LaBRI

Présentation scientifique

CoBalt

Guillaume Blin

cnrs

université de BORDEAUX

Bordeaux INP AQUITAINE

CoBalt : Axes de recherche

Recherche appliquée à la biologie et à la santé

Résoudre des problèmes informatiques émergeant de l'analyse de données biologiques

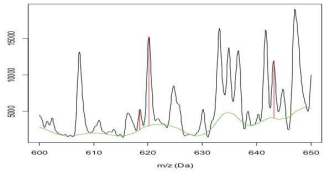
Données

Séquences biologiques

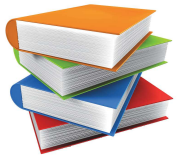
ACT ··· GATTACA ··· CGT ··· TTT ...CGGAT
AGT ··· ATTACAT ··· CCT ··· TTT ...CCGGT
GCT ··· TTGCTAT ··· CGT ··· AAA ...CGGAT
ACT ··· GACTTCA ··· GGT ···CTT ...CAGGT
AGT ··· ATTACAT ··· CCT ··· TTT ...CCGGT
CCT ··· TCGTCTC ··· CGT ··· ATA ...GGGAT

Structures biologiques

Mesures

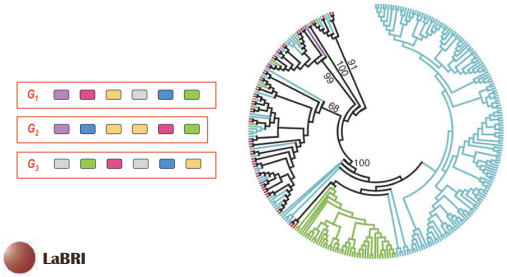


Connaissances

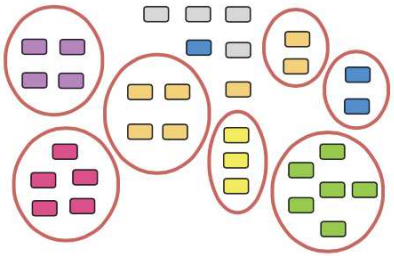


Questions biologiques ou de santé

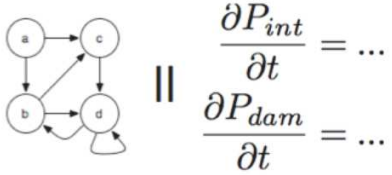
Evolution



Organisation



Modélisation de processus



LaBRI

cnrs université BORDEAUX

CoBalt : Axes de recherche

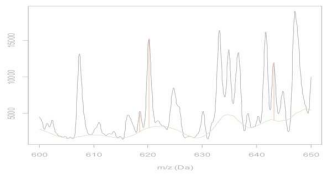
Données

Séquences biologiques

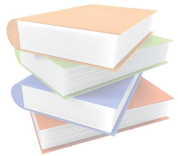
ACT ··· GATTACA ··· CGT ··· TTT ...CGGAT
AGT ··· ATTACAT ··· CCT ··· TTT ...CCGGT
GCT ··· TTGCTAT ··· CGT ··· AAA ...CGGAT
ACT ··· GACTTCA ··· GGT ···CTT ...CAGGT
AGT ··· ATTACAT ··· CCT ··· TTT ...CCGGT
CCT ··· TCGTCTC ··· CGT ··· ATA ...GGGAT

