Analyse comparée d'arbres taxonomiques

C. Réda
sous la direction de M. Nikolski & M. Raffinot

CGFB, équipe CBIB, Bordeaux

du 1er juin au 27 juillet 2016

Informations sur le stage

- Lieu : **Centre de Génomique Fonctionnelle** (CGFB)
- Equipe : **Centre de BioInformatique** (CBIB)
- Dates : du 1er juin au 27 juillet 2016

Projets de l'équipe CBIB

- Une dizaine de logiciels de bioinformatique, dont Tango;

Projets de l'équipe CBIB

- Une dizaine de logiciels de bioinformatique, dont Tango;
- Participation et 3^{eme} place au concours international **Dream Challenge** ;

Projets de l'équipe CBIB

- Une dizaine de logiciels de bioinformatique, dont Tango;
- Participation et 3^{eme} place au concours international **Dream Challenge** ;

- Participation au projet **Galaxy** ;

Projets de l'équipe CBIB

- Une dizaine de logiciels de bioinformatique, dont Tango;
- Participation et 3^{eme} place au concours international **Dream Challenge** ;

- Participation au projet **Galaxy** ;
- Collaborations avec :
  - l'**Inra** de Bordeaux,

Projets de l'équipe CBIB

- Une dizaine de logiciels de bioinformatique, dont Tango;
- Participation et 3^{eme} place au concours international **Dream Challenge** ;

- Participation au projet **Galaxy** ;
- Collaborations avec :
  - l'**Inra** de Bordeaux,
  - le **LaBRI**,

Projets de l'équipe CBIB

- Une dizaine de logiciels de bioinformatique, dont Tango;
- Participation et 3^{eme} place au concours international **Dream Challenge** ;

- Participation au projet **Galaxy** ;
- Collaborations avec :
  - l'**Inra** de Bordeaux,
  - le **LaBRI**,
  - **GenoToul Bioinformatique** à Toulouse,

Projets de l'équipe CBIB

- Une dizaine de logiciels de bioinformatique, dont Tango;
- Participation et 3^{eme} place au concours international **Dream Challenge** ;

- Participation au projet **Galaxy** ;
- Collaborations avec :
  - l'**Inra** de Bordeaux,
  - le **LaBRI**,
  - **GenoToul Bioinformatique** à Toulouse,
  - et l'hôpital **Pellegrin** à Bordeaux.

Traitement du matériel génétique brut

- Extraction de l'ADN par réaction chimique;

Traitement du matériel génétique brut

- Extraction de l'ADN par réaction chimique;
- Séquençage de l'ADN obtenu : obtention de la structure primaire de l'ADN.

Traitement du matériel génétique brut

- Extraction de l'ADN par réaction chimique;
- Séquençage de l'ADN obtenu : obtention de la structure primaire de l'ADN.

Next-Generation Sequencing (NGS)

Méthode rapide, relativement bon marché de séquençage de l'ADN, encline aux erreurs.

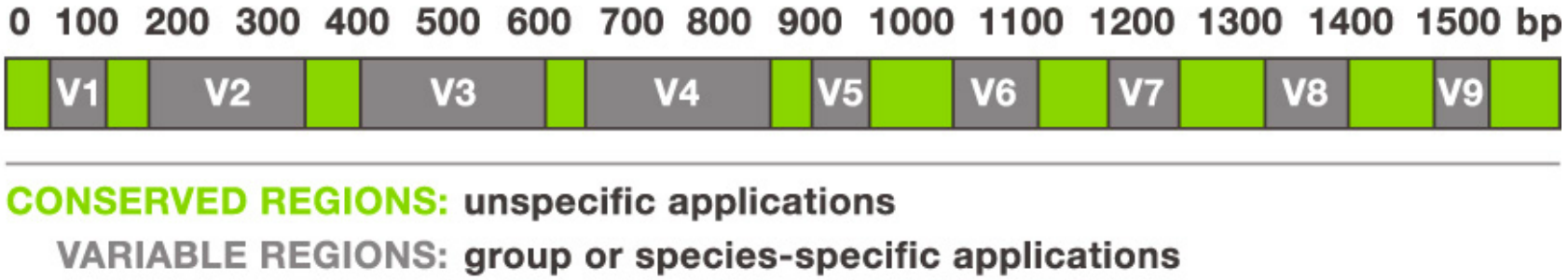
Donne des morceaux de séquences (reads) d'une longueur de 32 à 1 000 paires de bases.

