Université du Québec
Institut National de la Recherche Scientifique
Institut Armand-Frappier

Création d'un outil bio-informatique convivial pour l'étude des changements d'acides aminés au cours de l'évolution des symbiotes bactériens.

Par

Juan Francisco Guerra Maldonado

Mémoire ou thèse présentée pour l'obtention du grade de
Maître ès sciences (M.Sc.)
en Microbiologie Appliquée et Biotechnologie

Jury d'évaluation

Président du jury et examinateur interne	Jonathan Perreault INRS-Institut Armand-Frappier
Examinateur externe	Marc Monot Biomics Institut Pasteur
Directeur de recherche	Frédéric Veyrier INRS-Institut Armand-Frappier

DÉDICACE

À mes parents qui m'ont appris la persévérance et l'amour pour les sciences

À mes sœurs et frères qui m'ont aussi soutenu et encouragé tout au long de mon parcours.

À ma femme et mes enfants qui illuminent mon chemin.

REMERCIEMENTS

Je voudrais remercier mon Directeur de recherche Fréderic Veyrier qui m'a donné l'opportunité de travailler dans ce beau projet. Merci pour m'avoir immergé dans ce monde microscopique où j'ai eu la chance d'appliquer mes connaissances en bioinformatique et développer un nouvel outil au service de la science.

Un grand merci à tous mes amis du laboratoire, dans la vie on n'avance pas tout seul; et l'inspiration et motivation viennent toujours de partager ces expériences quotidiennes, tous les sourires et larmes font partie de l'être humain et nous poussent à nous améliorer.

RÉSUMÉ

Les bactéries sont présentes dans tous les écosystèmes grâce à de nombreux mécanismes adaptatifs dus à de multiples modifications dans leur ADN au cours de l'évolution. Dans certains cas, cette évolution a permis aux bactéries d'établir des liens étroits et permanents avec des organismes eucaryotes et de développer différents types de symbioses. Grâce à l'étude des relations phylogénétiques avec un outil bio-informatique (MycoHIT) créé par notre laboratoire, il a été possible d'identifier des évènements d'insertions ou de délétions gènes qui sont impliquées dans cette transition symbiotique. Cependant, ces résultats suggèrent qu'il doit y avoir d'autres évolutions plus subtiles sous la forme de changements d'acides aminés au niveau des régions conservées des protéines qui affectent leur fonctionnement. Pour répondre à cette problématique, un nouvel outil bio-informatique convivial qui détecte ces changements et fonctionne sur Windows, Linux et Mac OS a été développé. Notre outil, CAPRIB, effectue des analyses comparatives de protéines bactériennes et utilise les langages de programmation Perl, MySQL et Java. Une interface JAVA permet à l'utilisateur le control des opérations. Des scripts Perl extraient et filtrent l'information obtenue par BLAST pour pouvoir l'entreposer dans une base de données SQL. Brièvement, une bactérie de référence est utilisée. Chaque séquence de protéines de cette bactérie sera comparée par tBlastn à chaque génome à étudier. Chaque comparaison de chaque acide aminé pour chacune des protéines sera classifiée et stockée dans un tableau SQL. Java est ensuite utilisée pour gérer les processus d'extraction des comparaisons dans une interface graphique simple pour l'utilisateur. L'idée étant, par exemple, d'extraire les acides aminés conservés chez un groupe de bactéries ayant un phénotype commun mais qui seraient différents chez d'autres bactéries qui n'ont pas ce phénotype. Les changements d'acides aminés seront aussi associés à un score permettant de prédire leur impact sur la fonction des protéines. On présente ainsi une application de CAPRIB dans le projet « Évolution de la forme cellulaire bactérienne dans la famille *Neisseriaceae* ».

ABSTRACT

Bacteria are present in all ecosystems through many adaptive mechanisms due to multiple changes in their DNA during evolution. In some cases, this evolution has allowed the bacteria to establish close and permanent links with eukaryotic organisms and to develop different types of symbiosis. Through the study of phylogenetic relationships with a bioinformatics tool (MycoHIT) created by our laboratory, it was possible to identify events of insertions or deletions of genes that were involved in this symbiotic transition. However, these results suggest that there must be other more subtle changes in the form of amino acid changes in the conserved regions of the proteins that affect their functioning. To answer this problem, a new user-friendly bioinformatics tool that detects these changes and works on Windows, Linux and Mac OS has been developed. Our tool, CAPRIB, performs comparative analyzes of bacterial proteins and uses Perl, MySQL and Java programming languages. A JAVA interface allows the user control of operations. Perl scripts extract and filter information obtained by BLAST so that it can be stored in an SQL database. Briefly, a reference bacterium is used. Each protein sequence of this bacterium will be compared by TBlastN to each genome to be studied. Each comparison of each amino acid for each of the proteins will be classified and stored in an SQL table. Java is then used to manage comparison extraction processes in a simple graphical interface for the user. The idea being, for example, to extract amino acids conserved in a group of bacteria with a common phenotype but which would be different in other bacteria that do not have this phenotype. Amino acid changes will also be associated with a score to predict their impact on protein function. An application of CAPRIB is presented in the project "Evolution of the bacterial cellular form in the Neisseriaceae family".

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

AA	Acides aminés
ADN	Acide désoxyribonucléique
API	Interface de programmation applicative
ARN	Acide ribonucléique
Blast	Basic Local Alignment Search Tool Basic Local Alignment Search Tool
CDD	Conserved Domains Database
DDBJ	DNA Data Bank of Japan
E value	Expect value
ENA	European Nucleotide Archive
EX	Experimental Exchangeability
EX_{dest}	Destination Experimental Exchangeability
EX_{src}	Source Experimental Exchangeability
IR	Inverted repeat sequences
NCBI	National Center for Biotechnology Information
RefSeq	Reference sequence
SI	Séquence d'insertion

.

CHAPITRE 1 : INTRODUCTION

1.1 EVOLUTION BACTERIENNE

L'évolution biologique est un processus qui implique l'apparition et la fixation de nouveaux traits au fil du temps (Bazykin, 2015; Camps *et al.*, 2007). C'est ainsi que la variation est nécessaire pour changer et assurer l'adaptation des organismes à de nouveaux environnements. En ce sens, les bactéries sont présentes dans quasiment tous les écosystèmes terrestre grâce à leur adaptabilité due, entre autre, à ces modifications dans leur ADN au cours de l'évolution (Bang *et al.*, 2018; Hershberg, 2015). L'évolution et la sélection naturelle impliquent deux mécanismes qui sont 1) la variation entre individus et 2) la survie et la sélections des organismes les mieux adaptés (Tadrowski *et al.*, 2018). Cependant, il est à noter que la variation phénotypique des bactéries est causée par la variation génétique mais aussi par la variation adaptative (Rosenberg, 2001; Tadrowski *et al.*, 2018).

1.1.1 *Variation génétique*

La variation génétique appliquée par sélection naturelle est le produit de deux processus, le premier est la modification de l'ADN déjà présent et le deuxième l'incorporation et la recombinaison d'ADN étranger (Sniegowski *et al.*, 2000; Tadrowski *et al.*, 2018). Les mutations se présentent de plusieurs façons ou manières, par exemple comme des erreurs durant la réplication de l'ADN, lésions spontanées et par des éléments transposables (Griffiths, 2000; Maki, 2002). D'un autre côté, les bactéries peuvent aussi échanger horizontalement du matériel génétique par trois grands mécanismes : (1) la transformation, (2) la transduction et (3) la conjugaison (Didelot & Maiden, 2010; Frost *et al.*, 2005; Lin & Kussell, 2017), où le nouvel ADN sera intégré au génome par mécanismes de recombinaison homologue ou non homologue (Thomas & Nielsen, 2005).

1.1.1.1 Mutations spontanées

Erreurs durant la réplication

Le changement d'un nucléotide peut avoir lieu durant la réplication de l'ADN. Ce phénomène est dû aux tautomères des nucléotides qui mènent à un mauvais appariement (Griffiths, 2000). Si le changement s'effectue entre purines (adénine et guanine) ou entre pyrimidines (thymine et cytosine) alors on l'appelle transition. Par contre, si l'on change une purine pour une pyrimidine ou vice-versa alors on l'appelle transversion (Habibi Najafi, 2013)**.**

Lésions spontanées

Les lésions spontanées sont des dommages à l'ADN dans lequel se produisent la dépurination et la désamination. (Griffiths, 2000). La dépurination est observée en cas de rupture de la liaison entre la base et le désoxyribose qui mène à la perte de la guanine ou de l'adénine. (Habibi Najafi, 2013). La désamination de la cytosine produit de l'uracile. Les résidus d'uracile non réparés vont se coupler à l'adénine en réplication, ce qui entrainera la conversion d'une paire G-C en une paire A-T (une transition GC → AT) (Griffiths, 2000; Habibi Najafi, 2013).

Éléments génétiquement transposables

Les transposons ont été initialement détectés comme des éléments génétiques mobiles conférant souvent une résistance aux antibiotiques. Ces éléments sont constitués par des séquences de nucléotides inversement répétées (IR), d'un gène codant pour une protéine permettant l'insertion (transposasse) spécifique aux IR. Quelquefois plus complexes, ils peuvent être aussi composés de gènes qui codent, par exemple, pour une résistance aux agents antimicrobiens ou toxines. Les éléments SI et les transposons sont maintenant regroupés sous le même terme « éléments transposables ». Le processus de déplacement d'un endroit à un autre implique la transposition (insertions d'éléments transposables) et peut ainsi provoquer des mutations (Griffiths, 2000; Rappleye & Roth, 1997).

Résultats des mutations

Les mutations peuvent produire différents effets, par exemple, on pourra avoir un impact dans la transcription ou la transduction de la protéine (Pál & Papp, 2017). La modification d'une séquence de régulation ou une mutation silencieuse peut affecter l'efficacité de la traduction et la stabilité de l'ARNm (Gingold & Pilpel, 2011; Pál *et al.*, 2006). Le répertoire en gènes peut aussi être modulé en fonction des mutations, soit en générant des pseudogènes (par exemple par l'insertion d'une SI) (Abby & Daubin, 2007) ou au contraire en augmentant le nombre de gènes par duplications.

Lors de la transcription, l'ARN sera traduit par groupe de trois nucléotides ou codons. Le fait de changer un nucléotide peut avoir un impact sur le codon à transcrire et ainsi produire différents scénarios sur la protéine qui en résultera: (1) une mutation silencieuse, (2) une mutation faux-sens et (3) une mutation non-sens (Griffiths, 2000; Habibi Najafi, 2013). On aura une mutation silencieuse quand on change le codon, mais que l'acide aminé demeure le même, une mutation faux-sens lorsque l'acide aminé résultant est différent, finalement, une mutation non-sens quand la variation introduit un signal d'arrêt de la traduction, produisant une protéine tronquée (Habibi Najafi, 2013). Il existe aussi les mutations décalantes qui se présentent lorsqu'un ou une paire de nucléotides est inséré ou perdu dans l'ADN à transcrire, affectant le cadre de lecture (Griffiths, 2000).

Les mutations peuvent affecter les protéines de différentes manières. Par exemple, il est connu que le changement d'une lysine par une arginine dans la protéine S12 confère une résistance à la streptomycine dans plusieurs bactéries (Gingold & Pilpel, 2011; Torii *et al.*, 2003). Aussi, la variation de la lysine 11, de l'acide glutamique 146 et de l'acide aspartique 152 de la protéine MinD affectent la formation du dimère avec MinC (Zhou & Lutkenhaus, 2004). Une mutation peut avoir un impact positif dans le gain de fonction de la protéine, mais parfois ce changement occasionne une perte de stabilité (DePristo *et al.*, 2005; Soskine & Tawfik, 2010).

Les mutations se présentent occasionnellement et peuvent être transmises à la cellule fille. Si cette mutation affecte la croissance ou la survie de la cellule, on dit qu'il y a un effet sur le « fitness » (Gordo *et al.*, 2011). De plus, si le fitness décroit en dessous d'un seuil alors la mutation sera, au bout d'un certain nombre de générations, perdue de la population, sinon elle sera retenue (Pál *et al.*, 2006; Reva *et al.*, 2011). Au cours des cycles de réplication des bactéries, de nombreux nouveaux caractères se forment qui pourront être sélectionnés s'ils améliorent le fonctionnement cellulaire (Camps *et al.*, 2007)

1.1.1.2 Transfert horizontal

Les bactéries peuvent gagner de nouvelles informations génétiques grâce aux transferts horizontaux. Il est connu que les transferts horizontaux participent à l'expansion des réseaux métaboliques des bactéries (Pál *et al.*, 2005). Les mécanismes qui permettent les transferts se déclinant en trois grandes catégories : la transformation, la traduction et la conjugaison. Le processus de transfert commence lorsque l'ADN étranger est disponible et finit quand cet ADN est intégré par recombinaison ou en forme extra chromosomique et qu'il fait donc partie du matériel génétique de la cellule receveuse et de sa descendance (Chan *et al.*, 2009; Thomas & Nielsen, 2005).

Transformation

La transformation se présente lorsqu'il y a un transfert d'ADN dans une bactérie receveuse en état de compétence qui est coordonné par 20-50 protéines (Frost *et al.*, 2005; Griffiths, 2000; Thomas & Nielsen, 2005), l'ADN exogène provenant par exemple de la lyse d'une bactérie avoisinante.

Conjugaison

La conjugaison nécessite le contact entre deux cellules avec la formation d'un pont cytoplasmique permettant les échanges bactériens d'ADN plasmidique ou chromosomique. La bactérie donneuse possède la machinerie permettant le transfert (Frost *et al.*, 2005; Griffiths, 2000; Thomas & Nielsen, 2005).

Transduction

La transduction est un processus de transfert de l'ADN par l'intermédiaire d'un phage vecteur (Griffiths, 2000). A basse fréquence, les bactériophages peuvent incorporer des fragments de l'ADN de la cellule hôte et l'injecter après dans un nouvel hôte où il s'intègrera dans le chromosome par recombinaison (Frost *et al.*, 2005).

1.1.2 *Variation adaptative et plasticité phénotypique*

Grâce aux études de Cairns (1988), s'alimente le débat à savoir si les mutations sont le fruit du hasard ou si elles sont influencées par l'environnement. Alors, différentes recherches ont été réalisées (Hersh *et al.*, 2006; Ponder *et al.*, 2005; Rosenberg *et al.*, 1996) et ont permis de voir que les mutations ne sont pas issues d'un processus passif isolé dans la cellule, mais sont le résultat de plusieurs mécanismes faisant intervenir un grand nombre protéines (Nagel, 2007). La variation adaptive est un mécanisme par lequel l'environnement influence la fréquence d'apparition des variations génétiques. Même si l'environnement n'induit pas les changements, il est important pour la variation phénotypique (Thorson *et al.*, 2017). Par exemple, un stress peut réguler la transposition de certains éléments transposables et influencer le taux de mutations. Cette variation a été mise en évidence par Cairns et collaborateurs (Cairns *et al.*, 1988) lors de son expérience avec une souche mutante d'*Escherichia coli* lac^- (incapable de pousser sans ajout de lactose) dans un milieu sélectif non létal et du lactose comme source de carbone. Après quelques jours, des révertants sont apparus. Cairns (1988) a trouvé que quelques révertants apparaissaient dans la phase stationnaire et non dans la phase de croissance et il a conclu que ces mutations étaient induites par l'environnement. Cependant nous savons maintenant que c'est le taux de mutations qui est influencé et augmenté en phase stationnaire.

La plasticité phénotypique est le changement phénotypique sans changement génétique. En fait, il y a des cas où à partir du même génotype il y a plusieurs phénotypes observés, dus aux facteurs environnementaux et à l'histoire de l'organisme (Tadrowski *et al.*, 2018). Des réponses adaptatives rapides peuvent être

expliquées par exemple par différences épigénétiques (méthylation de l'ADN) influençant l'expression de gènes (Thorson *et al.*, 2017). Un autre exemple est démontré par une plus haute fréquence des erreurs dans la synthèse des protéines que celles des mutations génétiques. Les erreurs de lecture dans la transcription et traduction sont de l'ordre 10^{-3} à 10^{-4} tandis que pour les variations génétiques, la magnitude se situe entre 10^{-7} à 10^{-11} (Whitehead *et al.*, 2008; Yanagida *et al.*, 2015). Ces variations nous mènent vers une grande diversité phénotypique, laquelle est très importante dans la survie des bactéries face à différents environnements (Yanagida *et al.*, 2015).

1.1.3 *Adaptations complexes*

Les adaptations complexes sont des phénotypes qui requièrent des mutations multiples et spécifiques conférant un avantage fonctionnel (Kaessmann, 2010; Pál & Papp, 2017). Ces adaptations peuvent avoir lieu dans un gène où les multiples mutations peuvent stabiliser la structure de la protéine et en même temps promouvoir de nouvelles fonctions. Lorsque les adaptations touchent plus d'un gène alors il peut y avoir trois cas. Le premier cas est lorsque les changements entrainent la formation de nouvelles interactions entre molécules (Lynch & Hagner, 2015; Pál & Papp, 2017). Le deuxième cas est lorsqu'il y a la génération de voies métaboliques qui exigent le fonctionnement coordonné de plusieurs gènes en faisant un réseau de protéines (Pál & Papp, 2017; Yamada & Bork, 2009). Finalement, le troisième cas est où il y a formation de macrostructures comme le flagelle ou les canaux ioniques (Pál & Papp, 2017).

Il existe différentes approches pour étudier les adaptations complexes, comme les analyses phylogénétiques, la biologie moléculaire et l'élaboration de systèmes biologiques computationnels (Pál & Papp, 2017). Par exemple, les analyses de données génomiques de différentes espèces ou d'organismes d'une même famille ont montré que le "fitness" dépend de l'évolution de certaines positions dans le génome et que celles-ci changent avec le temps, modelant ainsi l'évolution des protéines (Bazykin, 2015). Dans les études bio-informatiques, la construction de réseaux de protéines met en évidence la relation entre phénotype et génotype pour avoir une vue complète et ainsi prédire quelles trajectoires évolutives particulières sont réalisées (Pál & Papp, 2017).

1.2 ANALYSES COMPARATIVES DE PROTEINES

Pour étudier l'évolution des protéines et leurs impacts phénotypiques, il est nécessaire de connaitre les relations phylogénétiques. Lorsque l'on connait les relations phylogénétiques qui lient plusieurs organismes, on recherche les caractères homologues qui sont partagés par différents organismes et, en mettant en relation ces caractères avec la phylogénie des organismes, on reconstruit l'histoire évolutive (Delsuc *et al.*, 2005).

Les analyses comparatives nous montrent différentes variations clés dans l'évolution et elles permettent entre autres, de faire le lien avec les phénotypes complexes (Abby & Daubin, 2007). Grâce à ces études, la compréhension et le fonctionnement des organismes ont progressé. Par exemple, lors de la comparaison de protéines d'*E. coli* MG1665 avec leurs orthologues dans d'autres souches d'*E. coli*, il a été montré que la pression sélective a modelé la disposition des gènes dans le chromosome bactérien dû à une grande quantité de pseudogènes présents chez *E. coli* MG1665 (Abby & Daubin, 2007). Un autre exemple est la construction de réseaux protéiques où on peut observer comment les protéines interagissent entre elles (Dos Santos Vasconcelos *et al.*, 2018; Noirot & Noirot-Gros, 2004; Papp *et al.*, 2011).

1.3 BASES DE DONNÉES BIOLOGIQUES

Grâce à la réalisation du séquençage de milliers d'organismes, nous pouvons désormais décoder l'évolution du vivant avec une précision sans précédent. Les séquences d'ADN ou de protéines sont entreposées dans des bases de données biologiques comme GenBank, qui contiennent des séquences de nucléotides pour 420 000 espèces (Sayers *et al.*, 2018). Cette base de données a été construite et distribuée par le Centre-américain pour les informations biotechnologiques (NCBI), et relie des informatiques biologiques comme la taxonomie, les génomes, les séquences de protéines, les structures, les domaines conservées et les annotations bibliographiques (PubMed) (Benson *et al.*, 2010; Sayers *et al.*, 2018).

GenBank participe avec les archives européennes de nucléotides (ENA), et la Banque de données génétiques du Japon (DDBJ) en tant que partenaire de la Collaboration internationale en bases de données de séquences de nucléotides INSDC, (Pevsner, 2009). L'uniformité et la disponibilité de l'information sont obtenues grâce aux échanges des données quotidiennement entre les partenaires de l'INSDC (Sayers *et al.*, 2018).

1.3.1 *Numéros d'accession*

Pour accéder efficacement à l'information d'intérêt, chaque enregistrement GenBank, composé à la fois d'une séquence et de ses annotations, est attribué à un identificateur unique appelé numéro d'accession qui est partagé entre les trois bases de données NCBI, ENA et DDBJ (Benson *et al.*, 2010).

1.3.2 *Séquence de référence (RefSeq)*

Le RefSeq nous offre la séquence la plus représentative pour chaque transcrit produit par un gène et pour chaque protéine normale obtenue. On pourra avoir beaucoup de numéros d'accession pour un gène, mais si le produit est le même on aura qu'un RefSeq (Pevsner, 2009). Cette référence nous permet d'être en relation avec d'autres bases de données comme celle de domaines conservés.

1.3.3 *Base de données de domaines conservés (CDD)*

Dans la NCBI, on trouve aussi la CDD. Cette base de données est une source d'annotation protéique constituée d'un ensemble de modèles d'alignement de séquences multiples et elle fournit des informations sur la séquence, la structure et la fonction, ainsi que sur les différents domaines (importés de plusieurs bases de données externes) (Marchler-Bauer *et al.*, 2017)

1.3.4 *Comparaison de séquences*

La comparaison de séquences d'ADN où de protéines nous permet de trouver des domaines ou motifs communs. Ainsi, on peut étudier la relation des protéines dans un organisme ou entre plusieurs organismes (Pevsner, 2009). L'un des premiers et plus importants algorithmes pour l'alignement globale de séquences est celui de Neddleman-Wunch (Needleman & Wunsch, 1970), où on commence par une matrice des séquences à comparer, puis on assigne des valeurs pour chaque changement et finalement on cherche l'alignement optimal. On peut le trouver implémenté en EMBOSS (European Molecular Biology Open Software Suite). Un autre algorithme important est celui de Smith et Waterman (Smith & Waterman, 1981) qui permet de faire un alignement local et il est très utile lorsqu'on veut aligner deux domaines protéiques. Ces algorithmes nous donnent une garantie d'obtenir un alignement optimal, mais leur exécution peut exiger un important temps computationnel : $O(m^2n)$ pour Neddleman-Wunch et de $O(mn)$ pour Smith-Waterman. Les approches heuristiques améliorent la performance, comme c'est le cas de FASTA (Lipman & Pearson, 1985) et BLAST (Altschul *et al.*, 1990). Dans le présent travail, on utilisera BLAST pour faire des alignements et ainsi comparer des séquences et qui génère un rapport avec des informations des séquences comparées (Figure 1).