Structures biologiques

Mesures



Connaissances

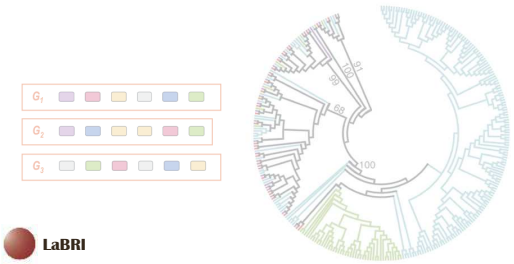


Informatique

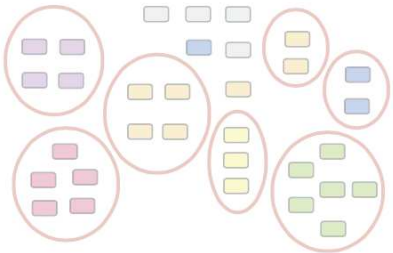
Expertise Bio	Manipulation données	Modèles informatiques bio-numériques	Algorithmique	Analyse	Programmation
ARNnc	*RFAM*	*Séquence*	*Algo du texte*	*Statistique*	
Réseaux de régulation	*BDD*	*Séquence arc-annotée*	*Comparaison*	*Spécificité/Sensabilité*	
Réseaux métaboliques	*Structures ARN*	*Arbre*	*Recherche de motifs*	*Complexité*	
Cancer	*ADN*	*Graphe*	*Alignement*		
Génomique	*Pseudogenes*	*Ordre partiel*	*Simulation*		
Protéine	*NGS*	*Automate de mode stochastique*			
...					

Questions biologiques ou de santé

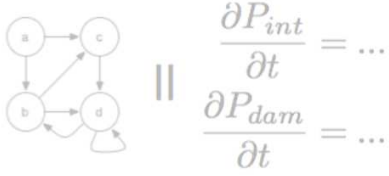
Evolution



Organisation



Modélisation de processus



LaBRI

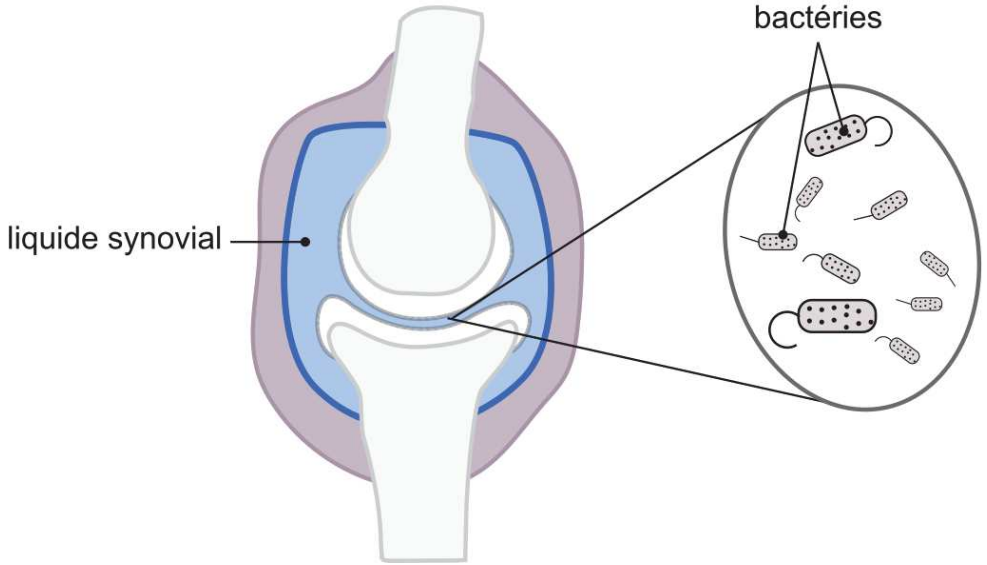
cnrs université BORDEAUX

CoBalt : Illustration de travaux

Arthrite infectieuse du nourrisson et de l'enfant

Inflammation due à l'invasion puis la prolifération d'un germe dans une articulation

Prévalence [2] : 23/100 000 habitants, supérieure chez les plus petits (< 2 ans) et chez les garçons



Données

Prélèvement de liquide synovial chez des patients sains et d'une cohorte de malades

Questions de santé

Hypothèse à valider : « Sans inoculation directe (plaie), les bactéries migrent des voies respiratoires vers l'articulation par voie hématogène (circulation sanguine). »