Identification des reads à des espèces

- Séquençage réalisé sur les gènes 16S;

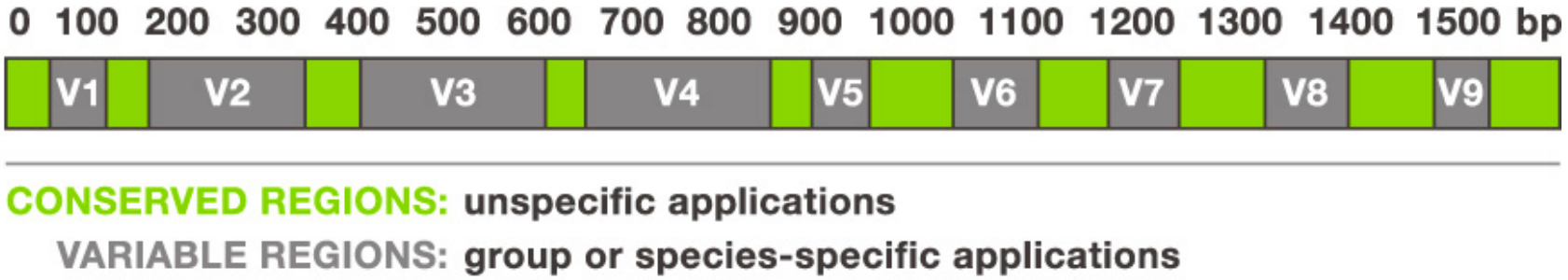
Identification des reads à des espèces

- Séquençage réalisé sur les gènes 16S;
- Limites des régions hyper variables difficiles à évaluer;

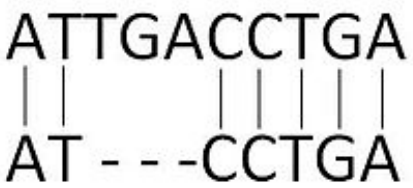


Identification des reads à des espèces

- Séquençage réalisé sur les gènes 16S;
- Limites des régions hyper variables difficiles à évaluer;



- Alignement des reads obtenus sur ces régions à des séquences de référence.