1.3.4.1 Alignement multiple

En vue de comprendre l'évolution des protéines, l'alignement de deux séquences peut faire ressortir des informations importantes : par exemple, on peut vérifier si ces protéines sont homologues ou non (Pevsner, 2009). Les alignements multiples sont une option plus complète puisque l'assemblage des alignements montre la relation biologique de différentes séquences. L'alignement multiple a beaucoup d'avantages, par exemple, il permet la caractérisation des familles multigéniques, la mise en évidence d'homologie entre les séquences, la prédiction des structures et de construire des phylogénies (Notredame, 2007). Des programmes qui font ces alignements sont ClustalW (Larkin *et al.*, 2007), Muscle (Edgar, 2004) et T-coffee (Notredame *et al.*, 2000). ClustalW et Muscle utilisent une méthode basée en matrice où on calcule les alignements entre paires de séquences considérées, ensuite on prend le meilleur alignement et puis on ajoute progressivement plus de séquences à l'alignement. Pour sa part, T-coffee est une méthode où on ajoute plus d'information dans l'évaluation et ainsi avoir des résultats plus précis (Notredame, 2007).

1.3.4.2 BLAST

BLAST est un outil performant de comparaison de séquences biologiques qui trouve des régions locales de similarité entre séquences. De nos jours, une grande quantité de données biologiques est disponible, mais BLAST ne suffit pas, à lui seul, pour traiter toutes sortes de requêtes liées aux similarités de séquence, de sorte que différentes variantes ont été développées comme BlastX, BlastN, BlastP, TBlastN, TBlastX, PSI_Blast (Tableau I). Chaque variant a ses propres paramètres, algorithmes et critères de performance (Altschul *et al.*, 1990; Kaur *et al.*, 2008).

BLAST est un outil très populaire pour faire des alignements, mais il fait partie des algorithmes heuristiques (Zhang *et al.*, 2000), c'est-à-dire une procédure qui progresse suivant des lignes empiriques pour parvenir à une solution, laquelle n'est pas garantie d'être optimale, mais suffisante pour atteindre nos buts (Brocchieri, 2001).

Tableau I Différentes variantes de Blast

Variant	Description
BlastX	Outil qui sert à comparer une séquence nucléotidique traduite en séquence de protéine contre une base de données de séquences de protéines.
BlastN	Outil qui sert à comparer une séquence nucléotidique contre une base de données de séquences nucléotidiques.
BlastP	Outil qui sert à comparer une séquence de protéine contre une base de données de séquences de protéines.
TBlastN	Outil qui sert à comparer une séquence de protéine contre une base de données de séquences nucléotidiques.
TBlastX	Outil qui sert à comparer une séquence nucléotidique traduite en séquence de protéine contre une base de données de séquences nucléotidiques traduites en séquences de protéines.
PSI-Blast	Outil qui sert à comparer une séquence protéique contre une banque de séquences protéiques. PSI pour « Position Specific Iterated». Cette variante réalise une recherche itérative dans laquelle les séquences trouvées après un cycle de recherche sont utilisées pour construire un score modèle pour le cycle suivant jusqu'à qu'il n'y en ait plus.

BLAST fait la comparaison d'une séquence appelée 'query' avec une autre nommée 'subject' qui sert de base de données. Ainsi, il commence par rechercher tous les mots, ou sous-peptides de longueur k (typiquement 3), existant dans la séquence de la protéine (query). On cherche ces mots (w) dans la séquence de la protéine à l'aide d'une matrice de substitution et chaque mot trouvé est mis dans une liste. Ces mots doivent avoir un score supérieur à un seuil T et doivent être liés au mot d'origine, au cas contraire, ils ne seront pas pris en compte. Puis, les mots et les positions sont gardés dans une table de hachage pour faciliter l'accès. Ensuite chaque match d'un mot de la séquence de la base de données avec un mot de la liste obtenue auparavant sert à commencer un alignement local. L'algorithme étend l'alignement à gauche et à droite de la séquence en calculant le score HSP (High-scoring Sequence Pair) à chaque résidu avec l'aide d'une matrix de substitution. L'algorithme s'arrête quand la valeur HSP obtenu diminue. Les valeurs HSP sont listées en ne retenant que ceux qui sont supérieures à un seuil S et on évalue leur signifiance. Seulement les HSP statistiquement significatifs sont retenus et utilisés pour préparer le rapport.

1.3.4.3 Homologie, similarité et identité

Lors de la comparaison de séquences, les termes comme « l'homologie », « la similarité » et « l'identité » sont très importants. L'homologie, parfois confondue maladroitement à similarité, est une conjecture, une hypothèse à tester (Koonin & Galperin, 2003). L'identité de deux séquences représente les nucléotides ou acides aminés qui ne changent pas, la similarité représente, en plus des résidus qui ne changent pas, une substitution conservative. C'est à dire lorsqu'il y a eu un changement pour un résidu fonctionnellement ou structurellement similaire. Le pourcentage de similarité lors de la comparaison entre deux séquences protéiques sera l'addition entre les acides aminés identiques plus les similaires. Si l'on étudie les acides aminées conservées il peut être plus utile de considérer le pourcentage d'identité, car celui de la similarité est la base de différents critères de relations des acides aminés (Pevsner, 2009). Deux séquences sont homologues si elles ont un pourcentage d'identité suffisant, ce % est arbitraire et peu varié en fonction des protéines comparées et du contexte. L'évidence forte proviendra de la combinaison des analyses structurales et évolutives (Pevsner, 2009).

1.3.4.4 E value

Lors de la comparaison de séquences, il peut y avoir une incertitude dans le fait d'avoir eu une homologie « par chance ». Pour quantifier ce risque, par exemple avec BLAST, on trouve le « Expect value » (E value) (Koonin & Galperin, 2003). E value est la meilleure approche statistique qui décrit le nombre de concordances qui ont un score particulier ou meilleur que celui obtenu par chance (Koonin & Galperin, 2003; Pevsner, 2009).

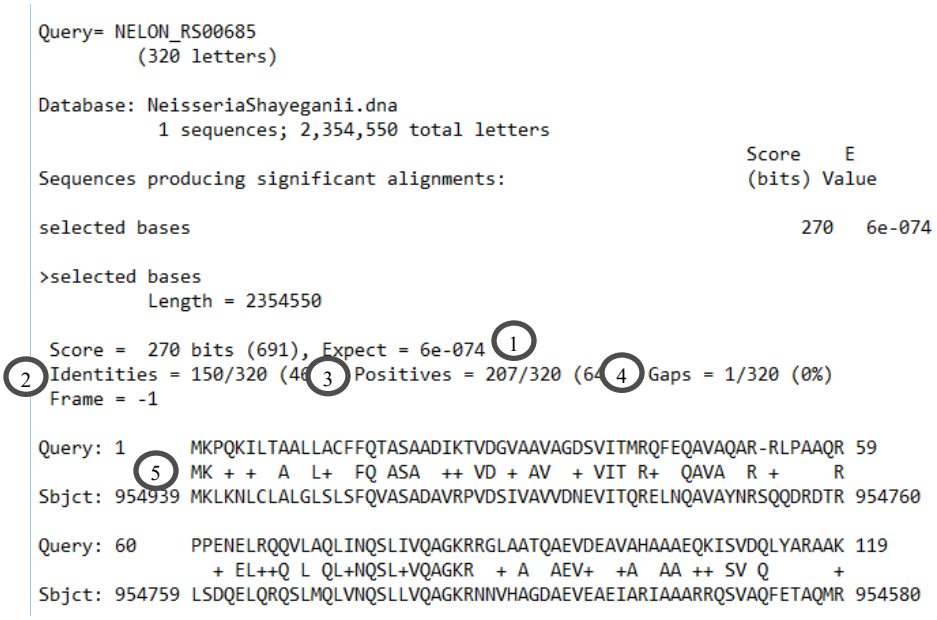


Figure 1 Informations d'un alignement.

Analyse TBlastN de la protéine NELON_RS00685 dans le génome de *N. shayeganii*, on observe quelques informations comme le Expect value (1), le pourcentage d'identité (2), similarité (3), les gaps(4) et l'alignement (5) où un caractère représente une identité, un signe plus est une similarité, un espace est une substitution pour un acide aminé différent et un tiret est un gap ou perte d'un acide aminé.

1.3.5 *Prédiction de l'impact d'un échange d'acides aminés*

Il existe des études qui mesurent l'interchangeabilité des acides aminés comme ceux de Grantham (1974) et Yampolsky (2005). Grantham (1974) prend trois facteurs, à savoir le volume moléculaire, la polarité et la composition. La distance de Grantham (Tableau II) nous montre alors l'effet de changer un acide aminé par un autre; il y aura plus d'impact lorsque la distance augmente. Par exemple, un changement entre une Ile – Leu avec une distance de Grantham de 5 aura moins d'impact qu'un changement entre Cys et Trp qui a une distance de Grantham de 215.

Tableau II Distance de Grantham

Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
110	145	74	58	99	124	56	142	155	144	112	89	68	46	121	65	80	135	177	**Ser**
	102	103	71	112	96	125	97	97	77	180	29	43	86	26	96	54	91	101	**Arg**
		98	92	96	32	138	5	22	36	198	99	113	153	107	172	138	15	61	**Leu**
			38	27	68	42	95	114	110	169	77	76	91	103	108	93	87	147	**Pro**
				58	69	59	89	103	92	149	47	42	65	78	85	65	81	128	**Thr**
					64	60	94	113	112	195	86	91	111	106	126	107	84	148	**Ala**
						109	29	50	55	192	84	96	133	97	152	121	21	88	**Val**
							135	153	147	159	98	87	80	127	94	98	127	184	**Gly**
								21	33	198	94	109	149	102	168	134	10	61	**Ile**
									22	205	100	116	158	102	177	140	28	40	**Phe**
										194	83	99	143	85	160	122	36	37	**Tyr**
											174	154	139	202	154	170	196	215	**Cys**
												24	68	32	81	40	87	115	**His**
													46	53	61	29	101	130	**Gln**
														94	23	42	142	174	**Asn**
															101	56	95	110	**Lys**
																45	160	181	**Asp**
																	126	152	**Glu**
																		67	**Met**

(Grantham, 1974)

Une étude de Yampolsy (2005) propose l'index EX ou son nom en anglais « Experimental Exchangeability » qui mesure l'effet relatif dans l'activité de la protéine (Yampolsky & Stoltzfus, 2005a). L'EX a été mesuré expérimentalement par l'échange de 9671 acides aminés dans 12 protéines. Les valeurs EX (Figure 2) sont d'une grande utilité, car on peut déterminer quel acide aminé est plus facile à être substitué, par exemple la valeur EX_{dest} pour l'alanine est 411 indique que c'est l'acide aminé qui peut substituer mieux les autres, la valeur EX_{src} pour la Lysine est de 409 ce sera le plus facile à substituer, mais sa valeur EX_{dest} égal à 225 montre qu'il substitue aux autres moins facilement. En d'autres termes, avoir l'alanine comme remplacement, aura un impact fonctionnel mineur par rapport à la lysine.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	EX_{src}
C	.	258	121	201	334	288	109	109	270	*383*	258	306	252	169	109	347	89	349	349	139	280
S	373	.	*481*	249	*490*	*418*	*390*	314	343	352	353	363	275	321	270	295	358	334	294	160	351
T	325	*408*	.	164	*402*	332	240	190	212	308	246	299	256	152	198	271	362	273	260	66	287
P	345	*392*	286	.	*454*	*404*	352	254	346	*384*	369	254	231	257	204	258	*421*	339	298	305	335
A	*393*	*384*	312	243	.	*387*	*430*	193	275	320	301	295	225	*549*	245	313	319	305	286	165	312
G	267	304	187	140	369	.	210	188	206	272	235	178	219	197	110	193	208	168	188	173	228
N	234	355	329	275	*400*	*391*	.	208	257	298	248	252	183	236	184	233	233	210	251	120	272
D	285	275	245	220	293	264	201	.	344	263	298	252	208	245	299	236	175	233	227	103	258
E	332	355	292	216	*520*	*407*	258	*533*	.	341	380	279	323	219	*450*	321	351	342	348	145	*363*
Q	*383*	*443*	361	212	*499*	*406*	338	68	*439*	.	*396*	366	354	*504*	*467*	*391*	*603*	*383*	361	159	*386*
H	331	365	205	220	*462*	370	225	141	319	301	.	275	332	315	205	364	255	328	260	72	303
R	225	270	199	145	*459*	251	67	124	250	288	263	.	306	68	139	242	189	213	272	63	259
K	331	376	*476*	252	*600*	*492*	*457*	*465*	272	*441*	362	*440*	.	*414*	*491*	301	*487*	360	343	218	*409*
M	347	353	261	85	357	218	*544*	*392*	287	*394*	278	112	135	.	*612*	*513*	354	330	308	*633*	307
I	362	196	193	145	326	160	172	27	197	191	221	124	121	279	.	*417*	*494*	331	323	73	252
L	366	212	165	146	343	201	162	112	199	250	288	185	171	367	301	.	275	336	295	152	248
V	382	326	*398*	201	*389*	269	108	228	192	280	253	190	197	*562*	*537*	333	.	207	209	286	277
F	176	152	257	112	236	94	136	90	62	216	237	122	85	255	181	296	291	.	332	232	193
Y	142	173	.	194	*402*	357	129	87	176	369	197	340	171	*392*	.	362	.	360	.	303	258
W	137	92	17	66	63	162	.	.	65	61	239	103	54	110	.	177	110	364	281	.	142
EX_{dest}	315	311	293	192	*411*	321	258	225	262	305	290	255	225	314	293	307	305	294	279	172	291

Figure 2 Mesure d'échange EX

Valeur EX(x1000) par point de départ (ligne) et point de destination (colonne)

(Yampolsky & Stoltzfus, 2005a)

1.4 LA FORME CELLULAIRE, UN PHÉNOTYPE COMPLEXE

La variété de formes chez les bactéries (ronde, allongée en bâtonnets, spirale) implique différents mécanismes de croissance et de division cellulaire (Pinho *et al.*, 2013). Elle est dépendante d'une macrostructure composée de sucre et de peptides appelée peptidoglycane (ou paroi). La paroi cellulaire, avec sa résistance face à la pression osmotique et sa flexibilité durant la croissance et la division, nécessite donc un métabolisme et une coordination complexe orchestrée par un système de multiples protéines spécifiques à chaque forme (den Blaauwen, 2018; Jiang *et al.*, 2015; Laddomada *et al.*, 2016; Pinho *et al.*, 2013; Typas & Sourjik, 2015).

1.4.1 *Biosynthèse de peptidoglycanes*

Le peptidoglycane est le composant principal de la paroi cellulaire des procaryotes. C'est un grand polymère composé d'unités de N-acétylglucosamine (GlcNAc) et d'acide N-acétylmuramique (MurNAc). Ce qui varie, c'est l'identité et la fréquence des acides aminés qui y sont ancrés, formant une chaîne tétra peptidique. La machinerie impliquée dans la synthèse du peptidoglycane est l'une des cibles les plus courantes de plusieurs antibiotiques. Par exemple la protéine PBP « Penicillin Binding Protein » qui travaille de concert avec RodA et FtsW dans la synthèse du peptidoglycane, est inhibée par les β-lactames (den Blaauwen, 2018).

Les précurseurs du peptidoglycane sont synthétisés dans trois compartiments cellulaires différents qui sont le cytoplasme, la membrane, et le périplasme (Figure 3). La phase initiale de la biosynthèse commence dans le cytoplasme avec les précurseurs de la paroi cellulaire, l'UDP-N-acétyle glucosamine (UDP-GlcNAc) et l'UDP N-acétylmuramique (UDP-MurNac) pour obtenir l'UDP-N-acétylmuramyl-penta peptide (UDP-MurNAc-pentapeptide) par une série de réactions catalysées séquentiellement par MurA, MurB, MurC, MurD, MurE et MurF. D'autres protéines interviennent dans le cytoplasme comme MreB qui appartient à la machinerie d'élongation et FtsZ qui forme l'anneau contractile dans le site de division cellulaire. Ensuite, dans la membrane interne, l'UDP-MurNAc-penta peptide et l'undecaprenyl phosphate sont transformés par la protéine membranaire MraY en Lipid I. Puis, MurG lie une molécule GlcNAc avec le Lipid I pour obtenir Lipid II qui sera transféré vers le periplasme par des flippases. Finalement dans le periplasme, les PBP incorporent le GlcNAc-MurNac-pentapeptide dans la cape de peptidoglycanes avec des réactions de glycosylation et de transpeptidation (Laddomada *et al.*, 2016; Matteï *et al.*, 2010; Nikolaidis *et al.*, 2014; White *et al.*, 2010). Le peptidoglycane forme un réseau covalent fermé qui assure la survie des cellules. Néanmoins la croissance et la division cellulaire impliquent des insertions et des délétions de molécules de façon dynamique et bien coordonnée avec les machineries d'élongation et de division (Laddomada *et al.*, 2016). Par exemple, l'interaction entre MreB et MurG montre un lien direct entre les protéines cytoplasmiques de la biosynthèse du peptidoglycane et la machinerie d'élongation (Matteï *et al.*, 2010; White *et al.*, 2010).

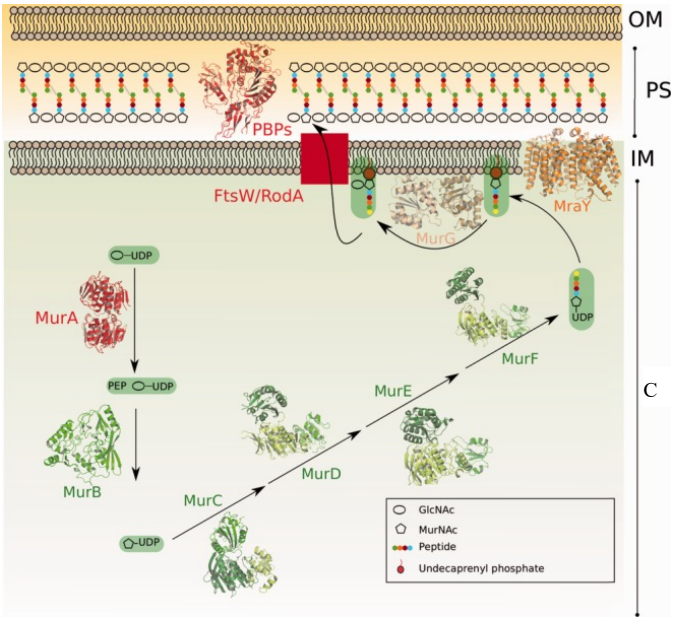


Figure 3 Représentation des étapes de la biosynthèse du peptidoglycane

OM, membrane externe; PS, periplasme, IM, membrane interne; C, cytoplasme, Tiré de (Nikolaidis *et al.*, 2014)

1.4.2 *Machinerie d'élongation*

La machinerie d'élongation (Figure 4**Erreur ! Source du renvoi introuvable.**) contrôle la croissance latérale de la cellule. Parmi les protéines jouant un rôle majeur, il y a MreB qui est un homologue de l'actine et polymérise de manière ATP-dépendante pour former des doubles brins antiparallèles étroitement liés à la membrane interne par le biais de l'extrémité amine de MreB (Typas & Sourjik, 2015). MreB sert d'échafaud et recrute d'autres protéines, comme MreC et MreD, importantes pour former des connexions avec des enzymes qui modifient le peptidoglycane (Pinho *et al.*, 2013; White *et al.*, 2010). La localisation de MreB et la formation de structures est affectées par RodZ, une protéine transmembranaire qui, une fois perdue, mène à la formation de structures non spiralées de MreB et puis à la perte de la forme cellulaire (Bendezú *et al.*, 2009). Beaucoup de bactéries qui n'ont pas MreB montrent une forme ronde et la coordination de l'élongation est dirigée par FtsZ, qui appartient à la machinerie de division, mais aussi par la protéine STK (Typas & Sourjik, 2015).

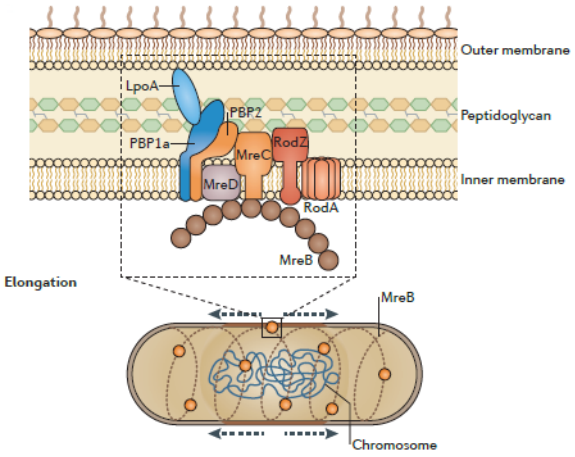


Figure 4 Machinerie d'élongation

Tirée de (Typas & Sourjik, 2015)

1.4.3 *Machinerie de division*

Dans le groupe de protéines impliqué dans la machinerie de division(Figure 5**Erreur ! Source du renvoi introuvable.**), c'est FtsZ qui est la protéine directrice et c'est autour d'elle que d'autres interactions s'établissent (Huang *et al.*, 2013). FtsZ est une protéine homologue de la tubuline, GTP dépendante et forme un anneau dans la moitié des cellules allongées, lequel génère la force contractile nécessaire pour la division cellulaire (Typas & Sourjik, 2015). Une fois que l'anneau est formé, il recrute d'autres protéines pour le complexe de division, comme FtsA et ZipA, qui sont nécessaires pour l'attachement de l'anneau à la membrane (Bernard *et al.*, 2007). FtsA contrôle aussi l'activité de FtsZ et facilite la formation et la constriction de l'anneau tandis que de nombreux facteurs (ZapA–ZapD) aident à sa dynamique de formation (Typas & Sourjik, 2015). Un grand complexe se forme avec l'arrivée d'autres protéines comme FtsEX, FtsK, FtsQ, FtsB, FtsL, FtsW, FtsI, FtsN et des enzymes qui modifient le peptidoglycane comme PBP3 et PBP1b (Bernard *et al.*, 2007; Typas & Sourjik, 2015).