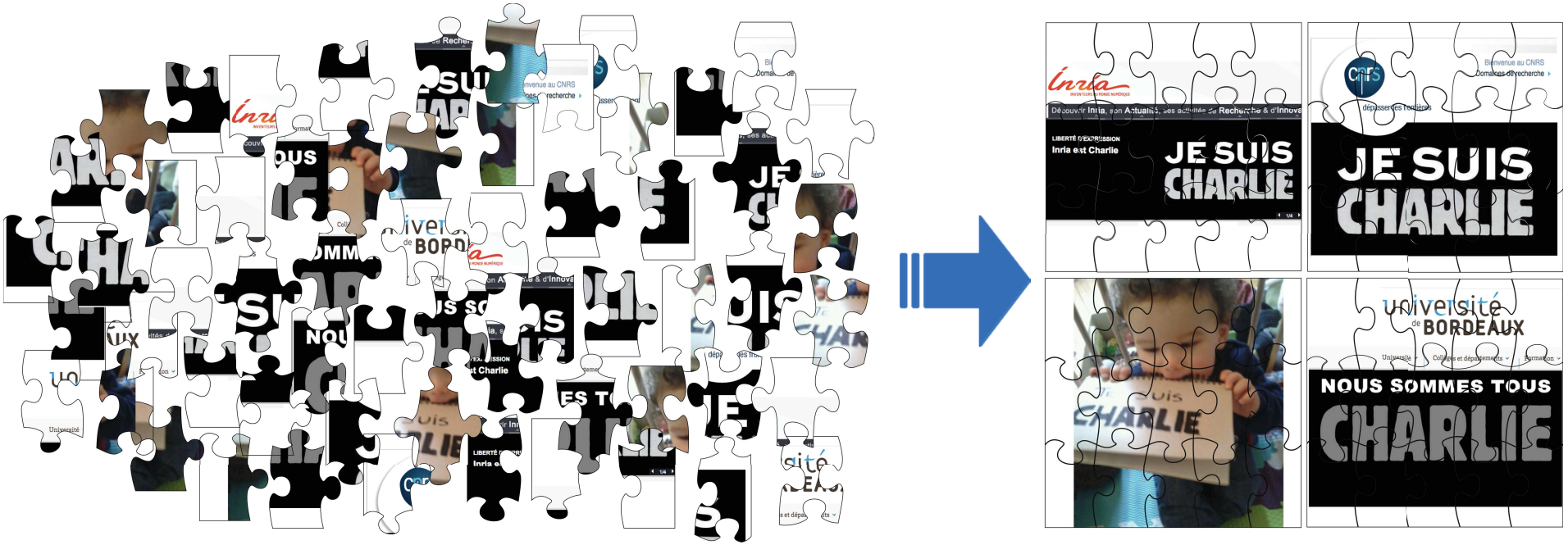
[2] Epidémiologie des infections ostéo-articulaires de l'enfant en France : analyses des données médico-administratives. Grammatico-Guillon et. al JNI 2012

LaBRI

CoBalt : Illustration de travaux

Métagénomique

Détermination et quantification de la composition en espèces d'un échantillon issu d'un environnement complexe



Données

Séquences ADN (*reads*) issues de technologies de séquençage haut-débit (*NGS*)

Informatique

Amélioration de TANGO [1]
Affectation taxonomique des *reads*

Questions de santé

Composition en espèces des données ?

[1] Further steps in TANGO: improved taxonomic assignment in metagenomics. Alonso-Alemany D. et al. Bioinformatics. 2014

CoBalt : Illustration de travaux

Pipeline

Quelques millions de *reads* courts (75 bases) par patient (données erronées, incomplètes, morcelées)