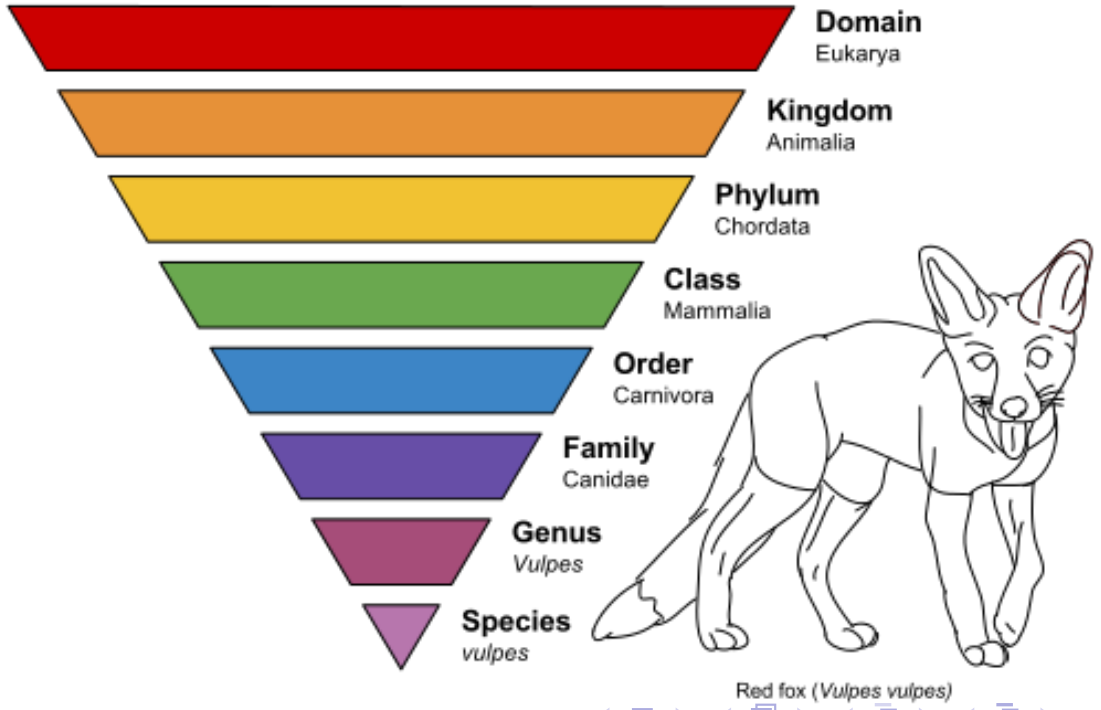


Arbres taxonomiques

Arbre taxonomique

Graphe connexe acyclique non orienté de hauteur bornée, correspondant à l'histoire évolutive du monde vivant.

```
R:Root
 K:Bacteria
  P:AC1
   C:LC-1
   C:SHA-114
  P:MSBL6
   C:KB1
   C:AT425_EubG1
  P:Verrucomicrobia
   C:Methylacidiphilae
    O:Methylacidiphilales
     F:LD19
     F:Methylacidiphilaceae
      G:LP2A
      G:Candidatus Methylacidiphilum
       S:Methylacidiphilum infernorum
   C:TP21
   C:Verruco-5
    O:RFP12
   C:Spartobacteria
    O:Spartobacteriales
     F:Spartobacteriaceae
      G:Chthoniobacter
       S:Chthoniobacter flavus
      G:MC18
      G:Candidatus Xiphinematobacter
   C:Verrucomicrobiae
    O:Verrucomicrobiales
     F:Verrucomicrobiaceae
      G:MSBL3
      G:Verrucomicrobium
```



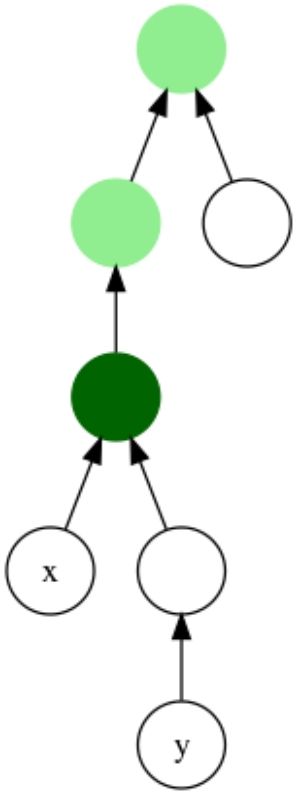
Arbres taxonomiques

Caractéristiques

- Plus un noeud est proche de la racine,
  moins son degré est grand;
- Nettement plus de feuilles que de noeuds internes;
- Assignation d'un read plus complexe qu'il n'y paraît.

Least Common Ancestor (LCA) de A et B

Dernier noeud de la partie commune des chemins de la racine jusqu'à A et B.



Problématiques de la métagénomique

Existence d'algorithmes :

Problématiques de la métagénomique

Existence d'algorithmes :

- améliorant l'alignement des reads aux séquences (Smith et Waterman, 1981);

Problématiques de la métagénomique

Existence d'algorithmes :

- améliorant l'alignement des reads aux séquences (Smith et Waterman, 1981);
- améliorant l'assignation dans l'arbre des reads matchés (Clemente et al., 2011 : l'outil **Tango**);

Problématiques de la métagénomique

Existence d'algorithmes :

- améliorant l'alignement des reads aux séquences (Smith et Waterman, 1981);
- améliorant l'assignation dans l'arbre des reads matchés (Clemente et al., 2011 : l'outil **Tango**);
- implémentant des mesures quantifiant la pertinence d'un arbre taxonomique (Robinson et Foulds, 1981)
- ...

Problématiques de la métagénomique

Existence d'algorithmes :

- améliorant l'alignement des reads aux séquences (Smith et Waterman, 1981);
- améliorant l'assignation dans l'arbre des reads matchés (Clemente et al., 2011 : l'outil **Tango**);
- implémentant des mesures quantifiant la pertinence d'un arbre taxonomique (Robinson et Foulds, 1981)
- ...

Mais...

Problème des paires les plus dissemblables

Entrée :

- Une matrice d'occurrence des assignations dans les échantillons.

Sortie : L'ensemble des paires d'échantillons les plus dissemblables.