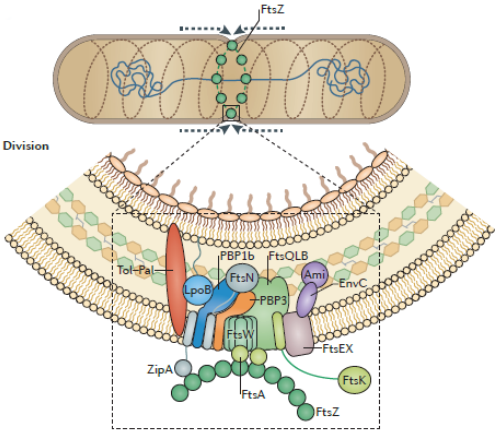


Figure 5 Machinerie de division.

(Typas & Sourjik, 2015)

CHAPITRE 2 : MÉTHODOLOGIE

2.1 PROBLEMATIQUE, HYPOTHESE ET OBJECTIFS

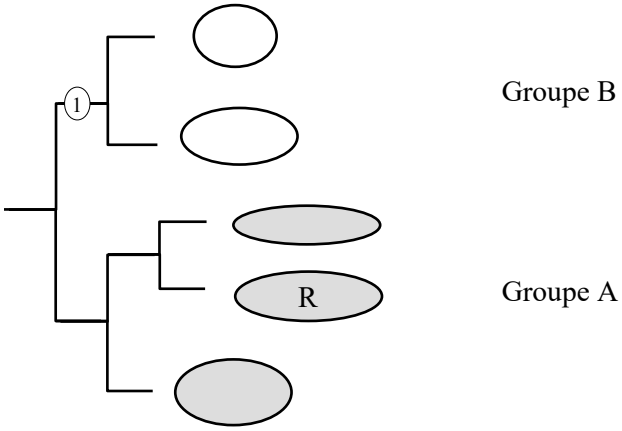
L'adaptabilité des microorganismes n'est plus à démontrer. Dans certains cas, des variations génétiques ont permis aux bactéries de s'adapter et d'établir des liens étroits et permanents (symbiose) avec des organismes eucaryotes (Moya *et al.*, 2008). Différentes relations peuvent avoir lieu, par exemple d'un côté on observe des relations bénéfiques (mutualisme) comme dans le microbiote humain où les bactéries aident l'hôte en favorisant l'homéostasie immunitaire, les réponses immunitaires et la protection contre la colonisation par des agents pathogènes (Pickard *et al.*, 2017). D'un autre côté, dans le cas du parasitisme, des bactéries dites pathogènes causent des dommages chez l'hôte (Ochman & Moran, 2001). La capacité à établir une symbiose avec un hôte des bactéries est en soi un phénotype et son étude dans un contexte évolutif peut nous permettre une meilleure compréhension des phénotypes complexes et des protéines impliquées. Le laboratoire du Dr. Veyrier tente donc de comprendre les évènements évolutifs ayant généré ces symbioses en utilisant différents modèles bactériens.

Dans le passé, le développement d'outils maison comme MycoHIT ont permis de répondre à certaines questions spécifiques au laboratoire et d'identifier des transferts horizontaux dans la genèse du pathogène *Mycobacterium tuberculosis* (Veyrier *et al.*, 2009a). Cet outil a par la suite servi à identifier plusieurs évènements de perte ou gain de gènes (Veyrier *et al.*, 2015a). Par exemple, il a pu être identifié des protéines perdues qui impactent sur le changement de forme des bactéries de la famille *Neisseriaceae* (Veyrier *et al.*, 2015a). Ces études (Veyrier *et al.*, 2009a; Veyrier *et al.*, 2015a), ont montré que si le gain ou la perte d'une protéine peut avoir un impact majeur sur l'apparition d'un phénotype complexe, ce n'est pas suffisant. Nous pensons que certaines autres variations qui n'ont pas été prises en compte comme c'est le cas des changements d'acides aminés au cours de l'évolution qui pourraient avoir un impact majeur. Plusieurs approches pourraient être entreprises pour étudier ces changements. Par exemple, nous pouvons penser aux alignements multiples de séquences de différentes espèces ayant ou non le phénotype, qui permettraient d'étudier l'évolution des protéines (Notredame, 2007). Par contre, cette approche nécessite de connaitre la protéine candidate et donc, étant donné qu'il y a beaucoup d'interactions, la liste de candidats pourrait être grande. De plus, il faut prendre en compte que les protéines peuvent créer de réseaux et il se peut qu'une mutation, même dans un domaine conservé, n'ait pas d'effet dans le phénotype dû à la nature robuste des réseaux protéiques (Typas & Sourjik, 2015). C'est ainsi qu'est née la nécessité de développer un programme qui teste toutes les protéines reportées d'un organisme pour prédire et extraire certains changements d'acides aminés dans certaines protéines qui pourraient avoir un lien avec l'apparition d'un phénotype donné.

L'hypothèse de départ se construit alors dans ce contexte évolutif:

« Soit T un arbre phylogénétique représentatif d'organismes génétiquement proches et « 1 » un nœud qui appartient à T, mais avec la particularité de représenter un évènement évolutif où il y a un changement de phénotype (Figure 6). Si le changement du phénotype est dû à un changement d'acide aminé et que cette variation est présente dans les organismes avec le nouveau phénotype, mais absente dans les autres, alors on pourra l'extraire et la prédire »

Figure 6 Apparition d'un phénotype au nœud 1.



R, Génome de référence; A, Groupe de référence; B groupe avec changement de phénotype; 1, nœud correspondant à l'évènement évolutif distinguant A et B.

L'objectif principal est alors de développer un outil bio-informatique convivial pour détecter les changements d'acides aminés sélectionnés au cours de l'évolution et qui peuvent affecter la fonction des protéines communes. Pour accompagner notre programme d'un exemple concret, on utilisera les l'exemple des organismes comparés lors de l'étude de la forme cellulaire dans la famille *Neisseriaceae* réalisée par Veyrier et al. (2015) et on appliquera une stratégie évolutive. Dans cet exemple, l'ancêtre commun à certaines bactéries (dont les deux espèces pathogènes *N. meningitidis* et *N. gonorrhoeae*) a subi un changement de forme; de la forme bacille à coque, ce qui correspondra à notre phénotype.

2.2 STRATÉGIE ÉVOLUTIVE

Dans le but d'extraire les variations qui auraient pu jouer un rôle dans le changement du phénotype lors d'un évènement évolutif, on extrairera les acides aminés dans une protéine qui sont conservés dans les organismes avec un phénotype mais absent chez ceux n'ayant pas le phénotype. Il existe plusieurs exemples dans la littérature de l'impact de tels changements. Par exemple, comme celui qui est présenté dans (Reva *et al.*, 2011), où la mutation en la protéine RAC1 change une alanine conservée dans la sous-famille #1 par un glutamate dans la sous-famille #2. Ce changement a eu un grand impact dans l'interface de liaison avec la protéine Tiam1 et il a changé le phénotype.

La première étape est d'effectuer des comparaisons par BLAST et on cherchera les acides aminés qui ont changé suite à cet évènement, mais pour ce faire il est nécessaire d'apporter quelques définitions basiques pour développer l'algorithme de travail.

Le fait d'avoir un évènement évolutif à l'origine d'un phénotype crée deux groupes, ici notés comme A et B. Soit le génome de **R**, un organisme de **R**éférence qui sera comparé avec les autres organismes par TBlastN, où on fera l'alignement entre toutes les protéines de R et chaque génome. Le groupe où se trouve R sera le groupe A (Figure 6).

Après avoir comparé les séquences protéiques de l'organisme de référence R avec les séquences génomiques des organismes des groupes A et B par TBlastN, on filtrera et gardera cette information dans une base de données pour finalement faire les opérations suivantes :

- IvsD : on cherche les acides aminés identiques dans le groupe A, mais différents dans le groupe B.
- ISvsD : on cherche les acides aminés identiques et similaires dans le groupe A, mais différents dans le groupe B
- IvsS : on cherche les acides aminés identiques dans le groupe A, mais similaires dans le groupe B.
- Gap In : On cherche les acides aminés perdus dans le groupe A pour une délétion.
- Gap Out : On cherche les acides aminés perdus dans le groupe B pour une insertion.
- Stop Codon : on cherche les codons stop apparus dans le groupe B.

Notre stratégie d'étude nécessitera des opérations d'ensembles et certaines bases pour leur implémentation. Un ensemble est une collection d'objets. Si l'ensemble ne contient pas d'éléments, on dit que c'est un ensemble vide et on le note par Ø. Si l'ensemble contient tous les éléments d'un domaine étudié, c'est l'ensemble de référence U. Le nombre d'éléments d'un ensemble est la cardinalité de l'ensemble. Les propriétés les plus importantes pour le présent travail sont l'associative, la commutative, la distributive et la neutralité. Entre les opérations, on trouve l'union et l'intersection. L'union de deux ensembles A et B est l'ensemble qui contient tous les éléments qui appartient à A et B et on la représente par $A \cup B$. L'intersection des ensembles A et B est l'ensemble qui contient seulement des éléments communs à A et B et on la note par $A \cap B$ (Johnsonbaugh, 2000) .

Tableau III Operations et propriétés d'ensembles

Propriété	Union	Intersection
Associative	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
Commutative	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Distributive	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
Neutralité	$A \cup Ø = A$ $A \cup U = U$	$A \cap U = A$ $A \cap Ø = Ø$

Donc, l'application des ensembles pour chaque combinaison dans notre programme est définie ci-dessous: soit des ensembles de positions d'acides aminées identiques (*K*), similaires (*L*), différents (*M*), position de départ d'un gap dans l'organisme de référence (*N*), position de départ d'un gap dans l'organisme comparé (*O*) et la position d'un codon d'arrêt (*P*); les opérations d'ensembles seront :

- IvsD : $K_A \cap M_B$
- ISvsD : $(K_A \cup L_A) \cap M_B$
- IvsS : $K_A \cap L_B$
- Gap In : $N_A \cap M_B$
- Gap Out : $(K_A \cup L_A \cup M_A) \cap O_B$
- Stop Codon: $(K_A \cup L_A \cup M_A) \cap P_B$

2.3 DONNÉS À L'ÉTUDE

Pour développer le programme, on a utilisé les séquences en format fasta des organismes de la famille *Neisseriaceae* du site du NCBI (Tableau IV). Le phénotype étudié a été le changement de forme lors d'un évènement évolutif présenté dans le nœud 1 (Figure 7). On a aussi utilisé différents organismes de référence, car on fait des comparaisons par blast, et on voulait établir le meilleur choix pour la référence. Les organismes utilisés pour construire les bases de données ont été *N. elongata* avec 2100 protéines*, K. kingae* avec 1833 protéines*, N. bacilliformis* avec 2235 protéines et *S. alvi* avec 2216 protéines.

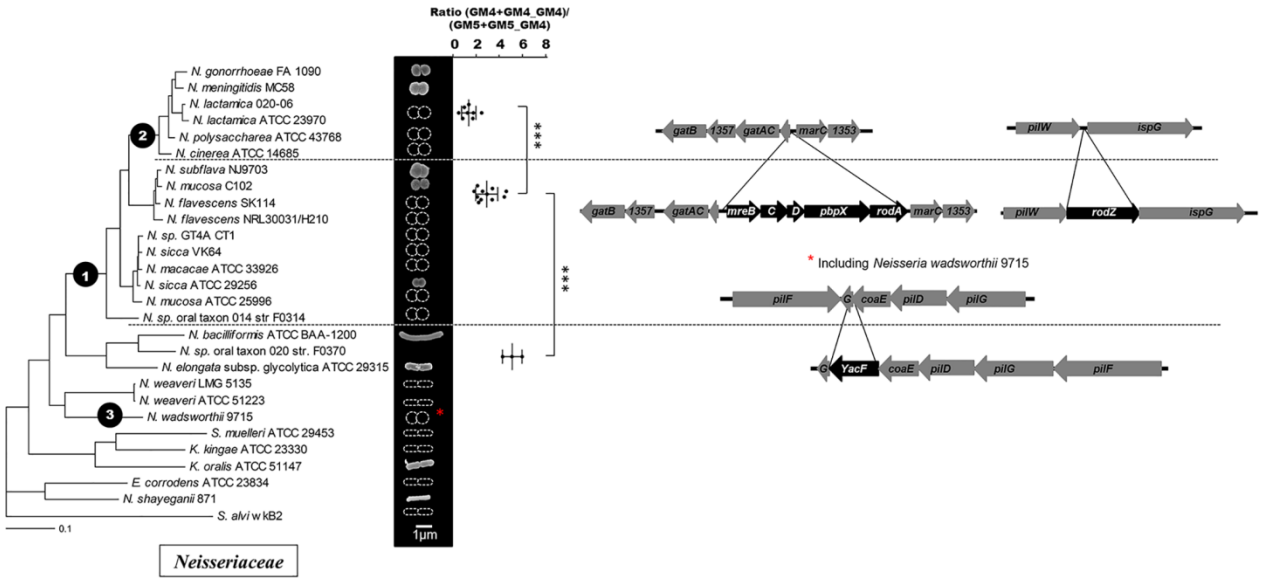


Figure 7 Arbre phylogénétique de la famille *Neisseriaceae*.

Un évènement évolutif au niveau du nœud 1 qui démarque le changement de forme cellulaire, qui montre la perte de du YacF ou ZapD appartenant à la machinerie de division cellulaire et un autre au niveau du nœud 2 où il y a une grande perte de gènes impliqués dans la machinerie d'élongation.

(Veyrier *et al.*, 2015a)

Tableau IV Organismes de la famille *Neisseriaceae* utilisés pour le développement de l'outil CAPRIB.

Groupe	Organisme	Assemblage
Coques	*Neisseria gonorrhoeae* FA 1090	ASM684v1
	Neisseria meningitidis MC58	ASM880v1
	Neisseria Lactamica 020-06	ASM19629v1
	Neisseria Lactamica ATCC 23970	ASM74196v1
	Neisseria polysacharea ATCC 43768	ASM17673v1
	Neisseria cinerea ATCC 14685	ASM17389v1
	Neisseria subflava NJ9703	ASM17395v1
	Neisseria mucosa C102	Neis_muco_C102_V1
	Neisseria flavescens SK114	ASM17527v1
	Neisseria flavescens NRL30031/H210	ASM17393v1
	Neisseria sp. GT4A CT1	Neisseria_sp_GT4A_CT1_V1
	Neisseria sicca VK64	NsiccaVK64v1.0
	Neisseria macacae ATCC 33926	ASM22086v1
	Neisseria sicca ATCC 29256	ASM17465v1
	Neisseria mucosa ATCC 26256	ASM17387v1
	Neisseria sp. Oral taxon 014 str F014	ASM9087v1
Bacilles	*Neisseria bacilliformis* ATCC BAA-1200	ASM19492v1
	Neisseria bacilliformis 914 NLAC	ASM106763v1
	Neisseria sp. Oral taxon 020 str F0370	ASM31823v2
	Neisseria elongata subsp. glycolytica ATCC 29315	ASM81803v1
	Neisseria weaveri LMG 5135	ASM22425v1
	Neisseria weaveri ATCC 51223	ASM22427v2
	Kingella kingae ATCC 23330	ASM21353v1
	Kingella oralis ATCC 51147	ASM16043v1
	Kingella denitrificans ATCC 33394	ASM19069v1
	Neisseria shayeganii 871	ASM22687v1
	Snodgrassella alvi	ASM60000v1

2.4 FLUX DU TRAVAIL

Le programme qui a été développé doit faire des comparaisons par BLAST, extraire, stocker et gérer les informations des rapports obtenus dans une base de données, produire un rapport avec les mutations trouvées et interagir avec la base de données de domaines conservés de la NCBI si l'utilisateur le souhaite (Figure 8). Un fichier d'aide est fourni en Annexe 2.

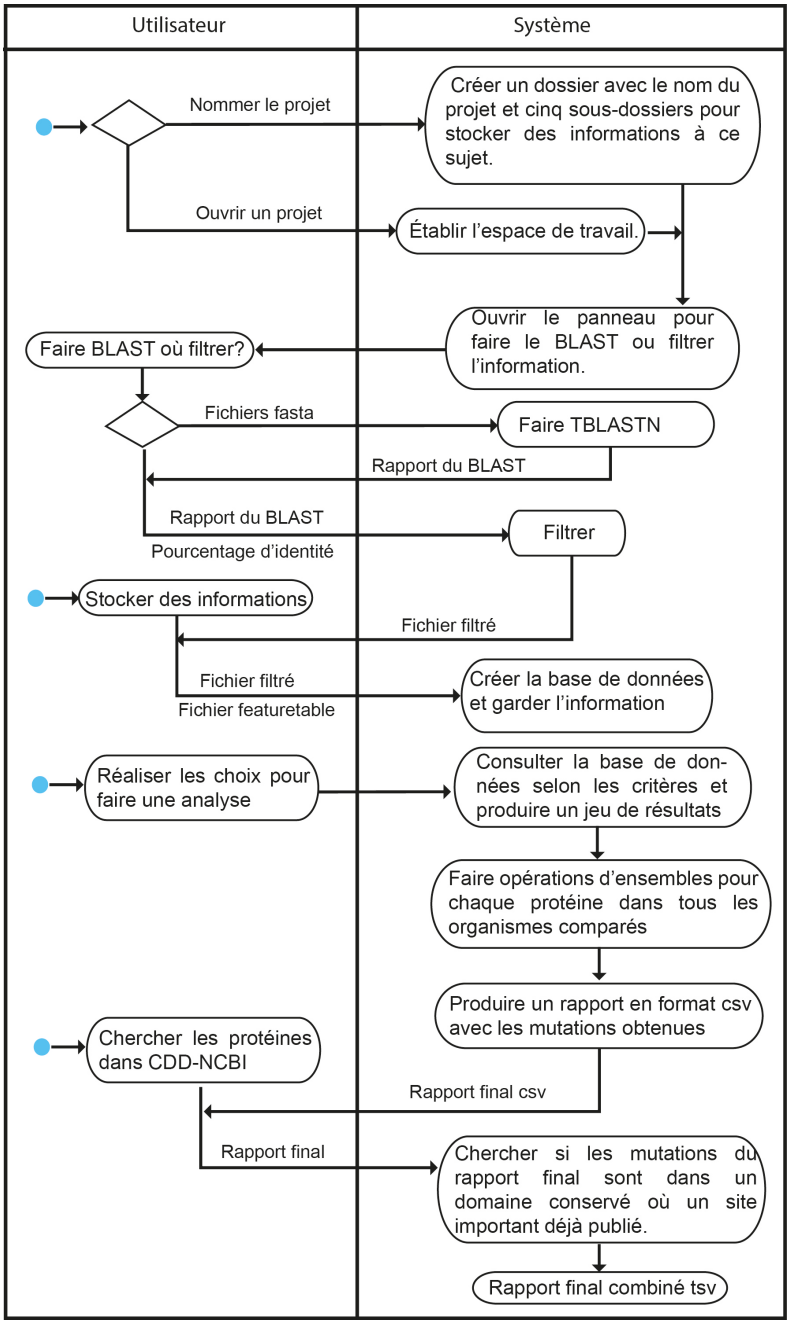


Figure 8 Diagramme d'activité du programme.

2.4.1 *Implémentation*

Le programme est développé en langage Perl, SQL, bash et Java. Perl est utilisé pour filtrer les rapports BLAST, mais aussi pour communiquer avec la CDD-NCBI et combiner ce résultat avec les fichiers de protéines candidates et, finalement, obtenir le rapport en format TSV. SQL nous sert à construire, gérer et avoir accès à la base de données, bash nous permet d'entretenir le dossier 'externalPrograms' où sont placés les scripts en Perl. Finalement, Java nous permet de gérer le tout grâce à différentes classes avec différentes fonctions et qui sont rangés en trois modules. Ainsi le premier module 'control' aura des classes à jouer qui feront des calculs, le deuxième 'modele' aura par exemple des classes qui interagissent avec la base de données et dans le troisième module 'vue' on aura les classes de l'interface graphique. Toutes ces classes sont dans le fichier Caprib.jar et on peut y accéder en décompressant ce fichier.

2.4.2 *Commencer un projet*

Pour commencer, l'utilisateur doit avoir un espace pour garder les fichiers fasta, les rapports du BLAST, les fichiers filtrés et les résultats. Pour ce faire, on crée la classe Project qui générera un dossier avec le nom du projet et quatre sous-dossiers : Fasta, Blast, Filtered et Results. Le premier servira à garder les fichiers fasta, le suivant pour les rapports BLAST, le troisième les fichiers pour alimenter la base de données et le dernier pour les rapports finaux avec les mutations. Il y aura aussi un message d'alerte lorsque l'usager veut créer un projet déjà existant ou bien s'il laisse le champ du nom vide. Finalement, il existe aussi l'option pour travailler sur un projet préalablement créé.

2.4.3 *TBLASTN*

Pour réaliser le TBLASTN, on utilise les commandes **makeblastdb** et **tblastn** de BLAST+ version 2.9.0+ (https://www.ncbi.nlm.nih.gov/books/NBK279690/). La première sert à construire la base de données de BLAST avec la séquence d'ADN de l'organisme à comparer en format FASTA. La deuxième commande sert à comparer les séquences des protéines de l'organisme de référence avec la base de données qu'on a créé précédemment. A noter que pour des raisons de compatibilité avec MycoHIT mais aussi pour des raisons de simplifications des résultats, le système crée un fichier fasta des protéines (combine.fasta) où l'on change la valeur du Refseq par le locustag. Ces opérations seront mises dans la classe PrepareBlastFiles qui devra identifier le système opératif et lancer un message une fois le BLAST fini.