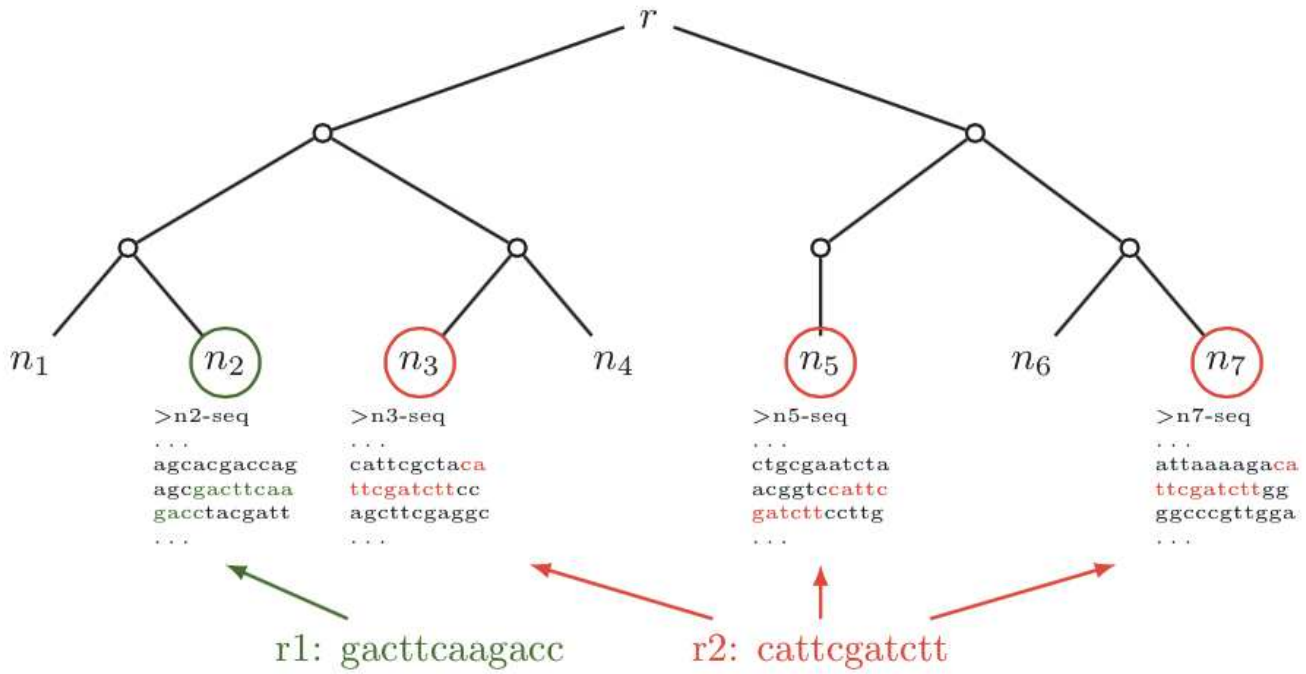
Filtrage

Alignement des *reads* sur une BDD de séquences des éventuels contaminants
Tout *read* susceptible de correspondre à du contaminant est écarté

Alignement de reads

Alignement des *reads* sur des séquences de référence d'une classification arborescente
Possible ambiguïté

TANGO



LaBRI

CoBalt : Illustration de travaux

Pipeline

Quelques millions de reads courts (75 bases) par patient (données erronées, incomplètes, morcelées)

