Machine learning in bioinformatics

6 novembre 2017

Résumé

Revue des méthodes de ML pour la bioinformatique : présentation des méthodes de modélisation (classification supervisée, clustering, modèles probabilistes, heuristiques stochastiques et déterministes pour optimisation). Plusieurs domaines d'application : génomique, protéomique, biologie des systèmes, évolution ...

1 Introduction

La hausse de la quantité de données en bio pose le problème de l'extraction des infos importantes de ces données.

Domaines principaux incluent du ML :

- Génomique : hausse exponentielle des données donc besoin de récupérer l'info utile.
  - Séquences du génôme : localisation et structure des gènes.
  - Identification d'éléments régulatoires (à l'origine de l'expression des gènes).
- Protéomique : prédiction de structures.
- Puces à ADN : il s'agit de données expérimentales donc :
  - besoin d'être pré-traitées pour être utilisables en ML ;
  - analyse (ex : identification de motifs, classification, etc)
- Biologie des systèmes : modélisation de processus à l'intérieur des cellules très complexe → réseaux génétiques, réseau de transduction de signal (=réponse de la cellule à l'info qu'elle reçoit) etc
- Evolution : les arbres phylogénétiques (représentation schématique de l'évolution des organismes) peuvent être reconstruits grâce au ML → avant, ils étaient contsruits à partir de caractéristiques morphologiques, métaboliques, etc et maintenant, on peut construire ces arbres en se basant sur les différents génomes (avec de l'alignement de séquences).

2 Classification supervisée

- Explication du principe de classification supervisée, de la mesure de performance (courbe ROC,cross-validation, bootstrap ...)
- Algorithmes utilisés :
  - Classifieur Bayésien
  - Régression logistique
  - Analyse discriminante
  - Arbres de classification
  - Pus proches voisins
  - Réseaux de neurones
  - Machines à vecteurs de support
  - Combinaison de classifieurs

3 Clustering

2 approches :

— Partitionnement : le but est d'obtenir une partition des données (comme avec l'algo des k-means et ses variantes) alors qu'on ne sait même pas combien de classes il y a.

— Hiérarchisation : ensemble de partitions représentées pas un arbre. Le nombre de partitions dépend du niveau de l'arbre qu'on observe (plus la profondeur regardée sera grande, plus il y aura de partitions).

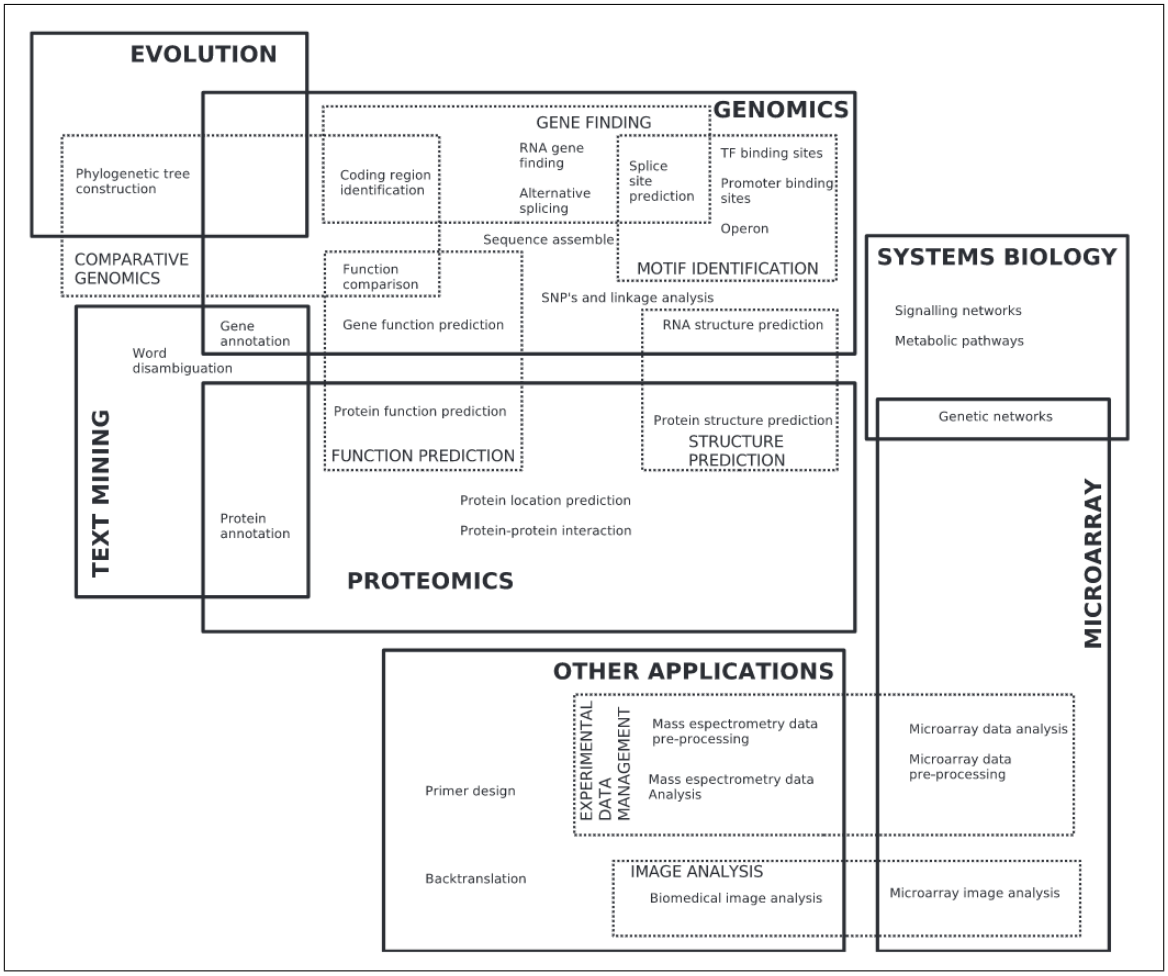

FIGURE 1 – Représenation schématique des dommaines de la bio où le ML s'applique

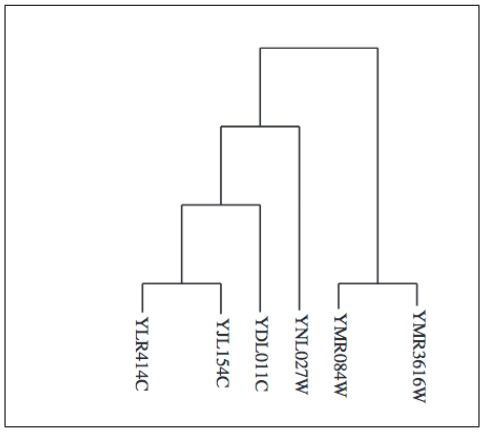

FIGURE 2 – Exemple de hiérarchisation.