2.4.4 *Filtrer*

Cette étape est très importante car il s'agit d'extraire les informations du rapport du BLAST. On l'a développé en langage Perl dans un but futur de faire une application combinée MycoHIT/CAPRIB. Pour extraire les informations, on utilise des expressions régulières qui identifieront la protéine, la valeur E-value, le pourcentage d'identité (%I), le pourcentage de similarité (%S), la longueur, la position des acides aminés identiques, la position des acides aminés similaires, la position des acides aminés différents, la position où commencent les gaps et la position d'un codon d'arrêt (Figure 10). L'usager a le choix d'établir un seuil minimal de %I et ainsi extraire les informations des protéines qui aient une valeur égale ou plus grande que celui-ci. Cependant, pour des petites bases de données il est préférable de mettre 0. Toute l'information sera mise dans la table SQL et l'utilisateur aura ensuite le choix de choisir un seuil lors des calculs. Cette information est extraite dans un fichier de sortie en format csv.

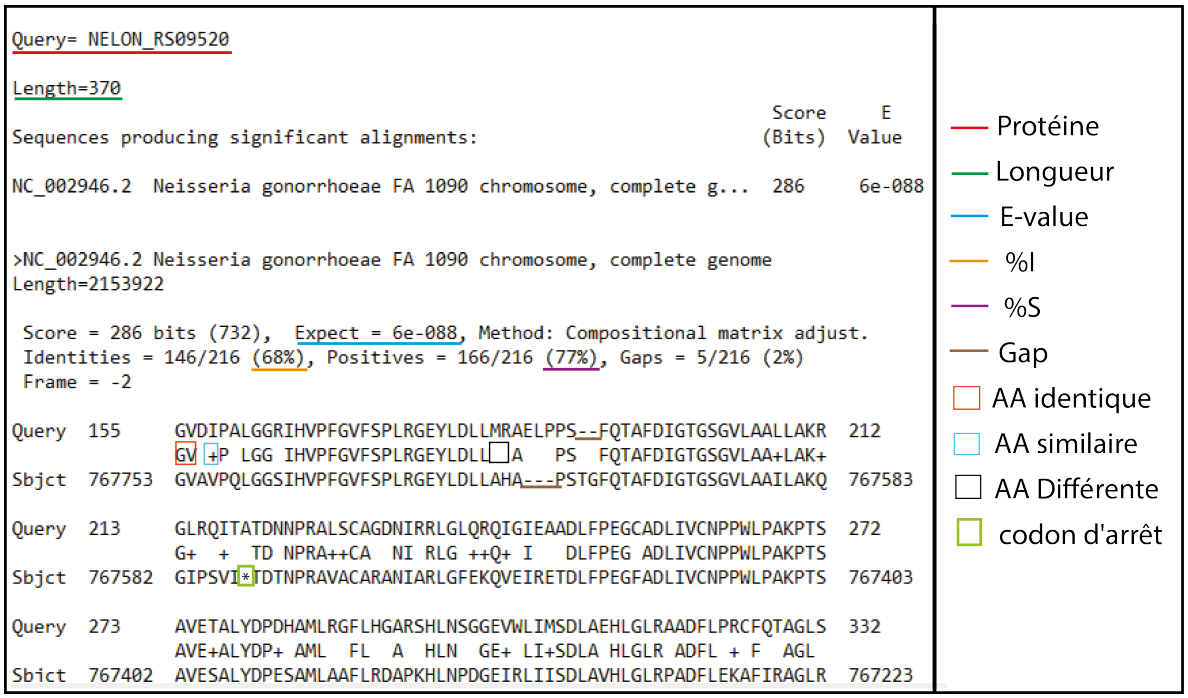


Figure 9 Informations à extraire d'un rapport BLAST

2.4.5 *Création de la base de données*

L'outil réalisera une base de données relationnelle pour stocker l'information filtrée. On utilisera le programme XAMPP version 7.3.0-0 qui est en libre accès. L'outil utilisera trois tableaux. Le premier pour garder des informations du fichier featuretable.txt de l'organisme de référence (comme les fonctions des

protéines, COG …), un autre pour garder les informations filtrées du BLAST et un dernier pour avoir un registre de chaque comparaison (Figure 10). Étant donné qu'il s'agit d'un travail de communication entre Java et XAMPP, l'outil vérifie d'abord la connexion et, en cas d'échec, un message d'alerte avertira à l'usager. Ensuite, l'utilisateur nomme la base de données avec un mot sans espaces et unique. Le système vérifie la validité du nom. Le pas suivant est de remplir les autres deux tableaux, c'est à dire d'introduire les fichiers filtrés et de faire la dernière vérification de sa validité. A noter, que l'on a laissé la possibilité d'effacer la base de données ou d'effacer un registre.

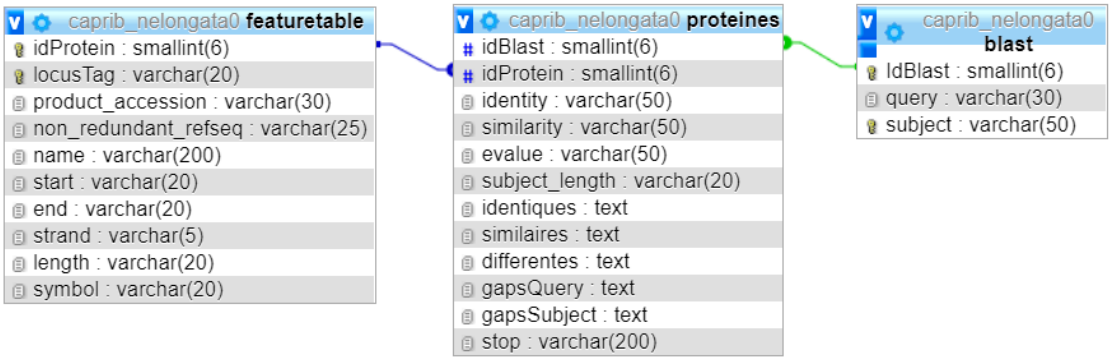


Figure 10 Diagramme de la base de données relationnelle

2.4.6 *Trouver les mutations*

La recherche des mutations qui peuvent avoir un rôle dans le changement d'un phénotype est faite par comparaison de l'information filtrée et gardée dans la base de données. C'est un processus complexe et on le découpe ici en cinq étapes pour une meilleure compréhension. La première étape commence comme dans la création de la base de données. L'outil doit vérifier la connexion, consulter quelles bases de données existant dans l'ordinateur et ensuite charger celle qui intéresse l'utilisateur.

La deuxième étape démarre avec la base de données choisie, l'utilisateur fait la requête des organismes qui sont dans le tableau blast (Figure 10). L'utilisateur peut maintenant classer les organismes en deux groupes selon le phénotype, groupe A et groupe B où les organismes qui partagent le même phénotype que la référence sont classés dans le groupe A.

La troisième étape consiste à établir un seuil de pourcentage d'identité. A noter, que l'utilisateur avait déjà filtré les résultats de blast auparavant. Il se peut que l'utilisateur souhaite changer ce seuil (>seuil initial). Nous avons donc rajouté une option ici pour ne pas perdre du temps à faire une autre base de données.

La quatrième étape est d'extraire de la base de données, l'information des organismes classés qui respectent le seuil fixé avec la classe QueryProteinsDB qui retourne un objet ResultSet. On obtiendra un jeu de résultats de tous les blast réalisés pour les organismes choisis. On travaillera donc seulement avec les protéines qui sont présentes dans tous les organismes.

Finalement, on utilise les résultats de cette requête pour réaliser des opérations d'ensembles entre les deux groupes. On a programmé six combinaisons entre les deux groupes, ou on cherche pour chaque protéine conservée les positions des AA communs dans le groupe A qui ont changé dans le groupe B (Tableau V). La classe qui fait ces opérations s'appelle Calculs et utilise une liste pour les organismes du groupe A et une autre pour le groupe B et la liste des protéines conservées au seuil fixé. En fonction de la combinaison choisie par l'utilisateur l'outil fait l'opération d'ensembles. Par exemple, si on choisit IvsD l'outil sélectionnera l'ensemble des positions des acides aminés identiques d'une protéine pour les organismes du groupe A, et fera une opération d'intersection pour voir quelles positions sont communes pour cette protéine dans tous les organismes A. L'outil fait la même opération pour cette protéine dans les organismes du groupe B mais cette fois avec la position des acides aminés différents. Finalement, on réalise l'intersection entre les deux ensembles obtenus qui correspond à la (ou les) position des changements (non-similaires) d'AA entre les bactéries du groupe A et B. L'outil produit un rapport en format CSV.

Tableau V Combinaisons entre les groupes A et B à réaliser en chaque protéine conservée.

Combinaison	Description
I vs D	Position des acides aminés identiques dans le groupe A qui ont changé pour un acide aminé différent dans le groupe B.
IS vs D	Position des acides aminés identiques ou similaires dans le groupe A qui ont changé pour un acide aminé différent dans le groupe B.
I vs S	Position des acides aminés identiques dans le groupe A qui ont changé pour un acide aminé similaire dans le groupe B.
GapIn	Position à laquelle commence un gap dans une protéine du groupe A.
GapOut	Position à laquelle commence un gap dans une protéine du groupe B.
Stop codon	Position à laquelle un acide aminé du group A a été changé par un codon d'arrêt dans le groupe B

2.4.7 *Création du rapport*

La partie la plus énergivore par rapport au temps d'exécution, est celle d'obtenir le rapport final car il doit contenir des informations en relation avec les organismes, les protéines et les mutations comme on décrit au Tableau VI ci-dessous.

Tableau VI Informations qui doivent être présentes dans le rapport final

Descripteur	Description
Organismes	• Organismes comparés et classés dans les groupes A et B
Protéines	• Quantité de protéines communes au seuil fixé. • Nombre de protéines candidates. • Locustag. • RefSeq. • Symbol (si rapporté au featuretable.txt). • Nom.
Mutations	• Nombre de mutations trouvées selon l'analyse • Positions des mutations • Acides aminés qui ont changé pour chaque position • Compter les mutations qui présentent le même changement pour chaque organisme. • Distance de Grantham • Valeur EX

Le plan à suivre est de créer la méthode pour produire un rapport .csv dans la classe Calculs où l'on trouve les organismes qui ont servi pour faire la requête avec QueryProteinsDB, l'information des protéines dans le ResultSet et la liste de protéines candidates avec ses mutations. Les distances de Grantham et valeur EX seront ajoutées dans un objet de la classe Subject qui est construit pour chaque organisme avec les informations du ResultSet. L'outil calcule et ajoute la valeur des distances pour chaque mutation (seulement pour les AA candidats) pour alléger le temps d'exécution (Figure 11).

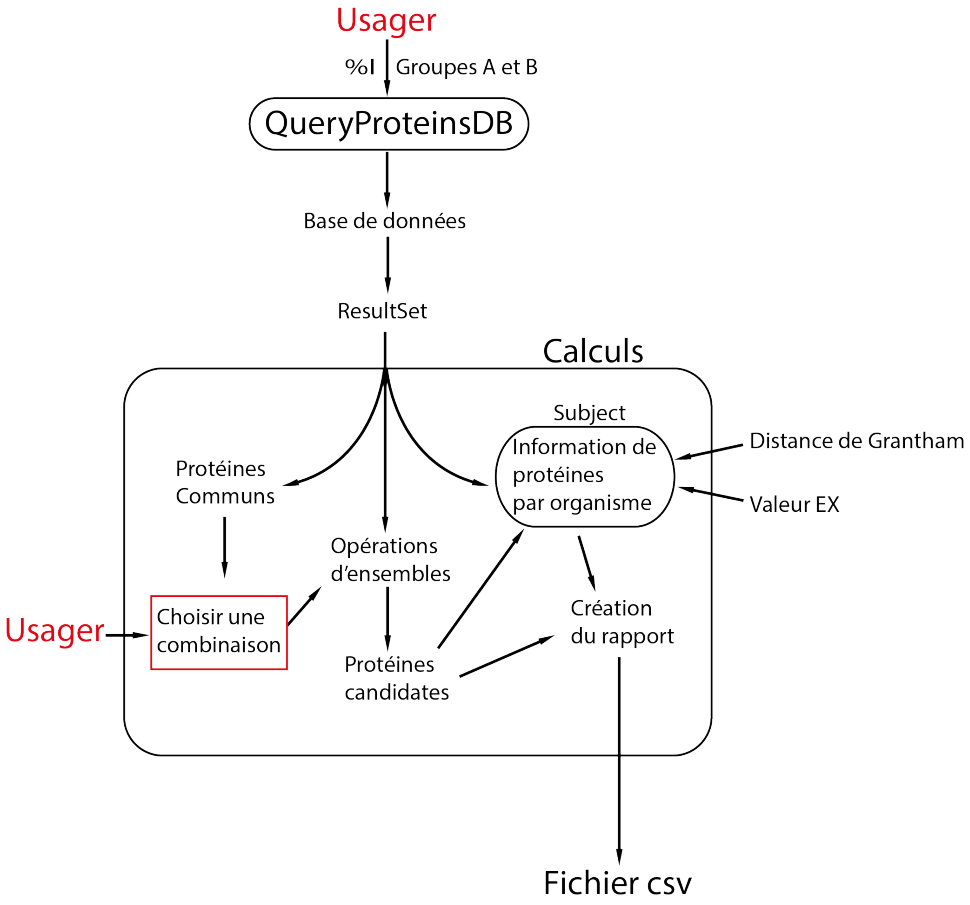


Figure 11 Obtention du fichier CSV.

Le processus commence par la requête de l'usager qui choisit un seuil de %I et les organismes qu'il souhaite comparer. Le jeu de données obtenues dans un objet ResultSet sert à faire des calculs comme obtenir les protéines communes, puis l'usager choisit une combinaison (Tableau V), ensuite le système fait les opérations d'ensembles pour obtenir un dictionnaire des protéines avec les positions des mutations trouvées. Après, un objet Subject est créé pour les protéines candidates et les organismes du groupe B, on récupère l'information du ResultSet et on ajoute les distances. Finalement, on compte les mutations et on imprime le fichier CSV.

2.4.9 *Communiquer avec la base de données de domaines conservés*

Une fois trouvée les protéines avec les mutations candidates suggérées, nous avons pensé qu'il serait utile de communiquer avec la base de données de domaines conservés de la NCBI (Marchler-Bauer *et al.*, 2017; Marchler-Bauer & Bryant, 2004). Ceci est en option. L'outil compare la position des AA de la liste des candidats et compare avec les domaines conservés. On obtient les domaines et des positions importantes, et cette opération peut se faire directement dans le site de la NCBI. L'avantage de procéder avec notre programme est qu'on vérifie si les positions obtenues appartiennent à un domaine conservé ou si la mutation touche un acide aminé important déjà rapporté. Pour le faire on utilise le code fournit par le site de la CDD NCBI https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml, on l'utilise dans la classe CddNcbi qui lance le script cdd.pl, récupère les résultats et les combine avec le fichier de protéines candidates .csv obtenu dans le numéral antérieur pour produire un fichier combiné en format TSV.

2.4.10 *Interface graphique*

On a utilisé les packages javax.swing et java.awt présents dans Java et avec lesquels on peut construire les éléments nécessaires pour l'interaction de l'usager. Le menu contient quatre sous-menus à savoir File, Database, Operations et Help. Le premier sert à créer un nouveau projet ou à ouvrir un projet déjà existant et lancer le panel pour faire une analyse BLAST ou Filtrer. Le deuxième aide l'utilisateur à créer et alimenter la base de données. Le troisième sous-menu sert à faire les combinaisons et communiquer avec la CDD NCBI et finalement, le dernier présente un tutoriel en format html.

Pour les opérations BLAST et filtrer, on utilise un objet JPanel, pour la base de données et pour les opérations, on utilise un objet JTabbedPane avec des listes pour avoir une première vue des résultats obtenus. Il a donc été nécessaire de programmer des classes accessoires qui nous aident à gérer tous les éléments souhaités.

t

CHAPITRE 3 RESULTATS

3.1 ARTICLE

Cet article a été soumis au journal « Bioinformatics » sous forme de note (application note) et était limité à deux pages. Malheureusement, il a été refusé. Entre la première soumission et les corrections du mémoire nous avons travaillé sur une nouvelle version qui est présentée en annexe.

CAPRI-B: A USER-FRIENDLY TOOL TO STUDY MUTATIONS IN PROTEINS DURING EMERGENCE OF A PHENOTYPE IN A BACTERIAL GENUS.

Juan F. Guerra Maldonado[1], Frederic J. Veyrier[1*]

[1] INRS-Institut Armand-Frappier, Bacterial Symbionts Evolution, Laval, Quebec, Canada

* Corresponding author : frederic.veyrier@iaf.inrs.ca

ABSTRACT

Summary: Capri-B is an easy-to-use bioinformatics tool to study evolutionary events in bacterial proteome through a graphical interface in java. It functions by comparing the protein sequences inside a given genus and generates a list of amino acid variations in the core proteome that could have occurred concomitant with the emergence of a phenotype. The output is a list of amino acid changes that correlate with an user-defined phenotype (such as, but not restricted to, cell-shape, antibiotic resistance, virulence), associated with indicators of their potential impacts. These indicators are scores of such as GRANTHAM substitution matrix score and EX value or their potential localisation in conserved domains (CDD).

Availability and Implementation: The software and documentation (tutorial and requirements) are available in https://github.com/BactSymEvol/Caprib. It is implemented in Java, Perl and SQL. Supported on Unix-like operating systems and MS Windows.

Contact: frederic.veyrier@iaf.inrs.ca

Supplementary information*:* Material included with the program.

1 Introduction:

Evolution and selection allowed bacteria to adapt to different ecological niches. Several mechanisms participated to this adaptation such as gene deletions, gene duplications or horizontal gene transfer among others. Multiple tools and strategies have been designed in the past to detect such event in the context of bacterial evolution (Veyrier *et al.*, 2009b). Nevertheless, while these genetic events were important, they were not sufficient to completely explain bacterial adaptation. In most of the cases, this strongly implies a need for secondary events to allow this evolution, independently of gene deletion/insertion, such as more subtle modifications like single amino-acid changes. These variations in the protein sequence may affect the function, the stability and the interactions with biomolecules, such as DNA, RNA, lipids or other proteins (Notredame, 2007; Reva *et al.*, 2011; Ventura *et al.*, 2007). The effect of these changes on the bacterial fitness will determine its selection or elimination for future generations (Gordo *et al.*, 2011).

If these changes are easy to detect in relatively closely related bacteria (intra-species, intra-bacterial complex) (Habibi Najafi, 2013), it is however, not the case for more ancestral events (such as changes intra-genus or intra-family). Identifying these ancestral amino acid changes could potentially reveal new genetic determinants regulating bacterial evolution. In order to detect these mutations, one approach could be to use bioinformatics tools such as ClustalW, Muscle and MAFFT (Notredame, 2007) that use multiple sequence alignment algorithms to find the biological relationship between several sequences. However, these programs require a targeted strategy implying the identification of protein candidates, while, it is expected that major changes in bacterial evolution will require reworking of the entire proteome (Typas & Sourjik, 2015). Multiple tools exist for large-scale protein comparisons but none has been developed - to the extent of our knowledge - that are based on a phylogenetic comparative sorting process to pinpoint amino-acids changes that may explain evolutionary modifications responsible for phenotypic changes at the

bacterial genus or family level. We have designed a tool, called Capri-B (Comparative Analyses of Proteins In Bacteria, Capri-B), that is able to perform such searches in protein sequences that are common to all bacteria in the database (herein a genus). Capri-B, that uses a JAVA interface for more convenience, permits to extract a list of amino acid changes that are correlated with the emergence of a given phenotype and it suggests those that could have the more drastic impacts. These changes may affect protein function, the nature of protein-protein interactions or gene regulators among others.

2 Materiel and methods:

With the phylogeny and genome sequence of bacteria inside a given genus, it was possible to compare genes content and detect insertion and deletion events at a specific node where a given phenotype has emerged. Our group has already designed a tool (MycoHIT) and established a methodology to identify these events (Radomski *et al.*, 2013; Veyrier *et al.*, 2009b; Veyrier *et al.*, 2015b; Veyrier *et al.*, 2011; Wang *et al.*, 2015). Although the present program, called Capri-B, can be applied in multiple contexts, we will illustrate it using the example of the coccoid transition, from an ancestral bacilli cell-shape, inside the *Neisseriaceae* family (Veyrier *et al.*, 2015b). We have described the evolutionary events that led to changes in cell shape in the ancestor of coccus *Neisseria* such as *N. meningitidis* or *N. gonorrhoeae*. We have established that this cell shape difference was the result of an event at a specific evolutionary node (node 1 in Fig 1). In our efforts to understand the evolutionary events that were implicated in the cell shape transition for these cocci, we have identified a specific genetic deletion (Veyrier *et al.*, 2015b). Using gene deletion in a rod-shaped ancestor *N. elongata*, we showed the importance of the YacF (renamed ZapD) in the coordination of the cell elongation and division (Veyrier *et al.*, 2015b). Nevertheless, this deletion

was not sufficient to explain the complete transition and we hypothesized that change in amino acids could have participated in this transition.

2.1 Defining the evolutionary Strategy

Briefly, our approach consists of comparing two groups of species separated by an evolutionary event that represents a change in phenotype (herein bacilli versus cocci). Using a reference dataset (for example protein sequences from *N. elongata*), we apply the tblastn tool (Altschul *et al.*, 1990) on all the genomes from the database (herein *Neisseria* species). The information of the blast is filtered and stored in a relational SQL database (such as, but not limited to, Amino acids (AA) identical with the references, AA different, AA similar, gaps, stop codons). To apply this evolutionary strategy, Capri-B perform this whole process using a JAVA interface connected with Blast +, Perl, SQL and the NCBI database of conserved domains (CDD NCBI). The first step is the tblastn, using the input files: fasta protein file (.faa), and the features file of the reference organism (.ptt that is provided by genbank), and the fasta containing the genomic sequences of the query organisms (.fna). With the obtained output reports, Perl is used to automatically parse and extract some information that will be save in the SQL database. This operation needs to be repeated for all query organisms.

2.2 Operations in Capri-B

Once the database is build, it is now possible to retrieve the global amino-acid changes that strictly correlated with the emergence of the phenotype (herein cell shape evolution) using different parsing strategies. The user classifies organisms in function of their phenotype (with group A being the organisms that present the same phenotype than the reference). In our example, bacilli are in group A (as is the reference *N. elongata*) and cocci in group B. We then dispose of several choices to combine the groups, such as IvsD, ISvsD, IvsS, GapIn, GapOut and StopCodon. As a example,

IvsD (identical vs different) will allow the extraction of position of amino acids that are identical in group A but that have changed for a different amino acid in group B. GapIn, GapOut and StopCodon, will extract position of gaps in group A, position of gaps in group B, stop codon positions in group B. The program is also able to compare this list with other databases (such as NCBI CDD Conserved Domains Database, which contains the annotation of sequences with the location of conserved domain footprints, and functional sites (Marchler-Bauer *et al.*, 2017; Marchler-Bauer & Bryant, 2004)) to extract the changes that are predicted to have the most potent functional impacts. Each Amino acid change is also associated with the GRANTHAM substitution matrix score (Grantham, 1974) and EX value (Yampolsky & Stoltzfus, 2005b) that could indicate how severe the variant will affect the protein.