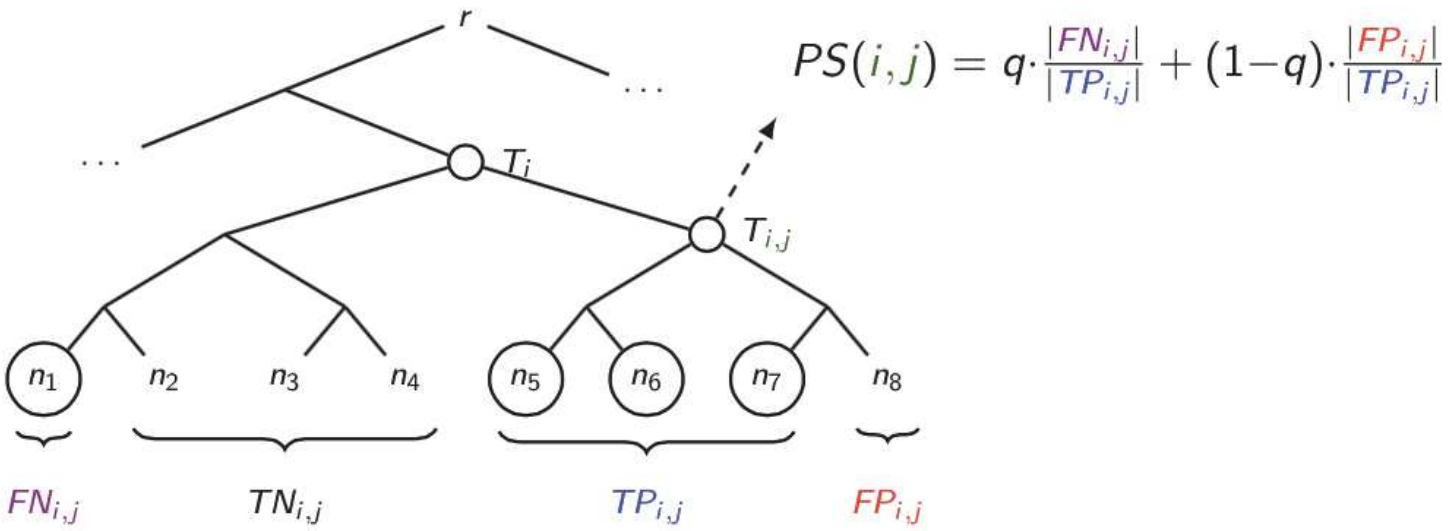
Filtrage

Alignement des *reads* sur une BDD de séquences des éventuels contaminants
Tout read susceptible de correspondre à du contaminant est écarté

Alignement de reads

Alignement des *reads* sur des séquences de référence d'une classification arborescente
Possible ambiguïté

TANGO



Résultat : la composition en espèces de la classification utilisée de l'environnement étudié

CoBalt : Illustration de travaux

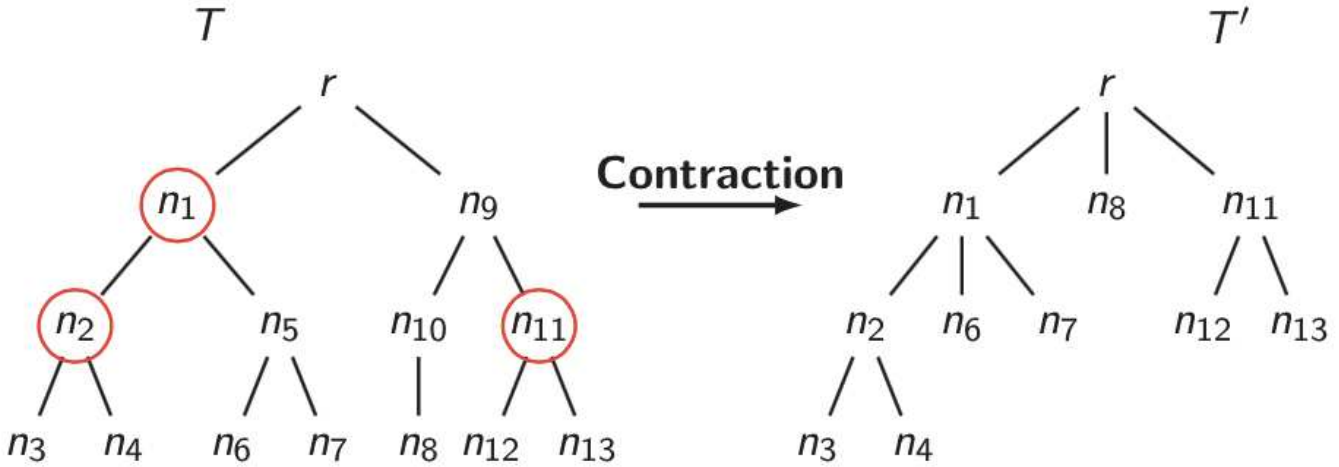
Limite de TANGO

Problème de passage à l'échelle (ici des arbres avec des millions de feuilles)
Il n'y a pas de classification arborescente consensus pour les bactéries (e.g. NCBI, Greengenes, Ribosomal Database Project)

