

STAT1 : Introduction au cours

Jacques van Helden
2017-09-05

Pourquoi un cours de proba / stat ?

- Bioinformatique = biologie + informatique + ...
 - modèles probabilistes (séquences, réseaux, évolution, ...)
 - méthodes statistiques (échantillonnage, inférence, tests d'hypothèse, classification, prédiction, ...)
- Génomique = analyse de données massives
 - \Rightarrow nécessité d'appliquer des méthodes statistiques pour extraire l'information pertinente à partir des données.
- Toute analyse de données repose sur des statistiques. On peut faire de bonnes statistiques ou de mauvaises, les formuler explicitement ou pas, mais on n'y échappe pas.

Exemples d'applications

- Recherche de **similarités de séquences** (BLAST):
 - modèles probabilistes d'alignements, significativité des "hits".
- **Détection de motifs dans les séquences**: signaux de régulation (ADN), domaines (protéines)
 - modèles de séquence, découverte et recherche de motifs.
- **Inférence phylogénétique**:
 - modèles évolutifs sous-jacents, méthodes de clustering, méthodes basées sur la vraisemblance.
- **Analyse du transcriptome**:
 - normalisation des données, tests différentiels d'expression, clustering, classification supervisée.
- **Analyse des réseaux**:
 - modèles génératifs de réseaux, motifs sur-représentés
- **Enrichissement fonctionnel de groupes de gènes**:
 - test de sur/sous-représentation
- ...

Compétences attendues

- **Formaliser** en termes de proba/stat une problématique bioinformatique/génomique initialement décrite en termes biologiques.
- Connaître les **distributions de probabilités** les plus utilisées en bioinformatique et génomique, comprendre leurs différences (conditions d'utilisation) et ressemblances (convergence, approximation).
 - Discrètes (binomiale, hypergéométrique, Poisson).
 - Continues (normale, Student, χ^2).
- Connaître les **tests statistiques** les plus courants (Student, Fisher, chi carré).
 - Savoir quel test utiliser pour répondre à quelle question.
 - Vérifier leurs conditions d'application (hypothèses de travail).
 - Formuler les hypothèses à tester (H_0 , H_A).
 - Mettre le test en application sur des données réelles.
 - Interpréter les résultats.

Vos attentes (discussion au cours)

- Quel est votre niveau de départ en proba/stat ?
- Qu'attendez-vous d'un cours de proba/stat pour la bioinfo/génomique ?
- Que craignez-vous ?

Approche pédagogique

- Approche classique: énoncé de la théorie suivie d'exercices illustratifs.
- Approche par résolution de problèmes:
 - En partant d'exemples concrets, on découvre la théorie
 - Approche progressive: on commence avec des cas intuitifs sur lesquels on bâtit ensuite les choses plus complexes.
 - Récapitulation: chaque cycle thématique se termine par une synthèse et une mise en perspective.
- Je peux faire les deux, à vous de choisir.
- On peut changer de stratégie en cours de semestre.

Autres cours de statistiques du M1 BBSG

- Probabilités et statistiques pour la biologie (<http://biologie.univ-mrs.fr/masterBBSG/index.php/fr/stages/m1/157-module-14.html>) (SBBAU16L - STAT1, 3 ECTS)
 - **Jacques van Helden**; obligatoire en M1 BBSG, contenu décrit ci-dessus
- Analyse statistique des données (<http://biologie.univ-mrs.fr/masterBBSG/index.php/fr/stages/m1/166-module-35.html>) (SBBAU12LB - ASD/STAT2, 3 ECTS)
 - **Annie Broglio**
 - Approfondissement des concepts statistiques + Apprentissage du langage R
 - Prérequis pour certains cours du second semestre, et fortement recommandé pour tous.
- Modélisation des séquences et des réseaux biomoléculaires (<http://biologie.univ-mrs.fr/masterBBSG/index.php/fr/stages/m1/177-module-63.html>) (SBBBU4AL - MSR, 3 ECTS)
 - **Badih Gatthas**
- Analyse multidimensionnelle des données génomiques (<http://biologie.univ-mrs.fr/masterBBSG/index.php/fr/stages/m1/178-module-64.html>) (SBBBU4BL, 3 crédits)
 - **Nil Venet (?)**