3 Example use Case

We provide example files in our package that include the reference (*N. elongata*) and few genomes. The user can test the software by comparing protein sequences between cocci (herein *N. meningitidis* and *N. gonorrhoeae*) and bacilli (herein *Kingella kingae* and *Snodgrassella alvi*) in the *Neisseriaceae* family as presented in figure 1. The users need to be aware that more genomes are necessary to obtain relevant results.

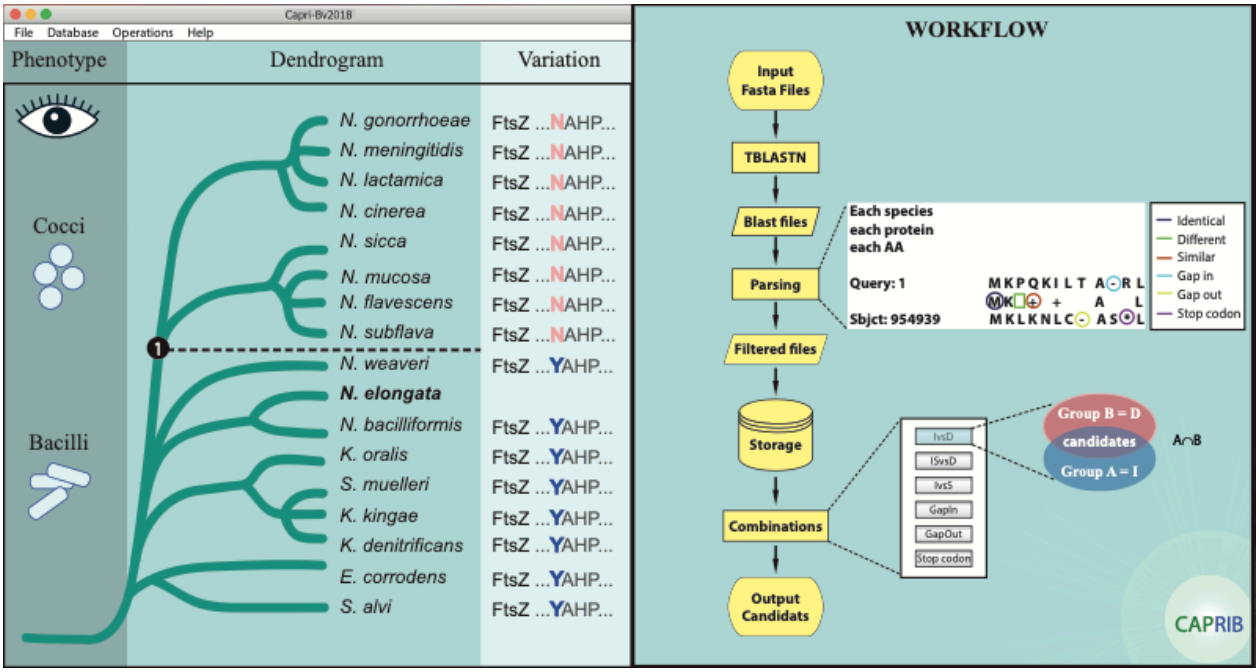
4 Conclusion

Capri-B is a license-free new user-friendly tool permitting comparison of proteins inside a genus and that allow prediction of amino acid changes that could impact protein function concomitant to the emergence of a given phenotype in this genus.

References

1. F. Veyrier, D. Pletzer, C. Turenne, M. A. Behr, Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis. *BMC Evol Biol* **9**, 196 (2009).
2. B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research* **39**, e118-e118 (2011).
3. M. Ventura *et al.*, Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum. *Microbiology and Molecular Biology Reviews : MMBR* **71**, 495-548 (2007).
4. C. Notredame, Recent Evolutions of Multiple Sequence Alignment Algorithms. *PLoS Computational Biology* **3**, e123 (2007).
5. I. Gordo, L. Perfeito, A. Sousa, Fitness Effects of Mutations in Bacteria. *Journal of Molecular Microbiology and Biotechnology* **21**, 20-35 (2011).
6. M. B. Habibi Najafi, *Bacterial Mutation; Types, Mechanisms and Mutant Detection Methods: a Review*. (2013), vol. 4, pp. 1857-7431.
7. A. Typas, V. Sourjik, Bacterial protein networks: properties and functions. *Nature Reviews Microbiology* **13**, 559 (2015).
8. N. Radomski *et al.*, atpE gene as a new useful specific molecular target to quantify Mycobacterium in environmental samples. *BMC Microbiol* **13**, 277 (2013).
9. F. J. Veyrier *et al.*, Common Cell Shape Evolution of Two Nasopharyngeal Pathogens. *PLoS Genet* **11**, e1005338 (2015).
10. F. J. Veyrier, A. Dufort, M. A. Behr, The rise and fall of the Mycobacterium tuberculosis genome. *Trends Microbiol* **19**, 156-161 (2011).
11. J. Wang *et al.*, Insights on the emergence of Mycobacterium tuberculosis from the analysis of Mycobacterium kansasii. *Genome biology and evolution* **7**, 856-870 (2015).
12. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990).
13. A. Marchler-Bauer, S. H. Bryant, CD-Search: protein domain annotations on the fly. *Nucleic Acids Research* **32**, W327-W331 (2004).
14. A. Marchler-Bauer *et al.*, CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research* **45**, D200-D203 (2017).
15. R. Grantham, Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-864 (1974).
16. L. Y. Yampolsky, A. Stoltzfus, The exchangeability of amino acids in proteins. *Genetics* **170**, 1459-1472 (2005).

Figure 1: Capri-B: from the phenotype to the amino-acids candidates. A) Example of the implementation of the evolutionary strategy. The user determines the phenotype and the node of evolution of this phenotype in the phylogenetic tree (herein cocci and bacilli). The user can then determine the two groups to compare by tblastn. (B) Schematic representation of the workflow for Capri-B analyses.



3.2 CAPRIB

On a développé CAPRIB qui nous permet de faire des analyses comparatives entre protéines conservées pour suggérer des changements d'acides aminés qui auraient pu avoir un rôle à jouer dans le changement d'un phénotype. Notre programme est composé par les dossiers 'externalPrograms', 'project' et 'tutorial', le fichier exécutable Caprib2019.jar ainsi que le fichier 'Requirements-Caprib.doc' avec l'information nécessaire pour le fonctionnement de notre programme (Figure 12). Dans le dossier 'externalPrograms' on a des scripts auxiliaires en langage perl et bash, le dossier 'project' a un sous-dossier pour chaque projet où l'on gère les fichiers fasta, blast, filtrées et les résultats des analyses. Finalement, le dossier 'tutorial' contient des fichiers en format html avec les informations du fonctionnement de CAPRIB.

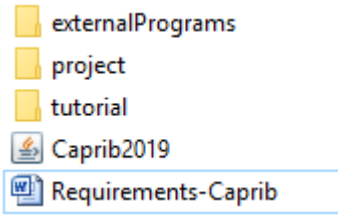


Figure 12 Elements du dossier Caprib

Si l'on exécute le fichier Caprib2019.jar le panneau principal s'ouvre et on y trouve un menu avec les options File, Database, Operations et Help (Figure 13) qui nous aideront à gérer les différentes opérations de CAPRIB présentées dans la Figure 8.

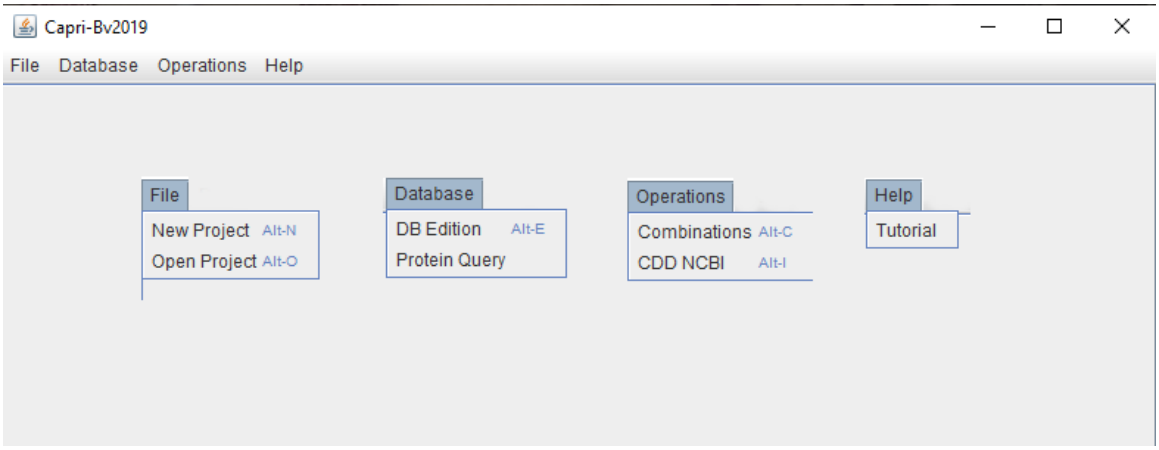


Figure 13 Interface graphique de CAPRIB.

Dans l'interface on trouve quatre sous-menus chacun avec ses options

3.2.1 *File*

Cette option nous permet de commencer le travail, et si l'on choisit 'New Project' une fenêtre s'ouvre pour introduire le nom du projet. L'usager ne peut pas laisser l'espace vide ou introduire un nom déjà existant ou un message d'alerte sortira pour le prévenir (Figure 15).

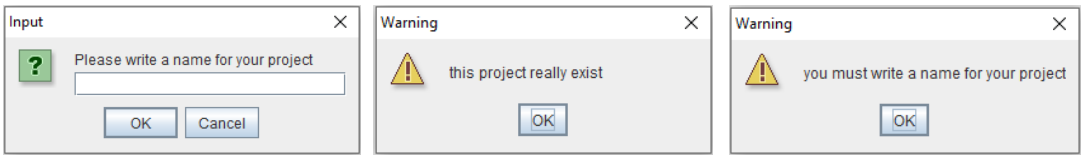


Figure 14 Création d'un nouveau projet.

Une fois le nom de l'espace de travail rempli, un fichier est créé et une fenêtre pour faire BLAST ou filtrer ses rapports émerge. Par exemple, si on nomme le projet 'NelongataTutorial', alors CAPRIB crée ce dossier et quatre sous dossiers comme on observe dans la Figure 16.

Figure 15 Création du dossier du projet.

Une fenêtre apparait (Figure 17) avec deux sections, dans la première nommée 'Prepare files', l'utilisateur a l'option de faire les analyses BLAST en introduisant des fichiers fasta de protéines et 'feature table' pour l'organisme de référence et le fichier fasta d'ADN pour l'organisme à comparer. La deuxième section 'Filter Blast' sert à filtrer les rapports BLAST par rapport au pourcentage d'identité'. Une fois que les informations sont entrées par l'usager alors il exécute avec le bouton 'Blast' ou 'Filter' selon le cas.

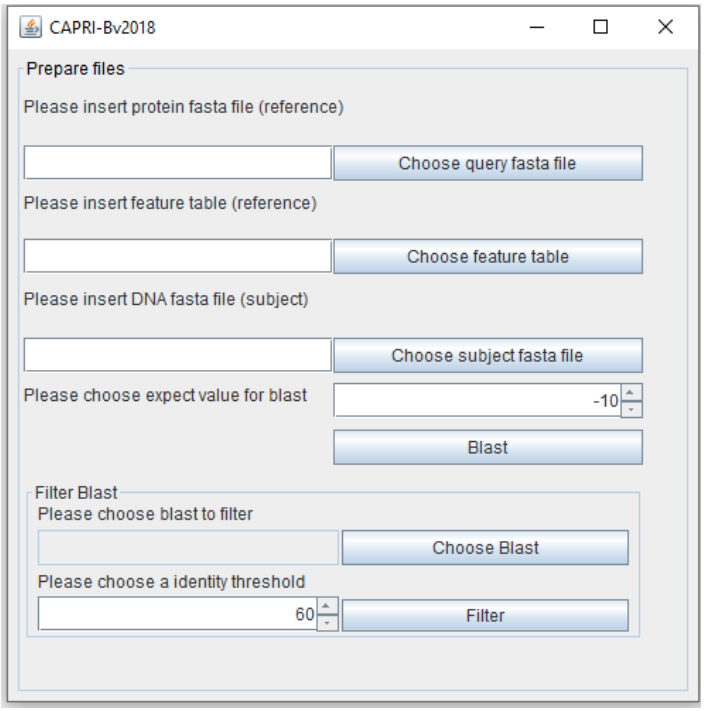


Figure 16 Panneau pour faire les opérations BLAST et filtration par rapport au pourcentage d'identité.

3.2.2 *Database*

Ce menu nous permet deux choix (Figure 13), le premier est 'DB Edition' qui sert à créer une base de données avec les fichiers filtrés. Le deuxième est 'Protein Query' qui nous permet de vérifier si une protéine est présente ou pas dans les organismes. Il faut préciser que pour toutes les opérations qui concernent la communication avec la base de données, CAPRIB établit d'abord si la connexion existe, le cas échéant un message d'avertissement apparait et ferme le programme (Figure 18).

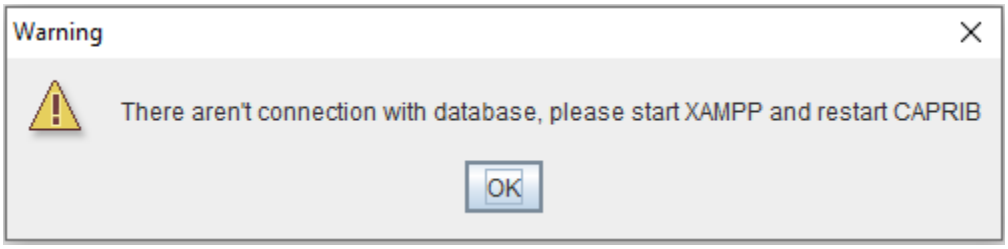


Figure 17 Vérification de la connexion entre CAPRIB et la base de données.

L'option 'DB Edition' ouvre un panneau 'Database' qui a trois sections (Figure 19), la première sert à créer la base de données, la deuxième à introduire les fichiers filtrés et la troisième pour effacer la base de données ou les données d'un organisme.

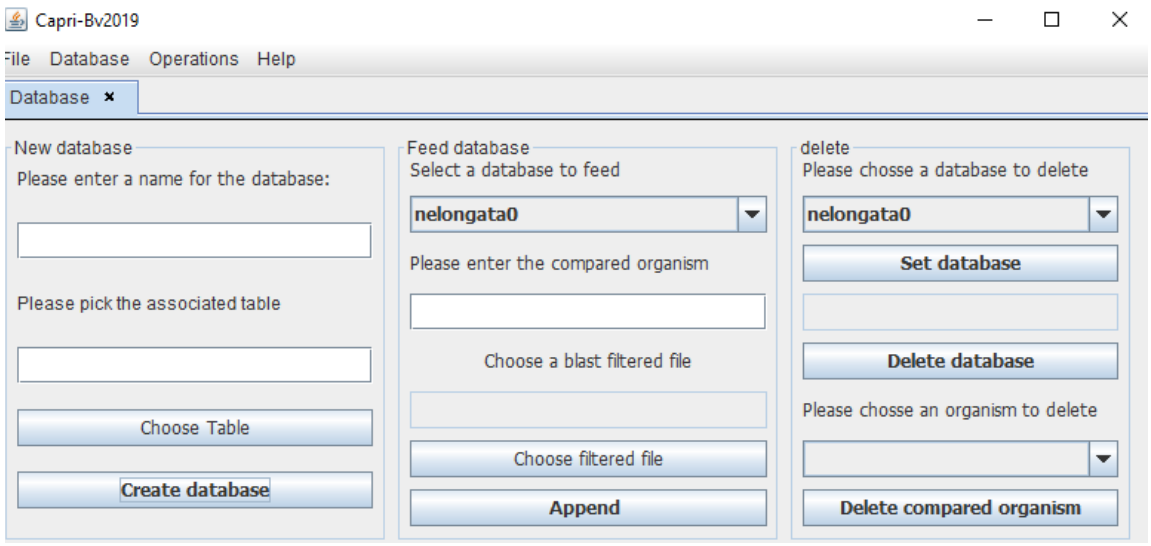


Figure 18 Panneau d'édition de la base de données.

Le nom de la base de données ne doit pas contenir d'espaces ou le caractère '_' réservé par CAPRIB dans l'identification des bases de données issues du même, néanmoins si l'usager le fait alors un message apparait (Figure 20).

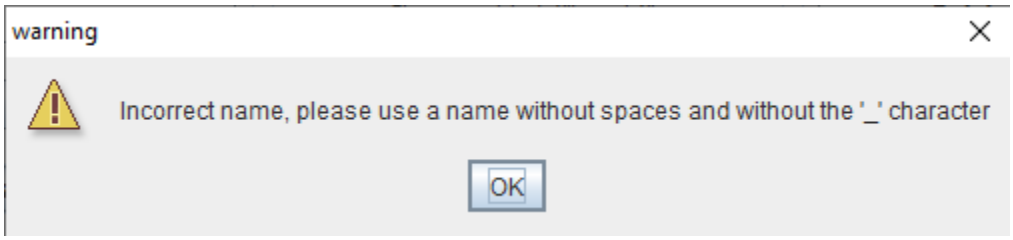
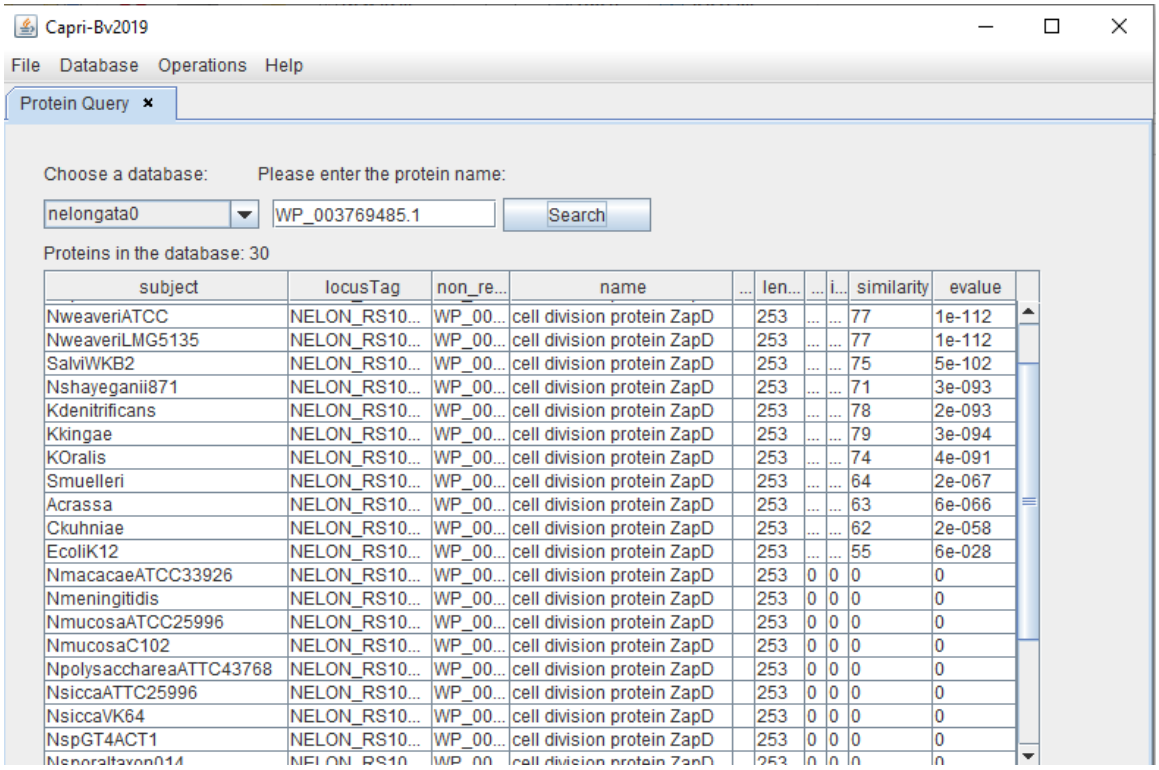


Figure 19 Message d'erreur pour l'entrée d'un nom non valide.

L'option 'Query Proteins' cherche dans une base de données une protéine à partir du locustag ou du refseq. Par exemple dans la Figure 21, on a cherché dans la base de données 'nelongata0' la protéine 'WP_0036985.1' et CAPRIB nous montre que dans les organismes comparés de la famille *Neisseriaceae,* cette protéine est présente dans les bacilles mais absente dans les coques corroborant ainsi les résultats de Mycohit en (Veyrier *et al.*, 2015a).



subject	locusTag	non_re...	name	...	len...	...	i...	similarity	evalue
NweaveriATCC	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	77	1e-112
NweaveriLMG5135	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	77	1e-112
SalviWKB2	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	75	5e-102
Nshayeganii871	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	71	3e-093
Kdenitrificans	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	78	2e-093
Kkingae	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	79	3e-094
KOralis	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	74	4e-091
Smuelleri	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	64	2e-067
Acrassa	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	63	6e-066
Ckuhniae	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	62	2e-058
EcoliK12	NELON_RS10...	WP_00...	cell division protein ZapD		253	...	...	55	6e-028
NmacacaeATCC33926	NELON_RS10...	WP_00...	cell division protein ZapD		253	0	0	0	0
Nmeningitidis	NELON_RS10...	WP_00...	cell division protein ZapD		253	0	0	0	0
NmucosaATCC25996	NELON_RS10...	WP_00...	cell division protein ZapD		253	0	0	0	0
NmucosaC102	NELON_RS10...	WP_00...	cell division protein ZapD		253	0	0	0	0
NpolysaccharеaATTC43768	NELON_RS10...	WP_00...	cell division protein ZapD		253	0	0	0	0
NsiccaATTC25996	NELON_RS10...	WP_00...	cell division protein ZapD		253	0	0	0	0
NsiccaVK64	NELON_RS10...	WP_00...	cell division protein ZapD		253	0	0	0	0
NspGT4ACT1	NELON_RS10...	WP_00...	cell division protein ZapD		253	0	0	0	0

Figure 20 Vérification de la protéine ZapD dans la famille *Neisseriaceae* avec 'Protein Query'.