Apports des membres de CoBalt

Passage à l'échelle par l'utilisation de structures de données adaptées
Ajout de la prise en compte de classifications multiples

Preprocessing Contraction des taxonomies suivant un ensemble de rangs taxonomiques
règne → embranchement → classe → ordre → famille → genre → espèce
animal → vertébrés → mammifères → carnivores → félin → félis → chat domestique



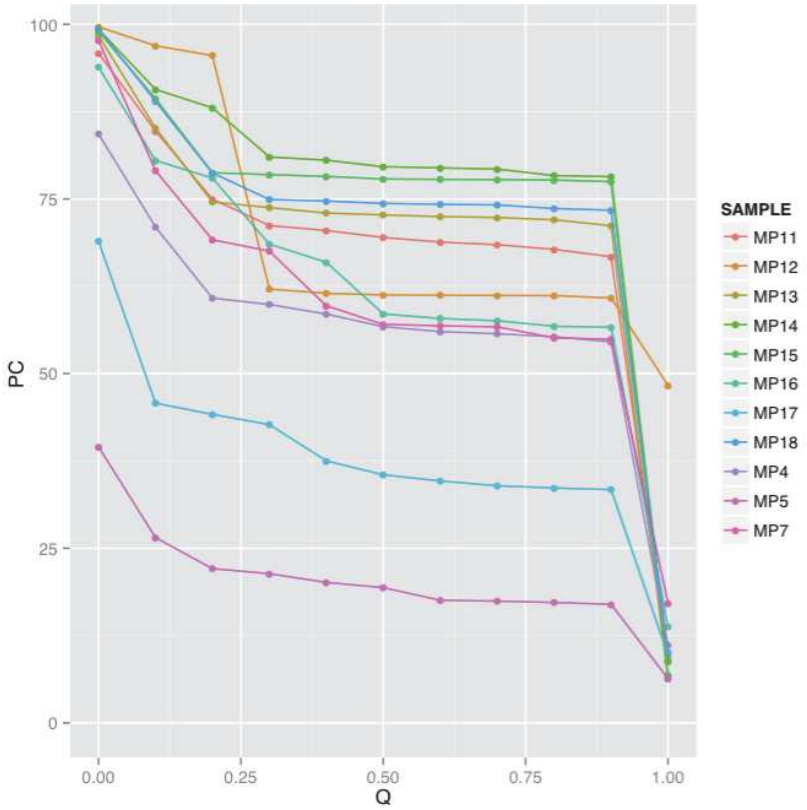
Alignement entre les taxonimies contractées (permettant une traduction de la solution)
Basé sur le LCA

CoBalt : Illustration de travaux

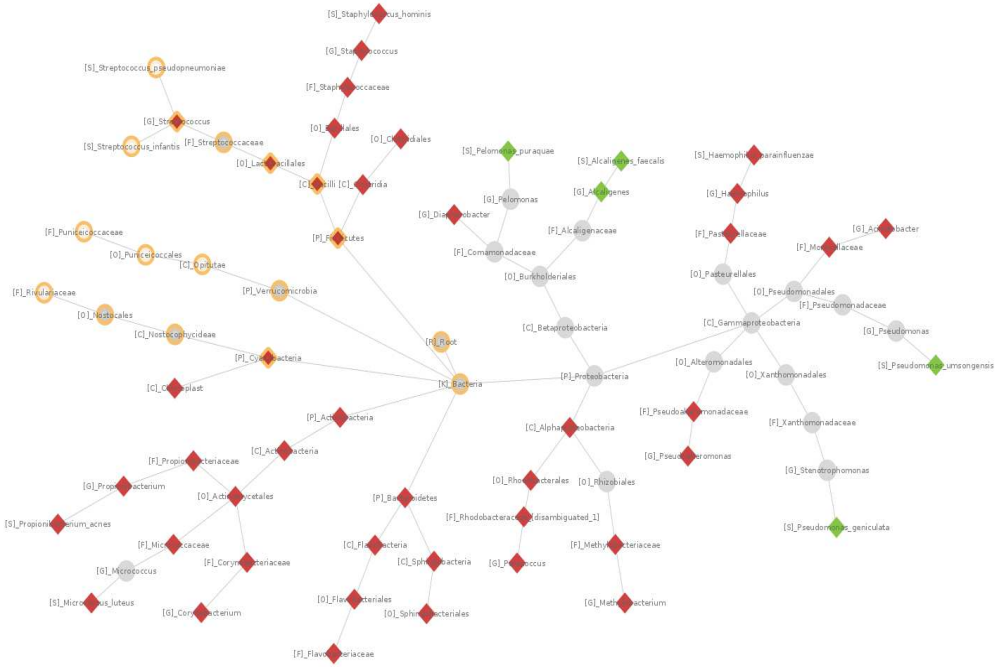
Réponses apportées à la question de santé