Problème de compatibilité de la classification

Entrée :

- Un sous-ensemble N de noeuds/bactéries;
- Un tableau contenant les noeuds matchés dans chaque échantillon;
- Un sous-ensemble M de métadonnées.

Sortie : Existe-t-il une correspondance entre N et M ?

Problème de meilleure classification

Entrée :

- Un tableau contenant les noeuds matchés dans chaque échantillon;
- Un sous-ensemble de métadonnées M.

Sortie : Un sous-ensemble N de noeuds tel que N ait une correspondance avec M.

Rappel des problèmes

- **Problème des paires les plus dissemblables**
  Sortie : L'ensemble des paires d'échantillons les plus dissemblables.

Distance calculée

Coefficient de similarité s

- Si $MD(G_1) - MD(G_2) = 0$:

$$s(G_1, G_2) = TR(G_1, G_2) + PR(G_1, G_2)$$

- Sinon :

$$s(G_1, G_2) = TR(G_1, G_2) + PR(G_1, G_2) - |MD(G_1) - MD(G_2)|$$

Distance calculée

Coefficient de similarité s

- Si $MD(G_1) - MD(G_2) = 0$:

$$s(G_1, G_2) = TR(G_1, G_2) + PR(G_1, G_2)$$

- Sinon :

$$s(G_1, G_2) = TR(G_1, G_2) + PR(G_1, G_2) - |MD(G_1) - MD(G_2)|$$

Coefficient de similarité $\overline{s}$

$$\overline{s}(G_1, G_2) = \frac{s(G_1, G_2) - E(s)}{\sigma(s)}$$

Problème : trouver des mesures pertinentes !

Rappel des problèmes

- **Problème de compatibilité de la classification**
  Sortie : Existe-t-il une correspondance entre N et M ?
- **Problème de meilleure classification**
  Sortie : Un sous-ensemble N de noeuds tel que N ait une correspondance avec M.

Quelques notions de *Machine Learning*

Machine Learning

Paradigme qui automatise la reconnaissance de certains motifs.

Quelques notions de *Machine Learning*

Machine Learning

Paradigme qui automatise la reconnaissance de certains motifs.

Apprentissage supervisé

Classification des données dans des catégories fixées, avec une connaissance *a priori* acquise sur un ensemble d'entraînement.

Le classificateur naïf bayésien

Classificateur naïf bayésien

Pour k classes disjointes de données $(C_i)_{1\leq i\leq k}$, des critères $(F_i)_{1\leq i\leq m}$, et une donnée d à classer, ayant $(F_i = x_i)_{1\leq i\leq m}$,
La classe de d est la classe C_j telle que:

$$P(C_j|F_1 = x_1, ..., F_m = x_m)$$
$$= \max_{h\in\{1,...,k\}} P(C_h|F_1 = x_1, ...F_m = x_m).$$

Calcul de la pertinence de la classification obtenue

Coefficient J de Youden

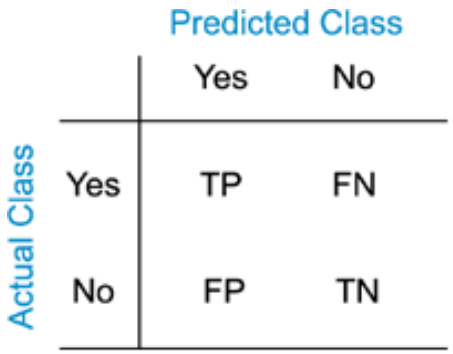
$$J(C) = \frac{TP(C)}{TP(C)+FN(C)} + \frac{TN(C)}{TN(C)+FP(C)} - 1.$$

		Predicted Class: Yes	Predicted Class: No
Actual Class	Yes	TP	FN
	No	FP	TN

Calcul de la pertinence de la classification obtenue

Coefficient J de Youden

$$J(C) = \frac{TP(C)}{TP(C)+FN(C)} + \frac{TN(C)}{TN(C)+FP(C)} - 1.$$



Coefficient modifié pour k classes $(C_i)_{1 \leq i \leq k}$

Classification "optimale" : $k - \sum_{i=1}^{k} J(C_i)$ minimal, positif.

