1} \leq p_{r2} \leq ... \leq p_{rk}$, les probabilités critiques ordonnées en utilisant les valeurs précédentes p_j^* $(j = 1, 2, ..., k)$:

$$\tilde{p}_{r1} = k p_{r1}$$

$$\tilde{p}_{rj} = max(\tilde{p}_{r(j-1)}, (k - j + 1) p_{rj}) \text{ pour } 2 \leq j \leq k$$

- Les probabilités critiques supérieures à 1 sont corrigées à 1.
- Les gènes considérés différemment exprimés sont finalement les gènes j tels que

$$\tilde{p}_{rj} \leq \alpha$$

Plan

Limites

- Ne prend pas en compte la structure de dépendance des gènes

  → Algorithme de Westfall et Young. *

- Même après permutations et corrections, les probabilités critiques sont peu fiables

  → Cette méthode permet de retrouver des résultats plus réalistes et de classer les gènes en fonction de leur potentiel intérêt.

- Survol de la méthode MTP : nombreuses autres possibilités.

- Après classement/sélection des gènes par MTP :

  1. Calcul d'une taille d'échantillon minimale nécessaire pour une validation externe
  2. Méthodes de mesure spécifique des expressions des caractéristiques (PCR, etc.)
  3. On retrouve une situation acceptable pour l'utilisation de la statistique :

$$\text{Nombre variables} <<< \text{Nombre de sujets}$$

*. Resampling-based multiple testing. Peter H. Westfall, S. Stanley Young, Wiley New York, 1993