Exploration et classification de données

Eric Normandeau - 2015-01-27

Club de Bioinformatique de l'IBIS

Objectif

- Partager connaissances en bioinformatique
- Similaire aux Interlabs IBIS
- Dimension analyses (analyses, programmes, code...)
- Exposer son expérience personnelle
- Pas simplement présenter des articles

Disponibilité

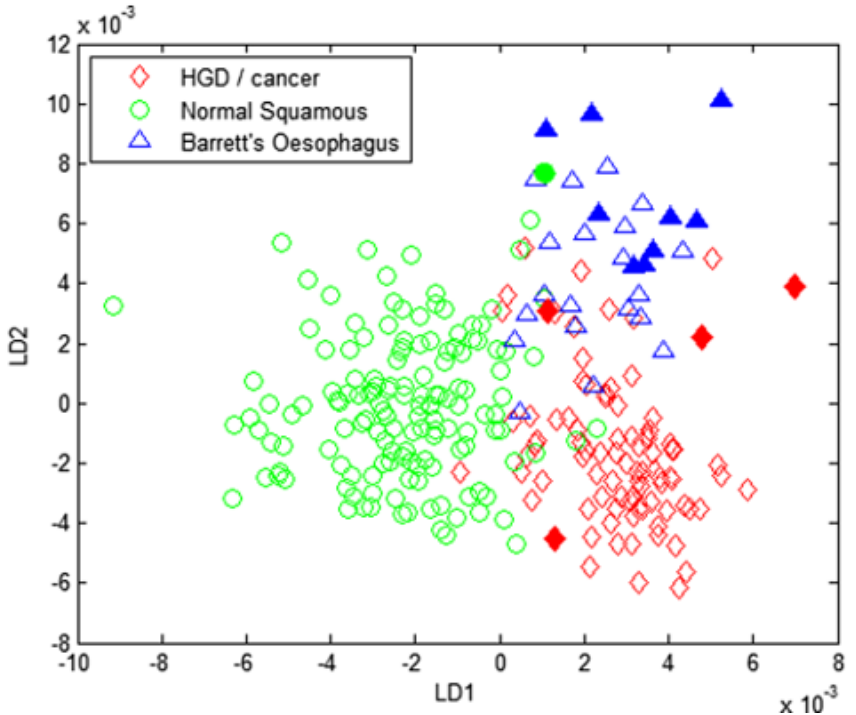
- Présentations et le matériel mis en ligne
- Page GitHub du Club de Bioinfo de l'IBIS

Exploration et classification de données

Objectifs

- Découvrir 3 techniques statistiques
- Exploration et de classification de données
- PCA, LDA, Random Forest
- Tester ces techniques dans R
- Faire un survol superficiel

PCA

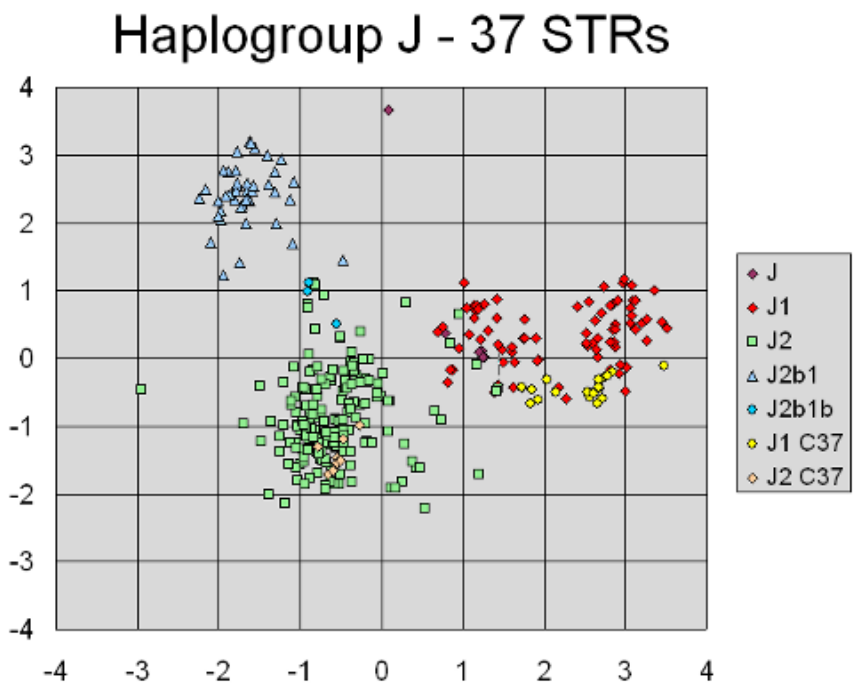

x 10^-3
HGD / cancer
Normal Squamous
Barrett's Oesophagus
LD2
LD1
x 10^-3

PCA

Description

- Explorer les sources de variances principales.
- Variables qui caractérisent des échantillons.
- Importance de différentes sources de variation.
- Aucune connaissances *a priori* des données.
- Non-supervisée (untrained).
- Visualiser le regroupement 'naturel' des données. ## Limites
- Compliqué : *p-valeur* des variables par axe.
- Plutôt, descriptif.

LDA


Haplogroup J - 37 STRs
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
J
J1
J2
J2b1
J2b1b
J1 C37
J2 C37

LDA

Description

- Classification d'individus (sains, infectés...).
- Connaissance *a priori* de la classification.
- Supervisée (trained).
- Problème de classification.
- Visualisation.

Limites

- Grand nombre d'échantillons d'entraînement.
- Moins approprié/puissant si grand nombre de variables.
- Données à classer doivent être similaires aux données test.

Random Forest

Random Forest

Description

- Méthode récente (algoritme publié en 2001).
- Classification ou régression.
- Connaissance *a priori* de la classification.
- Supervisée (trained).
- Problème de classification ou de régression.
- Visualisation.

Random Forest

Utilité en biologie/génétique/génomique

- Très approprié si le nombre de variables est très élevé.
  (Expression de gènes ou marqueurs SNPs)
- Classification d'individus en catégories.
  (Individus sains ou infectés)
  (Individus comme bio-marqueurs de pollution)
- Identifier variables importantes (valeur d'importance).
- Prédiction de réponse par régression.

Random Forest

Limites

- Méthode encore plutôt nouvelle.
- Valeur d'importance réduite pour marqueurs génétiques liés.
- Données à classer doivent être similaires aux données test.