Le **TaxoClassifier** : étapes d'une classification

Etant donnés un ensemble de métadonnées M (et donc des classes induites par M) et de noeuds N :

Le **TaxoClassifier** : étapes d'une classification

Etant donnés un ensemble de métadonnées M (et donc des classes induites par M) et de noeuds N :

1. Entraînement du classificateur;

Le **TaxoClassifier** : étapes d'une classification

Etant donnés un ensemble de métadonnées M (et donc des classes induites par M) et de noeuds N :

1. Entraînement du classificateur;
2. Pour chaque échantillon non assigné, calcul des probabilités postérieures à chaque classe;

Le classificateur naïf bayésien

Classificateur naïf bayésien

$$P(C_j|F_1 = x_1, ..., F_m = x_m)$$
$$= \max_{h \in \{1,...,k\}} P(C_h|F_1 = x_1, ...F_m = x_m).$$

Le **TaxoClassifier** : étapes d'une classification

Etant donnés un ensemble de métadonnées M (et donc des classes induites par M) et de noeuds N :

1. Entraînement du classificateur;
2. Pour chaque échantillon non assigné, calcul des probabilités postérieures à chaque classe;
3. Assignation de chaque échantillon à la classe qui maximise la probabilité précédente;

Le **TaxoClassifier** : étapes d'une classification

Etant donnés un ensemble de métadonnées M (et donc des classes induites par M) et de noeuds N :

1. Entraînement du classificateur;
2. Pour chaque échantillon non assigné, calcul des probabilités postérieures à chaque classe;
3. Assignation de chaque échantillon à la classe qui maximise la probabilité précédente;
4. Retour du coefficient J de Youden modifié associé à cette classification.

Le **TaxoClassifier** : étapes d'une classification

Etant donnés un ensemble de métadonnées M (et donc des classes induites par M) et de noeuds N :

1. Entraînement du classificateur;
2. Pour chaque échantillon non assigné, calcul des probabilités postérieures à chaque classe;
3. Assignation de chaque échantillon à la classe qui maximise la probabilité précédente;
4. Retour du coefficient J de Youden modifié associé à cette classification.

Problème(s) : Beaucoup d'hypothèses *a priori*, et de petites améliorations nécessaires pour gérer les cas limites !

Rappel des problèmes

- **Problème de meilleure classification**
  Sortie : Un sous-ensemble N de noeuds tel que N ait une correspondance avec M.

Notion de clustering

Apprentissage non supervisé

Identification des différentes classes de données, en étudiant la similarité entre les données.

Notion de clustering

Apprentissage non supervisé

Identification des différentes classes de données, en étudiant la similarité entre les données.

Clustering

Partition d'un ensemble de données, telle que :

- les parties obtenues minimisent la distance entre les objets d'un même groupe;
- la maximisent entre les objets de deux groupes différents.

Problème de clustering

Complexité du problème de partition en k *clusters* de *n* éléments

Le problème de *k*-partition de *n* éléments est NP-complet.

Problème de clustering

Complexité du problème de partition en k *clusters* de *n* éléments

Le problème de *k*-partition de *n* éléments est NP-complet.

Etapes de l'algorithme des K-moyennes

1. Initialisation des *k* clusters
2. Tant que les clusters sont modifiés pendant la boucle
   - Pour tout élément e de l'ensemble de départ
     - Déterminer le cluster C qui minimise la distance entre e et sa moyenne
     - Affecter e à C
     - Recalculer la moyenne de C

Notations utilisées dans **TaxoCluster**

Soit T l'arbre taxonomique complet. Pour un certain read i :

- M_i, ensemble de feuilles matchées par i;

Notations utilisées dans **TaxoCluster**

Soit T l'arbre taxonomique complet. Pour un certain read i :

- M_i, ensemble de feuilles matchées par i;
- T_i, sous-arbre de T enraciné au LCA des feuilles de M_i;

Notations utilisées dans **TaxoCluster**

Soit T l'arbre taxonomique complet. Pour un certain read i :

- M_i, ensemble de feuilles matchées par i;
- T_i, sous-arbre de T enraciné au LCA des feuilles de M_i;
- L_i, ensemble des feuilles de T_i;