3.2.3 *Operations*

Ce menu a deux options à savoir « Operations » et « CDD NCBI » (Figure 13). La première réalise les opérations d'ensembles pour obtenir les mutations dans les protéines conservées et générer le rapport en format CSV (Annexe 3). Par exemple dans la Figure 22 on charge les organismes de la base de données « nelongata0 » et puis on commence à classer les organismes selon le groupe auquel ils appartiennent. Les organismes qui partagent le phénotype avec l'organisme de référence sont au group A et les autres sont au groupe B. On peut aussi introduire le pourcentage d'identité voulu et lancer la requête avec le bouton « Query ». Lorsque CAPRIB finit ce processus, il apparait un message (Figure 22) et l'usager peut choisir une combinaison avec les boutons IvsD, ISvsD, IvsS, Gaps In, Gaps Out et Stop codon. CAPRIB produit un fichier en format CSV mais aussi une prévisualisation des résultats dans l'onglet 'Variations Table' (Figure 23).

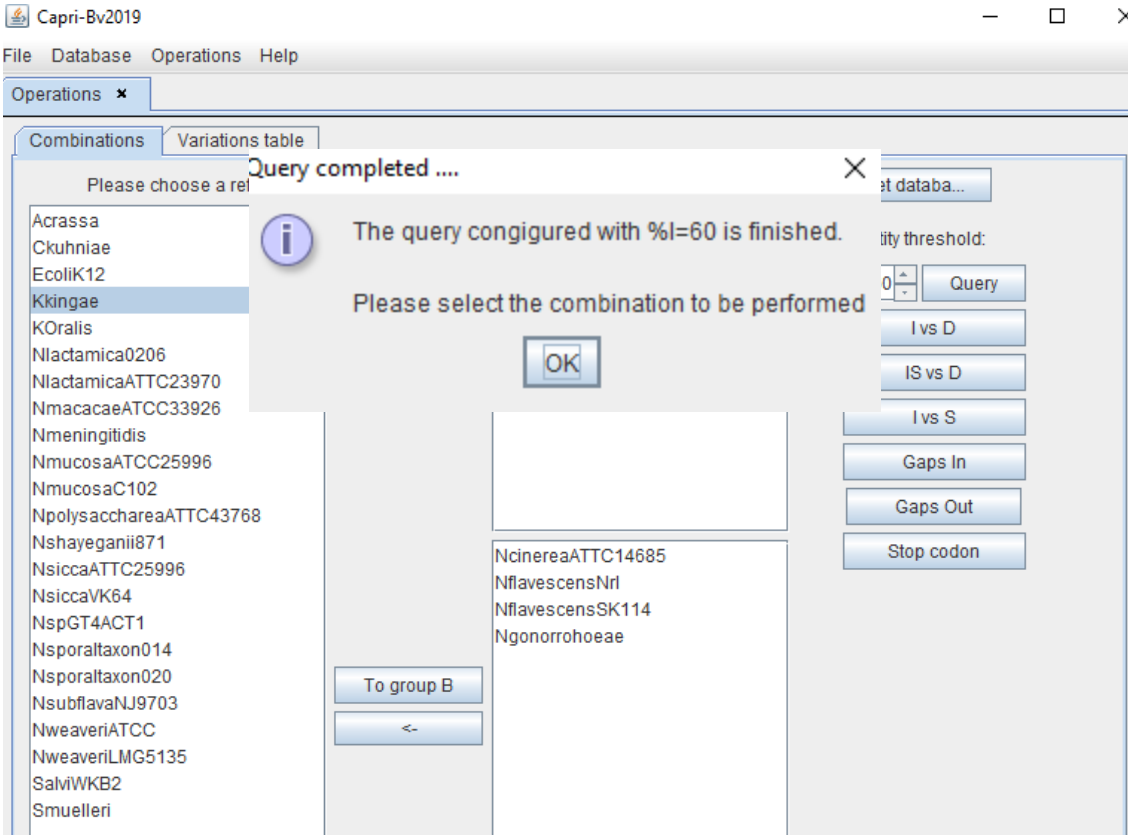
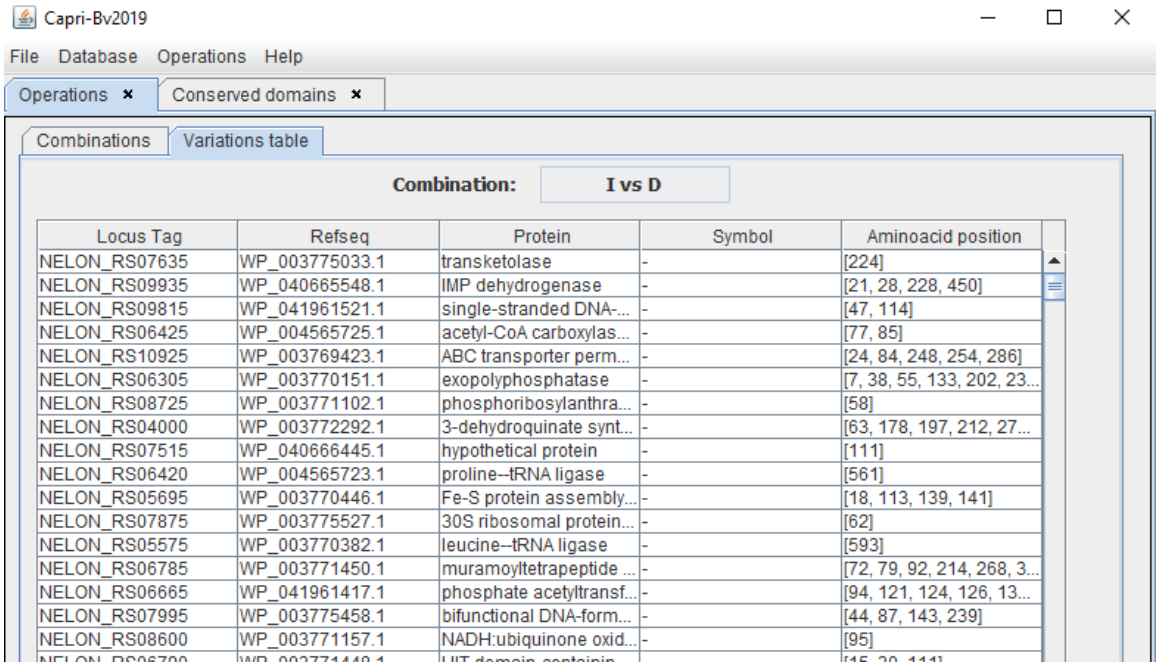


Figure 21 Panneau d'opérations.



Locus Tag	Refseq	Protein	Symbol	Aminoacid position
NELON_RS07635	WP_003775033.1	transketolase	-	[224]
NELON_RS09935	WP_040665548.1	IMP dehydrogenase	-	[21, 28, 228, 450]
NELON_RS09815	WP_041961521.1	single-stranded DNA-...	-	[47, 114]
NELON_RS06425	WP_004565725.1	acetyl-CoA carboxylas...	-	[77, 85]
NELON_RS10925	WP_003769423.1	ABC transporter perm...	-	[24, 84, 248, 254, 286]
NELON_RS06305	WP_003770151.1	exopolyphosphatase	-	[7, 38, 55, 133, 202, 23...
NELON_RS08725	WP_003771102.1	phosphoribosylanthra...	-	[58]
NELON_RS04000	WP_003772292.1	3-dehydroquinate synt...	-	[63, 178, 197, 212, 27...
NELON_RS07515	WP_040666445.1	hypothetical protein	-	[111]
NELON_RS06420	WP_004565723.1	proline--tRNA ligase	-	[561]
NELON_RS05695	WP_003770446.1	Fe-S protein assembly...	-	[18, 113, 139, 141]
NELON_RS07875	WP_003775527.1	30S ribosomal protein...	-	[62]
NELON_RS05575	WP_003770382.1	leucine--tRNA ligase	-	[593]
NELON_RS06785	WP_003771450.1	muramoyltetrapeptide ...	-	[72, 79, 92, 214, 268, 3...
NELON_RS06665	WP_041961417.1	phosphate acetyltransf...	-	[94, 121, 124, 126, 13...
NELON_RS07995	WP_003775458.1	bifunctional DNA-form...	-	[44, 87, 143, 239]
NELON_RS08600	WP_003771157.1	NADH:ubiquinone oxid...	-	[95]

Figure 23 Prévisualisation des résultats I vs D en CAPRIB.

La deuxième option, « CDD NCBI » sert à communiquer avec la base de données de domaines conservés de la NCBI (Figure 26). Cette base de données utilise le Refseq pour chercher les domaines conservés des proteines et produire une liste de domaines avec les positions, donc on l'utilise pour obtenir ces domaines et puis on vérifie. On utilise le fichier CSV obtenu auparavant pour lancer une requête avec le bouton « CDD search » et lorsque CAPRIB reçoit la réponse alors il combine les résultats obtenus avec certaine information du fichier CSV pour produire un fichier TSV (Annexe 3) et une prévisualisation de ce fichier apparait dans la partie droite du panneau (Figure 25). CAPRIB nous permet de chercher aussi de communiquer avec l'option « Features » de la NCBI avec le bouton « CDD NCBI » qui nous dirige au site d'internet pour télécharger le fichier features.txt. Ensuite on le charge en CAPRIB et avec le bouton « Merge » on le combine avec nos résultats du fichier CSV pour produire un nouveau fichier « features » combiné au format TSV.

.

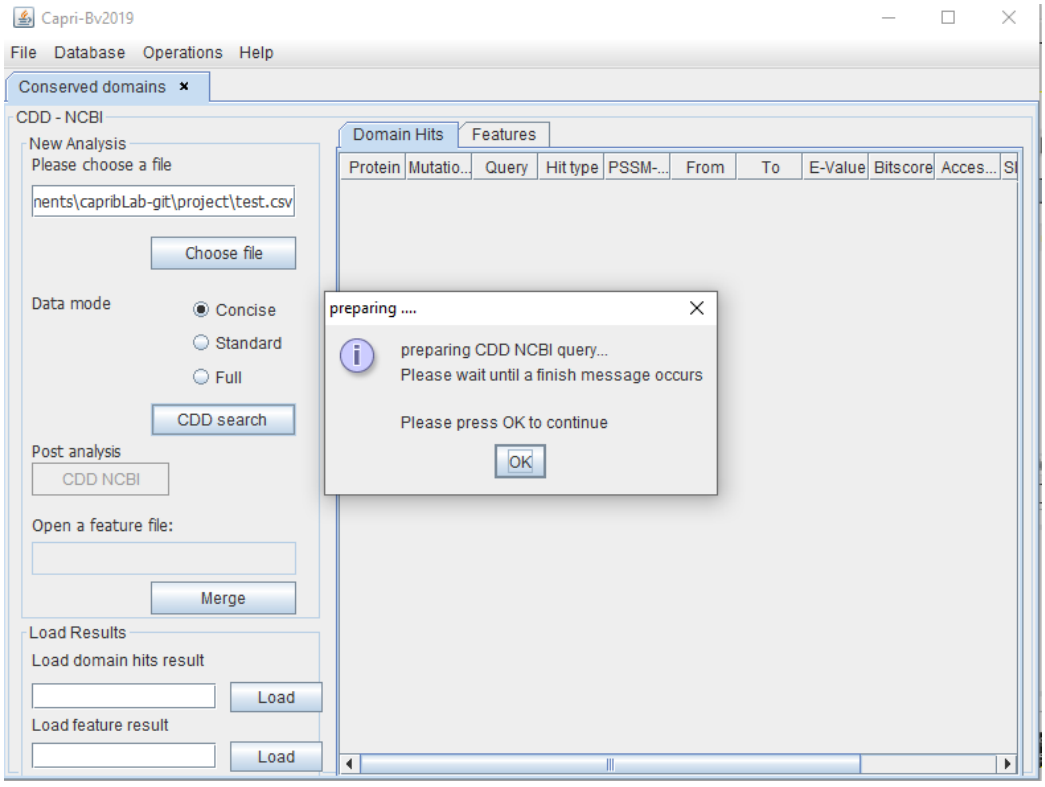


Figure 25 Panneau de communication avec CDD NCBI.

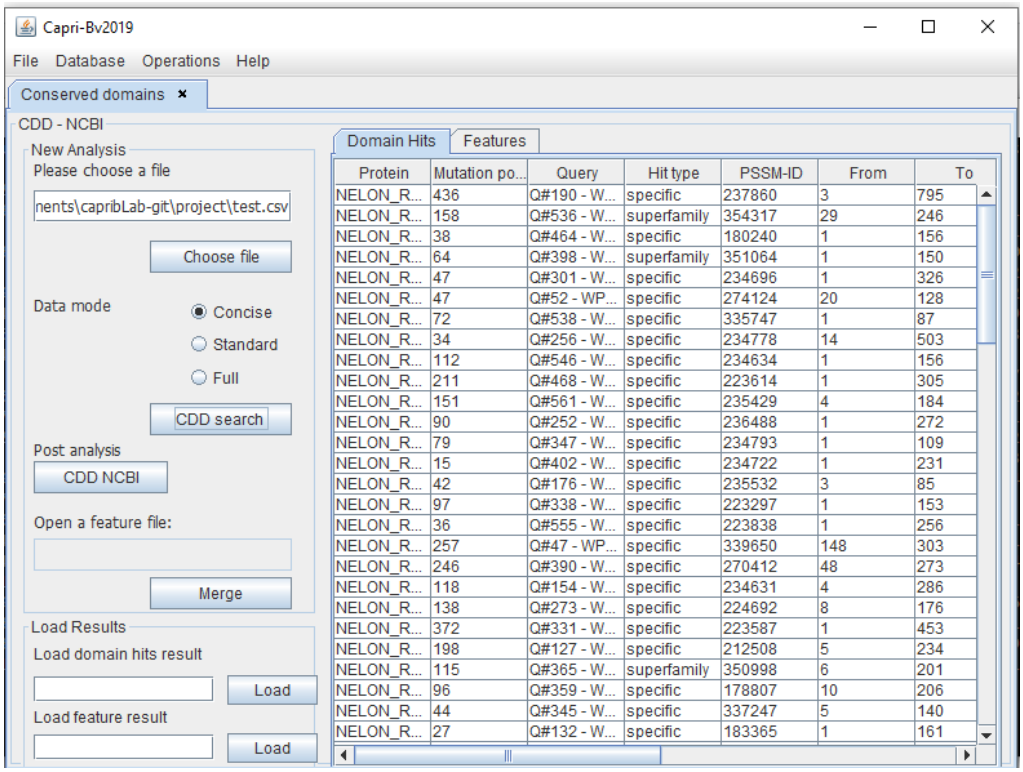


Figure 24 Prévisualisation du fichier TSV

3.2.4 *HELP*

Cette option nous mène vers un tutoriel en format HTML qui est présenté ici dans l'annexe 2.

3.3 ANALYSES COMPARATIVE (COQUES VS BACILLES) DES PROTÉINES CONSERVÉES DANS LA FAMILLE *NEISSERIACEAE* AVEC CAPRIB

3.3.1 *Comparaison en différents systèmes d'opération*

Avec *N. elongata* comme référence, on a construit des bases de données avec les 26 blast dans les systèmes d'opération Windows, Linux et MacOS et, dans le contexte coques vs bacilles, pour chaque combinaison on a obtenu les mêmes résultats ce qui nous indiquent que CAPRIB reproduit les résultats pour une référence dans les trois plateformes.

Tableau VII Nombre de mutations trouvées aux différentes combinaisons avec %I=60 et *N. elongata* comme référence

Combinaison	Windows	Linux	MacOS
IvsD	215	215	215
ISvsD	377	377	377
IvsS	270	270	270
Gaps In	7	7	7
Gaps Out	10	10	10
Stop codon	0	0	0

3.3.2 *Comparaison entre différentes références*

Les références à comparer présentent déjà des différences quant à la quantité de protéines rapportées dans leur fichier fasta (Tableau VIII), mais on a quand même réalisé les blast pour chercher les protéines conservées.

Tableau VIII Nombre de protéines d'après le fichier fasta de chaque organisme

Organisme	Nombre de Protéines
N. elongata	2100
K. kingae	1833
S. alvi	2216
N. bacilliformis	2235

Le premier pas de CAPRIB est d'établir le nombre protéines communes à tous les organismes comparés dans le seuil établi par l'usager. Dans le Tableau IX, on observe que quand on utilise *N. elongata* comme référence on obtient le majeur nombre de protéines communes avec un %I égal ou supérieur de 60. Quant au nombre de protéines candidates, on observe le même patron dans les différentes combinaisons sauf les gaps Figure 26.

Tableau IX Nombre de protéines communes avec un %I de 60

Organisme de référence	Nombre de protéines communs
N. elongata	583
N. bacilliformis	570
K. kingae	536
S. alvi	509

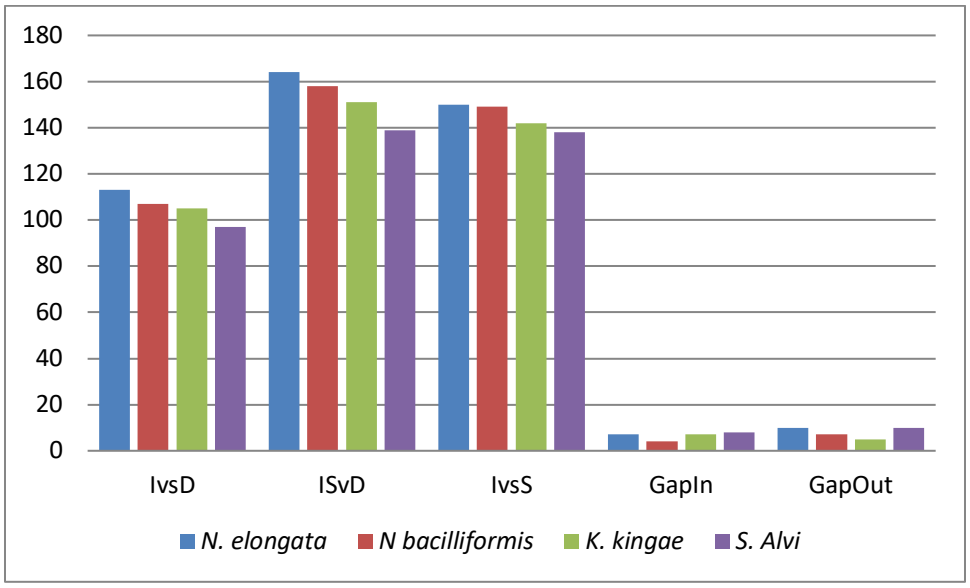


Figure 26 Nombre de protéines candidates pour les différentes combinaisons avec %I=60 avec les organismes de référence *N. elongata, K. kingae, S. alvi, N. bacilliformis*

3.4 LIMITATIONS, PROBLÈMES ET SOLUTIONS

On a trouvé différents situations pour améliorer en CAPRIB. D'abord, CAPRIB dépend de la qualité des séquences obtenues du NCBI. Si on a une mauvaise séquence de protéine pour l'organisme de référence ou d'ADN pour les autres organismes, alors on aura un résultat Blast qui peut fausser l'analyse. Il est dont fortement recommandé de vérifier la qualité des séquences de travail.

Ensuite, les résultats présentés dans la figure 27 nous montrent que le meilleur organisme de référence est *N. elongata* qui nous offre un meilleur nombre de protéines candidates avec un seuil de %I égal à 60. On remarque aussi que le nombre de candidats des organismes plus proches au nœud 1 (Figure 7) où il y a eu le changement de bacilles vers coques, comme *N. bacilliformis* et *N. elongata* ont plus de protéines conservées par rapport à *S. alvi* qui est l'organisme le plus éloigné phylogénétiquement.

Une limitation repose dans le théorème de base du programme. On détectera les mutations ponctuelles, ces changements aperçus au même temps dans le groupe B grâce à une opération d'intersection d'ensembles. Supposons une protéine quelconque « P » et suite à une mutation ponctuelle dans un évènement évolutif elle sera aperçue si et seulement si elle est présente dans tous les organismes du groupe B, si non la mutation ne sortira pas dans les candidates. Ce sera avantageux de faire une statistique pour savoir dans combien d'organismes du groupe B la mutation est présente.

Un autre point important à souligner est le fait que CAPRIB obtient le premier hit avec le meilleur e-value pour chaque protéine, néanmoins on pourrait modifier le code pour réaliser le blast et obtenir plus de hits et ensuite modifier le fichier « parseBlastFilter.pl », pour les extraire.

Quant aux codons stop qui sont détectés seulement dans le groupe B, il faut prendre en compte que les noms des groupes A et B sont arbitraires, on peut donc choisir un organisme de référence de chaque phénotype et ainsi trouver les AA identiques de chaque groupe, les codons stop de chaque groupe et ainsi de suite avec toutes les analyses. Par exemple, si l'on étudie le changement de forme entre bacilles vers coques et on veut savoir le codon stop pour les coques on utilise un organisme de référence des bacilles, au cas contraire on utilise un organisme de référence des coques.

Finalement, le fait d'utiliser TBlastN peut mener à de faux positifs et de faux négatifs. Par exemple un faux positif peut se trouver avec une correspondance dans un pseudo gène ou avec la moitié d'un gène et un faux négatif lorsqu'on établit le pourcentage d'identité, car la protéine peut avoir la mutation, mais elle n'est pas prise en compte parce qu'elle a un pourcentage d'identité plus faible, il est donc très important une fois qu'on a les candidates de faire un multi-alignement avec les protéines candidates afin de valider les résultats.

CHAPITRE 4 : CONCLUSION, PERSPECTIVES

CONCLUSIONS

On a développé un outil bioinformatique convivial, CAPRIB, qui prend « en entrée » des fichiers fasta et nous donne « en sortie » une liste de protéines conservées avec des mutations qui peuvent avoir un impact dans leur fonctionnement.

CAPRIB va nous permettre de chercher des variations d'acides aminés qui ont de l'impact sur les fonctions des protéines et qui sont en relation avec l'apparition d'un phénotype (virulence, résistance aux antibiotiques…).