Outils statistiques

Proportions d'assignation au niveau des feuilles pour les différents échantillons suivant la valeur de Q.



Noeuds de la taxonomie GreenGenes statistiquement différentiellement exprimés entre patients sains et malades.

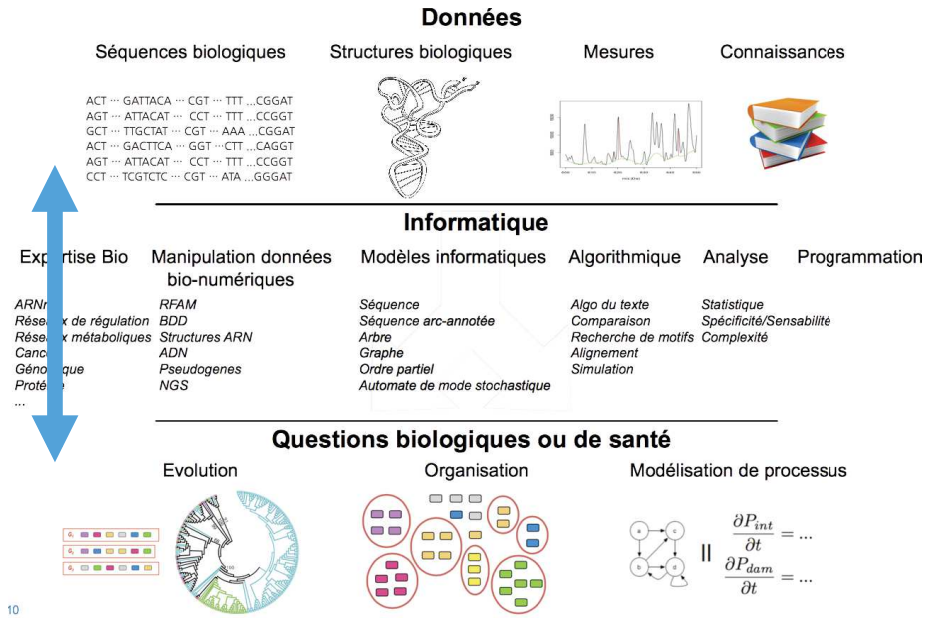


LaBRI

CoBalt : Illustration de travaux

Points saillants

Spectre de compétences élargi



Apport de solutions pratiques et adaptées

L'étude de la composition des données de la cohorte et des patients sains montre une signature chez les malades avec des bactéries connues.

Poursuite de l'étude avec des prélèvements pulmonaires pour conforter l'hypothèse de la migration.

Collaborations

Mise en place de collaborations internationales (Italie et Espagne).

L'ouverture vers de nouvelles questions et collaborations avec l'Hopital Pellegrin.

Valorisation

Ce projet de recherche et développement a été intégré en tant que service de la plateforme CBiB.

Le code est distribué sur un dépôt sourceforge (http://sourceforge.net/projects/taxoassignement)