Notations utilisées dans **TaxoCluster**

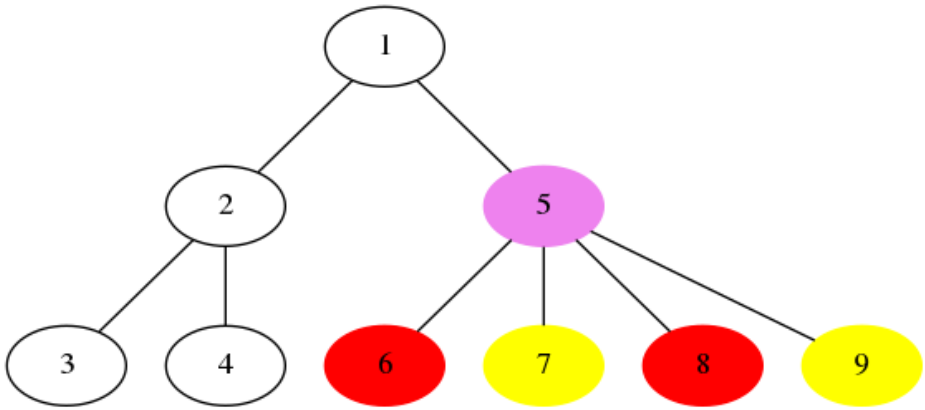
Soit T l'arbre taxonomique complet. Pour un certain read i :

- M_i, ensemble de feuilles matchées par i;
- T_i, sous-arbre de T enraciné au LCA des feuilles de M_i;
- L_i, ensemble des feuilles de T_i;
- N_i, tel que $L_i = M_i \sqcup N_i$.

Notations utilisées dans **TaxoCluster**

Soit T l'arbre taxonomique complet. Pour un certain read i :

- M_i, ensemble de feuilles matchées par i;
- T_i, sous-arbre de T enraciné au LCA des feuilles de M_i;
- L_i, ensemble des feuilles de T_i;
- N_i, tel que $L_i = M_i \sqcup N_i$.



Distances utilisées dans **TaxoCluster**

"Distance des ensembles de matches"

Pour deux reads i et j,

$$d_{matched}(i,j) = |M_i| + |M_j| - 2 * |M_i \cap M_j|.$$

Distances utilisées dans **TaxoCluster**

"Distance des ensembles de matches"

Pour deux reads i et j,

$$d_{matched}(i,j) = |M_i| + |M_j| - 2 * |M_i \cap M_j|.$$

"Distance de l'arbre de consensus"

Pour deux reads i et j, et un paramètre $q \in [0;1]$,

$$d_{consensus}(i,j) = |L_i| + |L_j| - q * (|N_i \cap M_j| + |N_j \cap M_i|) - |M_i \cap M_j|.$$

Distance de l'arbre de consensus

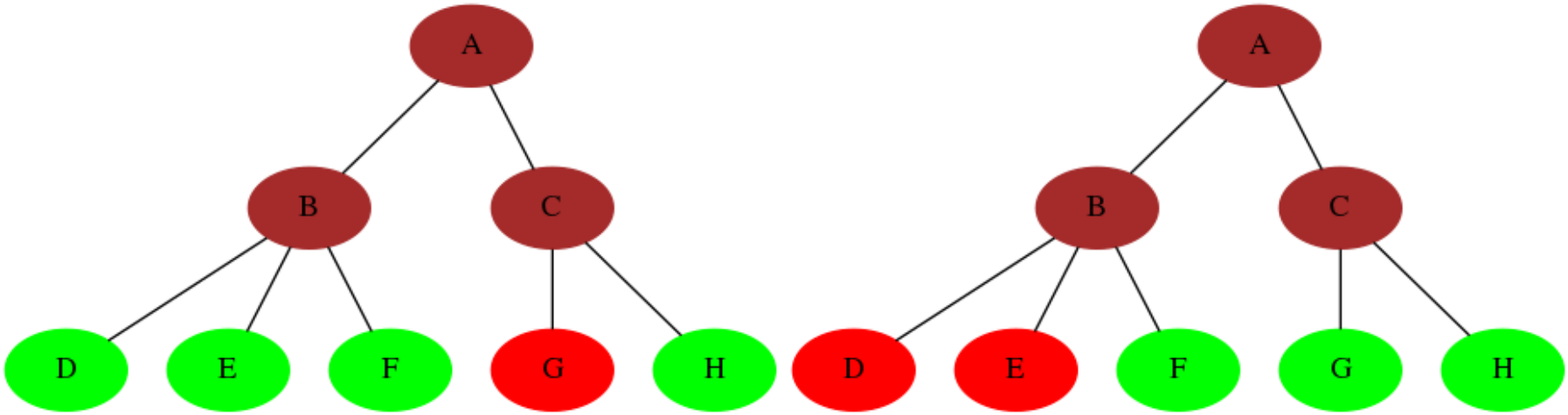


Figure: Quand $q = 0$, $d_{consensus}(T_1, T_2) = 5 + 5 - 0 - 2 = 8$.
Quand $q = 1$, $d_{consensus}(T_1, T_2) = 5 + 5 - 1 \times (1 + 2) - 2 = 5$