Pour avoir des bons résultats, il faut toujours vérifier la qualité des séquences du travail, avoir un organisme de référence pour chaque phénotype et réaliser un alignement multiple des protéines candidates obtenues.

Pour une nouvelle version de CAPRIB ce serait très important d'inclure une statistique des positions des acides aminés qui ont changé, et ainsi on pourra avoir une probabilité d'identifier des mutations compensatoires, car le fait de considérer l'intersection absolue entre les deux groupes est une limite du programme

Finalement, puisque CAPRIB cherche les protéines conservées prenant comme base le pourcentage d'identité de l'analyse blast, alors l'organisme de référence doit être proche du nœud où le changement du phénotype a eu lieu pour avoir un nombre maximal de protéines à étudier.

PERSPECTIVES

Il y existe encore deux points importants qui restent à développer, comme l'information qui a changé lors d'un évènement évolutif et l'interface graphique. Quant au premier, le programme développé auparavant, MycoHit (Veyrier *et al.*, 2015a), nous montre les insertions et délétions de gènes et CAPRIB les mutations qui peuvent avoir un impact dans la fonction des protéines conservées dans un évènement évolutif alors il sera important de développer un programme qui détecte les changements dans des éléments régulateurs comme des zones inter-géniques et des ARN régulateurs.

Il serait très intéressant aussi de voir si la position des protéines candidates a changé au cours de l'évolution, car il a été montré que l'ordre n'est pas disposé au hasard, mais il peut être une conséquence d'un évènement évolutif (Tamames *et al.*, 2001).

Développer une autre classe en Java pour analyser résultats de la requête faite para QueryProteinsDB avec le but d'obtenir une analyse statistique. Cette classe devrait produire une grande trame de données qui pourrait être traité dans un programme d'analyses statistiques. On y trouverait des mutations par protéine et par organisme et on pourrait même les grouper par phénotype.

Au niveau de l'interface graphique, il serait très pratique d'implémenter des graphiques avec les statistiques d'acides aminés qui ont changé et/ou la visualisation des mesures comme Grantham et EX, car ils permettraient un meilleur profit de l'information obtenue.

La structure de la base de données créée avec CAPRIB permet de l'adapter à MycoHit, comme on a pu observer avec le sous-menu 'Protein Query' il suffit d'ajouter une autre classe en java qui fasse la requête et montre les résultats des protéines présentes dans le groupe A mais absentes dans le B.

On utilise seulement l'information des AA différents et similaires pour le groupe B, mais on pourrait utiliser ces données pour le groupe A et ainsi déterminer dans toute la base de données quelles sont les mutations communes et les rares. Dès qu'on les retrouve après les analyses faites par CAPRIB, on aura une autre information mise à part les mesures de Grantham et EX pour mieux choisir dans les protéines candidates.

Grâce au jeu de données de la famille *Neisseriaceae,* on a pu développer CAPRIB et les résultats seront étudiés en profondeur par des biologistes qui pourront tester les protéines candidates qui ont eu une relation avec le phénotype pour les valider.

CHAPITRE 5 BIBLIOGRAPHIE

Abby S & Daubin V (2007) Comparative genomics and the evolution of prokaryotes. *Trends in Microbiology* 15(3):135-141.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215(3):403-410.

Bang C, Dagan T, Deines P, Dubilier N, Duschl WJ, Fraune S, Hentschel U, Hirt H, Hülter N, Lachnit T, Picazo D, Pita L, Pogoreutz C, Rädecker N, Saad MM, Schmitz RA, Schulenburg H, Voolstra CR, Weiland-Bräuer N, Ziegler M & Bosch TCG (2018) Metaorganisms in extreme environments: do microbes play a role in organismal adaptation? *Zoology (Jena)* 127:1-19.

Bazykin GA (2015) Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. *Biol Lett* 11(10).

Bendezú FO, Hale CA, Bernhardt TG & de Boer PA (2009) RodZ (YfgA) is required for proper assembly of the MreB actin cytoskeleton and cell shape in E. coli. *EMBO J* 28(3):193-204.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J & Sayers EW (2010) GenBank. *Nucleic Acids Res* 38(Database issue):D46-51.

Bernard CS, Sadasivam M, Shiomi D & Margolin W (2007) An altered FtsA can compensate for the loss of essential cell division protein FtsN in Escherichia coli. *Molecular microbiology* 64(5):1289-1305.

Brocchieri L (2001) Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol* 59(1):27-40.

Cairns J, Overbaugh J & Miller S (1988) The origin of mutants. *Nature* 335:142.

Camps M, Herman A, Loh E & Loeb LA (2007) Genetic constraints on protein evolution. *Crit Rev Biochem Mol Biol* 42(5):313-326.

Chan CX, Beiko RG, Darling AE & Ragan MA (2009) Lateral transfer of genes and gene fragments in prokaryotes. *Genome Biol Evol* 1:429-438.

Delsuc F, Brinkmann H & Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6(5):361-375.

den Blaauwen T (2018) Is Longitudinal Division in Rod-Shaped Bacteria a Matter of Swapping Axis? *Front Microbiol* 9:822.

DePristo MA, Weinreich DM & Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* 6(9):678-687.

Didelot X & Maiden MC (2010) Impact of recombination on bacterial evolution. *Trends Microbiol* 18(7):315-322.

Dos Santos Vasconcelos CR, de Lima Campos T & Rezende AM (2018) Building protein-protein interaction networks for Leishmania species through protein structural information. *BMC Bioinformatics* 19(1):85.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32(5):1792-1797.

Frost LS, Leplae R, Summers AO & Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology* 3:722.

Gingold H & Pilpel Y (2011) Determinants of translation efficiency and accuracy. *Mol Syst Biol* 7:481.

Gordo I, Perfeito L & Sousa A (2011) Fitness Effects of Mutations in Bacteria. *Journal of Molecular Microbiology and Biotechnology* 21(1-2):20-35.

Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185(4154):862-864.

Griffiths AJF (2000) *An introduction to genetic analysis.* W.H. Freeman, New York, 7th. xvii, 860 p. p. http://www.ncbi.nlm.nih.gov/books/NBK21766

Habibi Najafi MB (2013) *Bacterial Mutation; Types, Mechanisms and Mutant Detection Methods: a Review.* 1857-7431 p

Hersh MN, Morales LD, Ross KJ & Rosenberg SM (2006) Single-strand-specific exonucleases prevent frameshift mutagenesis by suppressing SOS induction and the action of DinB/DNA polymerase IV in growing cells. *J Bacteriol* 188(7):2336-2342.

Hershberg R (2015) Mutation—The Engine of Evolution: Studying Mutation and Its Role in the Evolution of Bacteria. *Cold Spring Harbor Perspectives in Biology* 7(9):a018077.

Huang KH, Durand-Heredia J & Janakiraman A (2013) FtsZ ring stability: of bundles, tubules, crosslinks, and curves. *J Bacteriol* 195(9):1859-1868.

Jiang C, Caccamo PD & Brun YV (2015) Mechanisms of bacterial morphogenesis: evolutionary cell biology approaches provide new insights. *Bioessays* 37(4):413-425.

Johnsonbaugh R (2000) *Discrete Mathematics.* Prentice Hall PTR. 621 p

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res* 20(10):1313-1326.

Kaur H, Amandeep S & Singh P (2008) *Comparison of Variants of BLAST.*

Koonin EV & Galperin MY (2003) *Sequence - evolution - function : computational approaches in comparative genomics.* Kluwer Academic Publishers, Boston. xiii, 461 p. p

Laddomada F, Miyachiro MM & Dessen A (2016) Structural Insights into Protein-Protein Interactions Involved in Bacterial Cell Wall Biogenesis. *Antibiotics (Basel)* 5(2).

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ & Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947-2948.

Lin M & Kussell E (2017) Correlated Mutations and Homologous Recombination Within Bacterial Populations. *Genetics* 205(2):891-917.

Lipman DJ & Pearson WR (1985) Rapid and sensitive protein similarity searches. *Science* 227(4693):1435.

Lynch M & Hagner K (2015) Evolutionary meandering of intermolecular interactions along the drift barrier. *Proc Natl Acad Sci U S A* 112(1):E30-38.

Maki H (2002) Origins of Spontaneous Mutations: Specificity and Directionality of Base-Substitution, Frameshift, and Sequence-Substitution Mutageneses. *Annual Review of Genetics* 36(1):279-303.

Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Geer LY & Bryant SH (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research* 45(Database issue):D200-D203.

Marchler-Bauer A & Bryant SH (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Research* 32(Web Server issue):W327-W331.

Matteï PJ, Neves D & Dessen A (2010) Bridging cell wall biosynthesis and bacterial morphogenesis. *Curr Opin Struct Biol* 20(6):749-755.

Moya A, Peretó J, Gil R & Latorre A (2008) Learning how to live together: genomic insights into prokaryote–animal symbioses. *Nature Reviews Genetics* 9:218.

Nagel R (2007) "LA MUTACIÓN ADAPTATIVA". POLÉMICAS Y MECANISMOS. *Journal of Basic & Applied Genetics* 18(1):51-59.

Needleman SB & Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3):443-453.

Nikolaidis I, Favini-Stabile S & Dessen A (2014) Resistance to antibiotics targeted to the bacterial cell wall. *Protein Sci* 23(3):243-259.

Noirot P & Noirot-Gros MF (2004) Protein interaction networks in bacteria. *Curr Opin Microbiol* 7(5):505-512.

Notredame C (2007) Recent Evolutions of Multiple Sequence Alignment Algorithms. *PLoS Computational Biology* 3(8):e123.

Notredame C, Higgins DG & Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1):205-217.
Ochman H & Moran NA (2001) Genes Lost and Genes Found: Evolution of Bacterial Pathogenesis and Symbiosis. *Science* 292(5519):1096.
Pál C & Papp B (2017) Evolution of complex adaptations in molecular systems. *Nat Ecol Evol* 1(8):1084-1092.
Pál C, Papp B & Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics* 37:1372.
Pál C, Papp B & Lercher MJ (2006) An integrated view of protein evolution. *Nature Reviews Genetics* 7:337.
Papp B, Notebaart RA & Pál C (2011) Systems-biology approaches for predicting genomic evolution. *Nat Rev Genet* 12(9):591-602.
Pevsner J (2009) *Bioinformatics and functional genomics.* Wiley-Blackwell, Hoboken, N.J., 2nd. 1 texte électronique p. http://onlinelibrary.wiley.com/book/10.1002/9780470451496 Accès réservé UdeM
Pickard JM, Zeng MY, Caruso R & Núñez G (2017) Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease. *Immunol Rev* 279(1):70-89.
Pinho MG, Kjos M & Veening JW (2013) How to get (a)round: mechanisms controlling growth and division of coccoid bacteria. *Nat Rev Microbiol* 11(9):601-614.
Ponder RG, Fonville NC & Rosenberg SM (2005) A switch from high-fidelity to error-prone DNA double-strand break repair underlies stress-induced mutation. *Mol Cell* 19(6):791-804.
Radomski N, Roguet A, Lucas FS, Veyrier FJ, Cambau E, Accrombessi H, Moilleron R, Behr MA & Moulin L (2013) atpE gene as a new useful specific molecular target to quantify Mycobacterium in environmental samples. *BMC Microbiol* 13:277.
Rappleye CA & Roth JR (1997) Transposition without transposase: a spontaneous mutation in bacteria. *Journal of Bacteriology* 179(6):2047-2052.
Reva B, Antipin Y & Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research* 39(17):e118-e118.
Rosenberg SM (2001) Evolving responsively: adaptive mutation. *Nat Rev Genet* 2(7):504-515.
Rosenberg SM, Harris RS, Longerich S & Galloway AM (1996) Recombination-dependent mutation in non-dividing cells. *Mutat Res* 350(1):69-76.
Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD & Karsch-Mizrachi I (2018) GenBank. *Nucleic Acids Res* 10.1093/nar/gky989.
Smith TF & Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* 147(1):195-197.
Sniegowski PD, Gerrish PJ, Johnson T & Shaver A (2000) The evolution of mutation rates: separating causes from consequences. *Bioessays* 22(12):1057-1066.
Soskine M & Tawfik DS (2010) Mutational effects and the evolution of new protein functions. *Nat Rev Genet* 11(8):572-582.
Tadrowski AC, Evans MR & Waclaw B (2018) Phenotypic Switching Can Speed up Microbial Evolution. *Sci Rep* 8(1):8941.
Tamames J, González-Moreno M, Mingorance J, Valencia A & Vicente M (2001) Bringing gene order into bacterial shape. *Trends Genet* 17(3):124-126.
Thomas CM & Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 3(9):711-721.
Thorson JLM, Smithson M, Beck D, Sadler-Riggleman I, Nilsson E, Dybdahl M & Skinner MK (2017) Epigenetics and adaptive phenotypic variation between habitats in an asexual snail. *Scientific Reports* 7(1):14139.
Torii N, Nozaki T, Masutani M, Nakagama H, Sugiyama T, Saito D, Asaka M, Sugimura T & Miki K (2003) Spontaneous mutations in the Helicobacter pylori rpsL gene. *Mutat Res* 535(2):141-145.
Typas A & Sourjik V (2015) Bacterial protein networks: properties and functions. *Nature Reviews Microbiology* 13:559.

Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF & van Sinderen D (2007) Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum. *Microbiology and Molecular Biology Reviews : MMBR* 71(3):495-548.

Veyrier F, Pletzer D, Turenne C & Behr MA (2009a) Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis. *BMC Evolutionary Biology* 9(1):196.

Veyrier F, Pletzer D, Turenne C & Behr MA (2009b) Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis. *BMC Evol Biol* 9:196.

Veyrier FJ, Biais N, Morales P, Belkacem N, Guilhen C, Ranjeva S, Sismeiro O, Péhau-Arnaudet G, Rocha EP, Werts C, Taha M-K & Boneca IG (2015a) Common Cell Shape Evolution of Two Nasopharyngeal Pathogens. *PLOS Genetics* 11(7):e1005338.

Veyrier FJ, Biais N, Morales P, Belkacem N, Guilhen C, Ranjeva S, Sismeiro O, Pehau-Arnaudet G, Rocha EP, Werts C, Taha MK & Boneca IG (2015b) Common Cell Shape Evolution of Two Nasopharyngeal Pathogens. *PLoS Genet* 11(7):e1005338.

Veyrier FJ, Dufort A & Behr MA (2011) The rise and fall of the Mycobacterium tuberculosis genome. *Trends Microbiol* 19(4):156-161.

Wang J, McIntosh F, Radomski N, Dewar K, Simeone R, Enninga J, Brosch R, Rocha EP, Veyrier FJ & Behr MA (2015) Insights on the emergence of Mycobacterium tuberculosis from the analysis of Mycobacterium kansasii. *Genome biology and evolution* 7(3):856-870.

White CL, Kitich A & Gober JW (2010) Positioning cell wall synthetic complexes by the bacterial morphogenetic proteins MreB and MreD. *Mol Microbiol* 76(3):616-633.

Whitehead DJ, Wilke CO, Vernazobres D & Bornberg-Bauer E (2008) The look-ahead effect of phenotypic mutations. *Biol Direct* 3:18.

Yamada T & Bork P (2009) Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nat Rev Mol Cell Biol* 10(11):791-803.

Yampolsky LY & Stoltzfus A (2005a) The exchangeability of amino acids in proteins. *Genetics* 170(4):1459-1472.

Yampolsky LY & Stoltzfus A (2005b) The exchangeability of amino acids in proteins. *Genetics* 170(4):1459-1472.

Yanagida H, Gispan A, Kadouri N, Rozen S, Sharon M, Barkai N & Tawfik DS (2015) The Evolutionary Potential of Phenotypic Mutations. *PLoS Genet* 11(8):e1005445.

Zhang Z, Schwartz S, Wagner L & Miller W (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7(1-2):203-214.

Zhou H & Lutkenhaus J (2004) The switch I and II regions of MinD are required for binding and activating MinC. *J Bacteriol* 186(5):1546-1555.

CHAPITRE 6 ANNEXES

ANNEXE 1

BEFORE STARTING

CAPRIP is a user-friendly tool implemented in java but useS Perl to parse and filter blast results and MySQL to store them. Consequently, before to start to use CAPRIB a minimal Ram 4 Gb, Internet access and the installation of additional open source software are required. Choose a location in your computer and unzip caprib.zip

JAVA

https://www.java.com/fr/download/manual.jsp

Choose your platform (e.g. Windows 10, Ubuntu, Mac OSX).

For Mac you need to install additionally JDK file from:

https://www.oracle.com/technetwork/java/javase/downloads/index.html

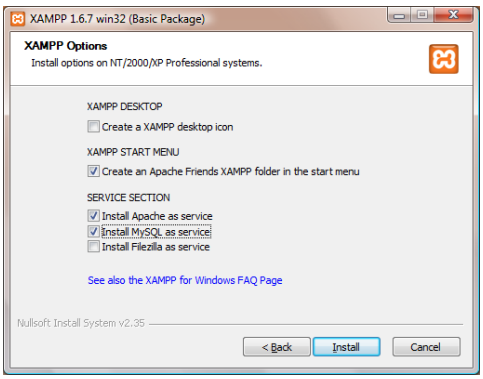
Follow instructions.

XAMPP

https://www.apachefriends.org/index.html

Figure 27 XAMPP Installation options for Windows

- Choose operating system platform (e.g. Windows, Linux, OS X)
- For Windows and Linux download Xampp version 7.3.0-0
- For OS X download Xampp version 5.6.39
- Download and follow instructions
- Choose install XAMPP and Apache as shown in Figure 28



Start Xampp form the executable.

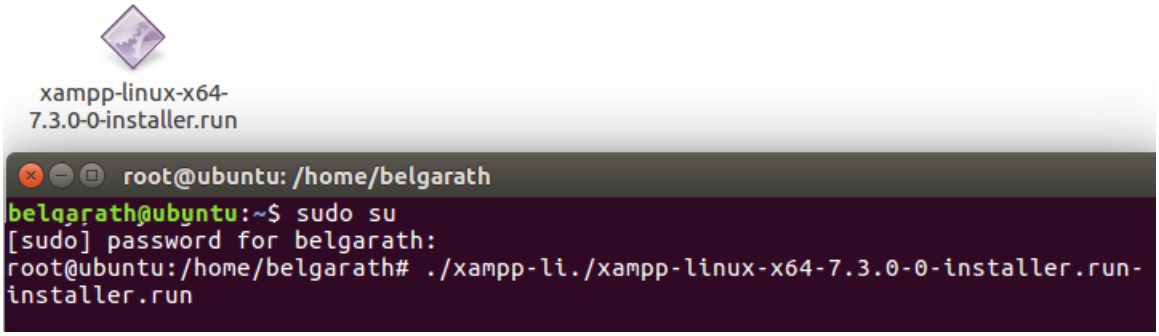
To install Xampp in linux, open a terminal in the download folder and type:

sudo su

Then enter your password and finally type:

./xampp-li./xampp-linux-x64-7.3.0-0-installer.run-installer.run (Figure 28)

Figure 28 Xampp installation in Ubuntu



To start Xampp from console: */opt/lampp/lampp start*

To stop Xampp from console*: /opt/lampp/lampp stop*

PERL

https://www.activestate.com/activeperl/downloads

Choose your platform (e.g. Windows 10, Ubuntu, Mac OSX) and follow instructions. Normally Ubuntu and Mac OSX have perl in their operative system, you can check its version in your terminal typing:

perl -v

BLAST+

CAPRIP use Blast+ version 2.9.0 to compare protein fasta files with a nucleotides database, you can install it in Windows following instructions in:

https://www.ncbi.nlm.nih.gov/books/NBK52637/

Click the link to download installer from:

ftp://ftp.ncbi.nih.gov/blast/executables/blast+/2.9.0/

START CAPRIB

Windows: double click in the Caprib2018.jar

Linux/Mac: type in the terminal: java -jar Caprib2018.jar

ANNEXE 2

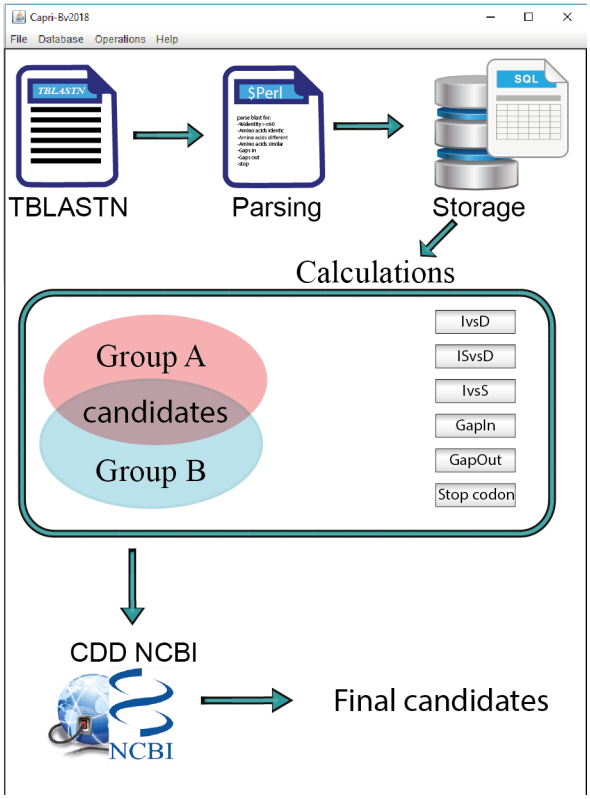
TUTORIEL DE CAPRIB (http://fveyrier.profs.inrs.ca/Download/tutorial/index.html)

Home

Caprib Tutorial

Caprib is a bioinformatics tool that allows researchers to study evolutionary events by comparing whole-bacterial-genome sequences. Overall, this tool generates a list of proteins with amino acid substitutions that are predicted to have impacted biological functions at a given node of evolution. These substitutions may explain the emergence of a given phenotype for a group of bacteria at this particular node of evolution.