Le programme **TaxoCluster**

1. Application de l'algorithme des K-moyennes avec $d_{matched}$

Le programme **TaxoCluster**

1. Application de l'algorithme des K-moyennes avec $d_{matched}$
2. Suppression des éléments les moins pertinents

Le programme **TaxoCluster**

1. Application de l'algorithme des K-moyennes avec $d_{matched}$
2. Suppression des éléments les moins pertinents
3. Application de l'algorithme des K-moyennes avec $d_{consensus}$

Le programme **TaxoCluster**

1. Application de l'algorithme des K-moyennes avec $d_{matched}$
2. Suppression des éléments les moins pertinents
3. Application de l'algorithme des K-moyennes avec $d_{consensus}$
4. Comparaison des clusters obtenus avec ceux induits par les métadonnées

Implémentation

- Implémentés en Python 2.9.7;

Implémentation

- Implémentés en Python 2.9.7;
- Disponibles sur GitHub;

Implémentation

- Implémentés en Python 2.9.7;
- Disponibles sur GitHub;
- **TaxoCluster** et **TaxoClassifier** encore en développement.

Evaluation

- Evaluation en comparant les résultats obtenus par les différents logiciels et les résultats statistiques de l'étude de l'hôpital Pellegrin;

Evaluation

- Evaluation en comparant les résultats obtenus par les différents logiciels et les résultats statistiques de l'étude de l'hôpital Pellegrin;
- Résultats donnés par **TaxoTree** (première méthode) confirmant les résultats statistiques;

Evaluation

- Evaluation en comparant les résultats obtenus par les différents logiciels et les résultats statistiques de l'étude de l'hôpital Pellegrin;
- Résultats donnés par **TaxoTree** (première méthode) confirmant les résultats statistiques;
- Manque de temps pour les tests de **TaxoClassifier** et **TaxoCluster**.

Evaluation

- Evaluation en comparant les résultats obtenus par les différents logiciels et les résultats statistiques de l'étude de l'hôpital Pellegrin;
- Résultats donnés par **TaxoTree** (première méthode) confirmant les résultats statistiques;
- Manque de temps pour les tests de **TaxoClassifier** et **TaxoCluster**.
- En théorie : méthode de TaxoCluster meilleure que les autres;

Evaluation

- Evaluation en comparant les résultats obtenus par les différents logiciels et les résultats statistiques de l'étude de l'hôpital Pellegrin;
- Résultats donnés par **TaxoTree** (première méthode) confirmant les résultats statistiques;
- Manque de temps pour les tests de **TaxoClassifier** et **TaxoCluster**.
- En théorie : méthode de TaxoCluster meilleure que les autres;
- En théorie : pire complexité temporelle pour **TaxoCluster** (pire cas).

Bilan et perspectives

1. Propositions de trois méthodes pour répondre au(x) problème(s);

Bilan et perspectives

1. Propositions de trois méthodes pour répondre au(x) problème(s);
2. Utilisables sur un ordinateur de puissance moyenne;

Bilan et perspectives

1. Propositions de trois méthodes pour répondre au(x) problème(s);
2. Utilisables sur un ordinateur de puissance moyenne;
3. Nécessité d'autres tests.

Sources

- **Flexible taxonomic assignment of ambiguous sequencing**, J. Clemente, J. Jansson et G. Valiente, *BMC Bioinformatics*, 2011.
- **Impact de l'antibiothérapie sur le microbiote intestinal chez l'enfant atteint de mucovisidose**, R. Enaud, *Université de Bordeaux, CHU Pellegrin*, 2016.
- **Understanding Machine Learning: From Theory to Algorithms**, S. Shalev-Shwartz et S. Ben-David, *Cambridge University Press*, 2014.
- ...