Caprib operations

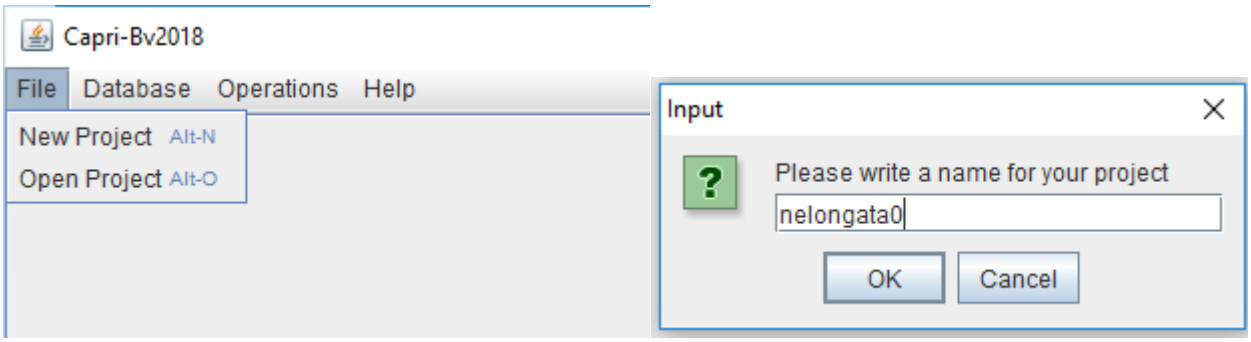


File menu

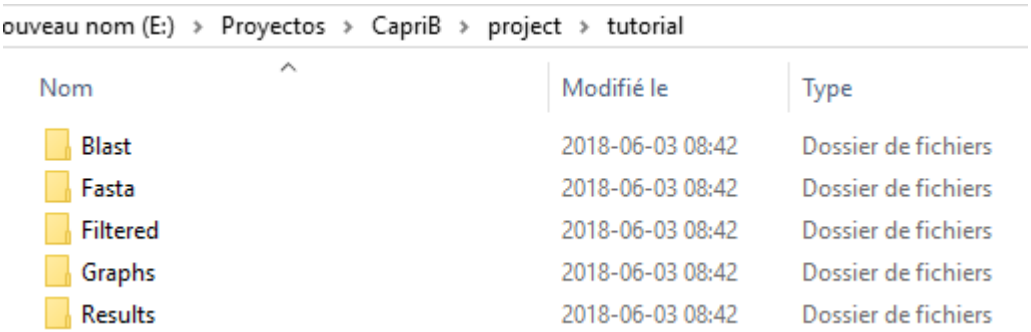
In this menu the user have two choices, create a new project or open an existing one.

New project

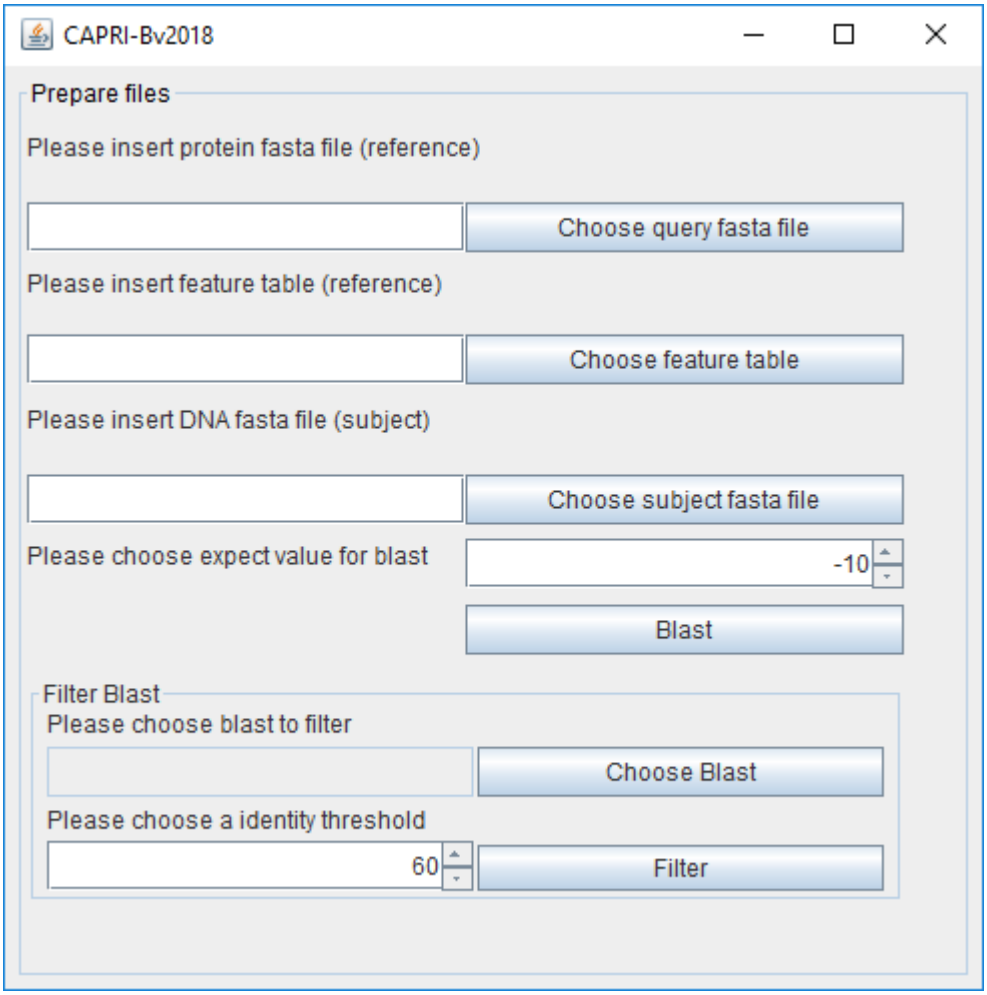
By clicking new project in the File menu, a window will appear that has to be filled with a name of the user choice (do not use spaces). In our example, "nelongata0" project was used.



Following this, a directory will be created to store all information. This directory is composed of five folders: Blast, Fasta, Filtered, Graphs and Results in the path /caprib/project/nelongata0/



A new window will appear to start blast and to filter the blast reports. Of note, all created files (fasta, blast and filtered files) will be saved in their respective folders.



Open Project

This option is used when the user have other analysis to do on an existing project.

Then, the user can choose the existing project.

Blast & Filter

To start the blast, the user will need input files: the protein fasta file and the feature file for the reference organism. *Neisseria elongata* is used in our example. The files are provided for training purposes in the tutorial folder named respectively:

GCF_000818035.1_ASM81803v1_protein.faa

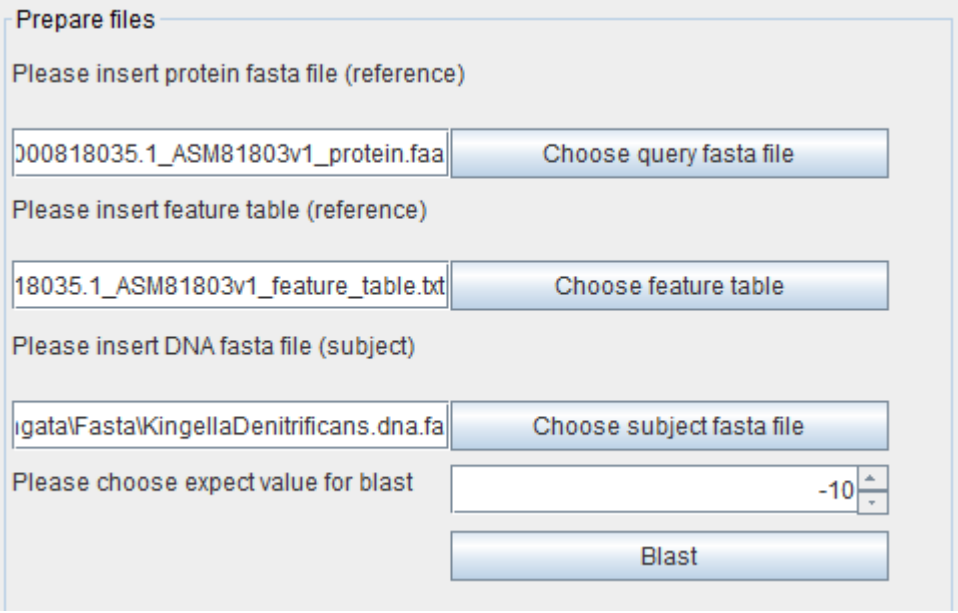
GCF_000818035.1_ASM81803v1_feature_table.txt

For other reference organisms these files (.faa and _feature_table.txt) are generally available on the NCBI FTP: (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/). The third output file needed is the gDNA sequence (.dna) as a fasta files for the other organism that will be compared (herein *Kingella denitrificans*) with the reference. This file can also be found in the NCBI FTP. The user can choose the expect (E) value (for explanation on this value see:

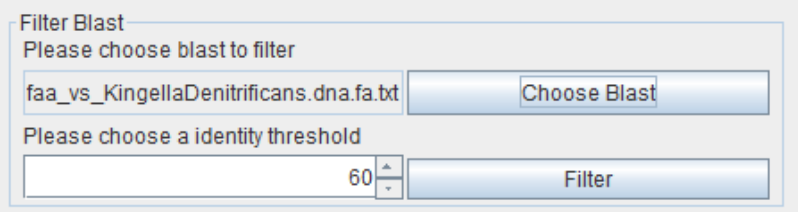
https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#expect)

It is expressed in e^(value enter by the user). By clicking "Blast" the tblastn analyses will start.

The report of these analyses is stored in the Blast folder.

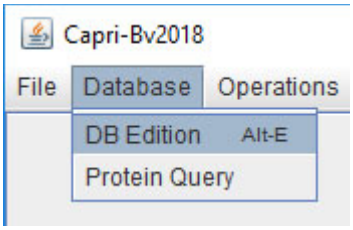


The user must do this operation for all genomes that will be compared with the reference. At this point, the user can filter the blast results (---vs---.txt in the "blast" folder) separately to exclude non-conserved proteins if needed. The % of identity score threshold is chosen by the user. A threshold of 0% will include all proteins and threshold can be defined latter. Again, this operation needs to be done for all genomes. The output created file will be stored in the "Filtered" folder.



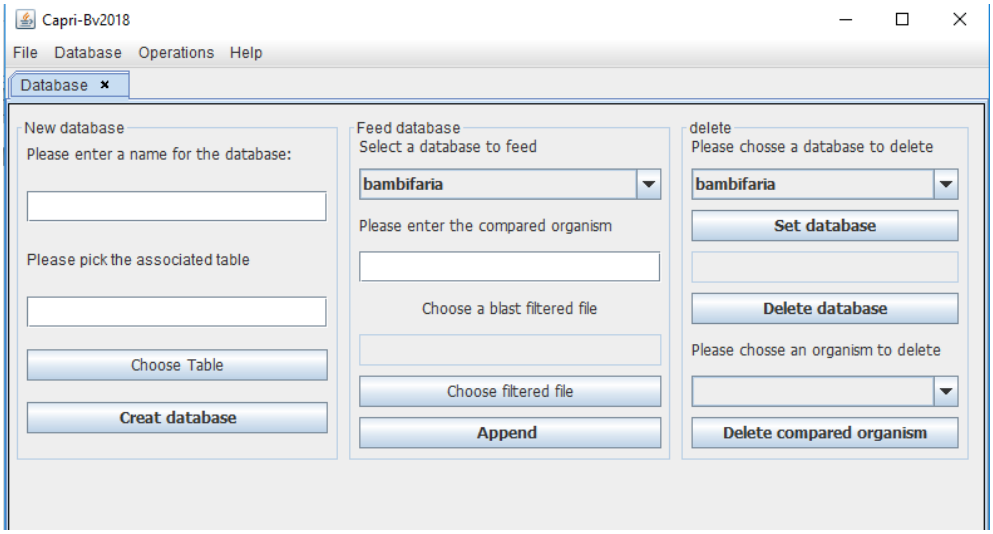
Database menu

In this menu the user has two choices, DB Edition and Protein query.



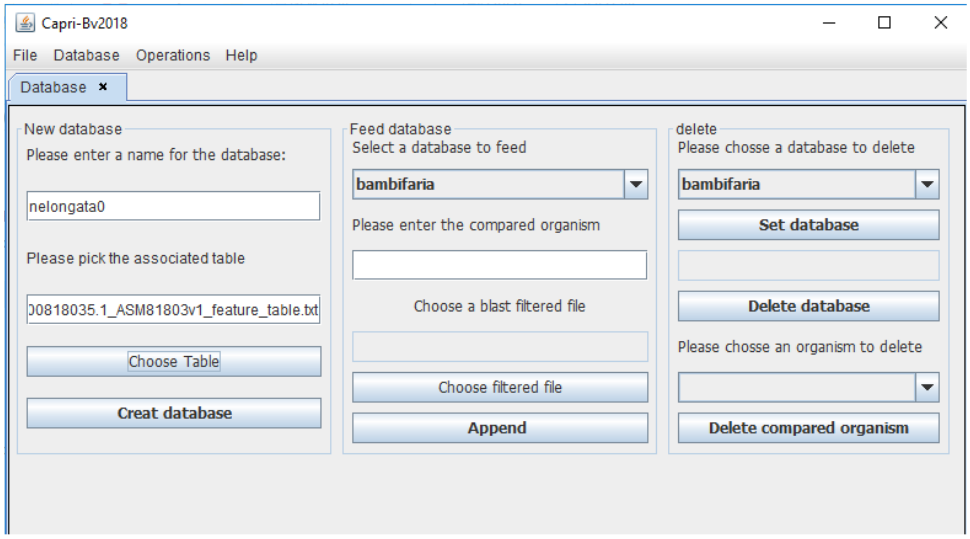
DB Edition

This is used to build databases. After the user chooses this option a panel will appear with three options: create the database, feed the database or delete one entry or a whole database.

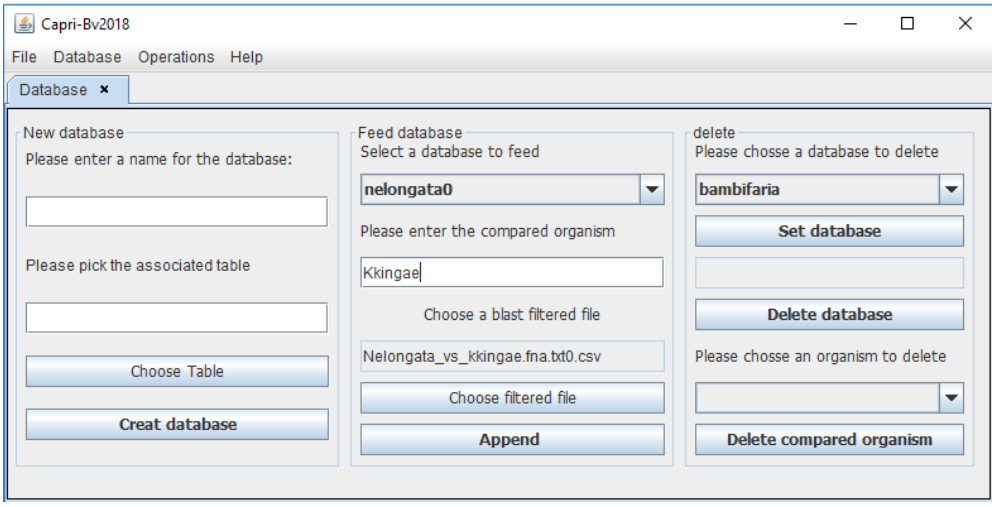


To build the database, the user needs to give a name to the database (usually the name of the reference) and associate its table file (_feature_table.txt in the "Fasta" folder), in this example for *N. elongata* as reference:
name: nelongata0

-GCF_000818035.1_ASM81803v1_feature_table.txt



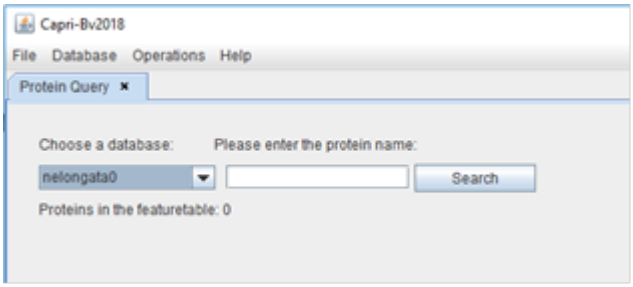
To feed the database (herein nelongata0), the user will add the name of this organism to append (without spaces) and choose the corresponding filtered file (.csv in the "Filtered" folder).

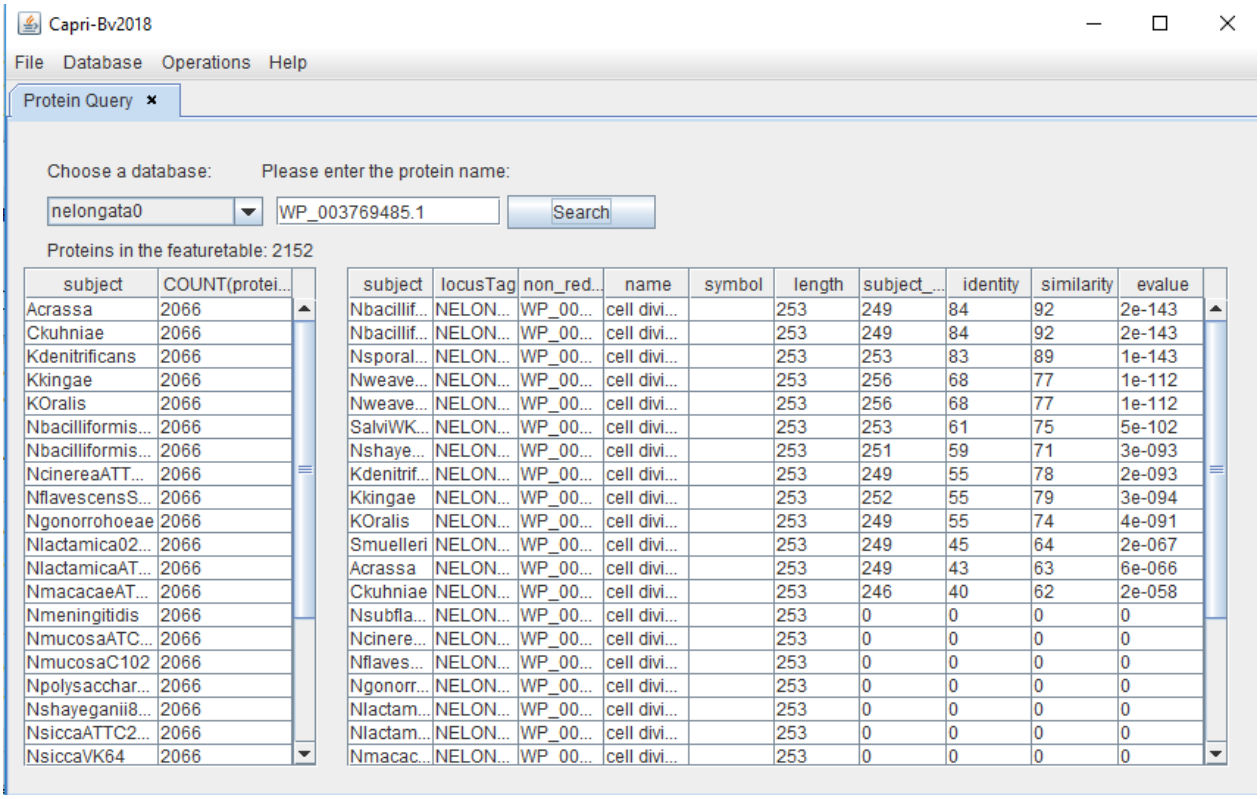


The last option is to delete an organism or the whole database, the user select the required database and then click on "set database" to provoke the changes.

Protein Query

This option is to query the database in order to find information on a given protein of interest. The user can simply type the ref.seq or the locus tag to obtain the results.



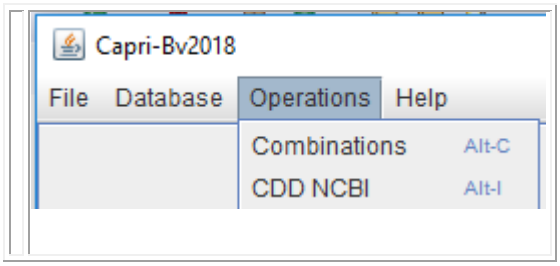


subject	COUNT(protei...
Acrassa	2066
Ckuhniae	2066
Kdenitrificans	2066
Kkingae	2066
KOralis	2066
Nbacilliformis...	2066
Nbacilliformis...	2066
NcinereaATT...	2066
NflavescensS...	2066
Ngonorrohoeae	2066
Nlactamica02...	2066
NlactamicaAT...	2066
NmacacaeAT...	2066
Nmeningitidis	2066
NmucosaATC...	2066
NmucosaC102	2066
Npolysacchar...	2066
Nshayeganii8...	2066
NsiccaATTC2...	2066
NsiccaVK64	2066

subject	locusTag	non_red...	name	symbol	length	subject_...	identity	similarity	evalue
Nbacillif...	NELON...	WP_00...	cell divi...		253	249	84	92	2e-143
Nbacillif...	NELON...	WP_00...	cell divi...		253	249	84	92	2e-143
Nsporal...	NELON...	WP_00...	cell divi...		253	253	83	89	1e-143
Nweave...	NELON...	WP_00...	cell divi...		253	256	68	77	1e-112
Nweave...	NELON...	WP_00...	cell divi...		253	256	68	77	1e-112
SalviWK...	NELON...	WP_00...	cell divi...		253	253	61	75	5e-102
Nshaye...	NELON...	WP_00...	cell divi...		253	251	59	71	3e-093
Kdenitrif...	NELON...	WP_00...	cell divi...		253	249	55	78	2e-093
Kkingae	NELON...	WP_00...	cell divi...		253	252	55	79	3e-094
KOralis	NELON...	WP_00...	cell divi...		253	249	55	74	4e-091
Smuelleri	NELON...	WP_00...	cell divi...		253	249	45	64	2e-067
Acrassa	NELON...	WP_00...	cell divi...		253	249	43	63	6e-066
Ckuhniae	NELON...	WP_00...	cell divi...		253	246	40	62	2e-058
Nsubfla...	NELON...	WP_00...	cell divi...		253	0	0	0	0
Ncinere...	NELON...	WP_00...	cell divi...		253	0	0	0	0
Nflaves...	NELON...	WP_00...	cell divi...		253	0	0	0	0
Ngonorr...	NELON...	WP_00...	cell divi...		253	0	0	0	0
Nlactam...	NELON...	WP_00...	cell divi...		253	0	0	0	0
Nlactam...	NELON...	WP_00...	cell divi...		253	0	0	0	0
Nmacac...	NELON...	WP_00...	cell divi...		253	0	0	0	0

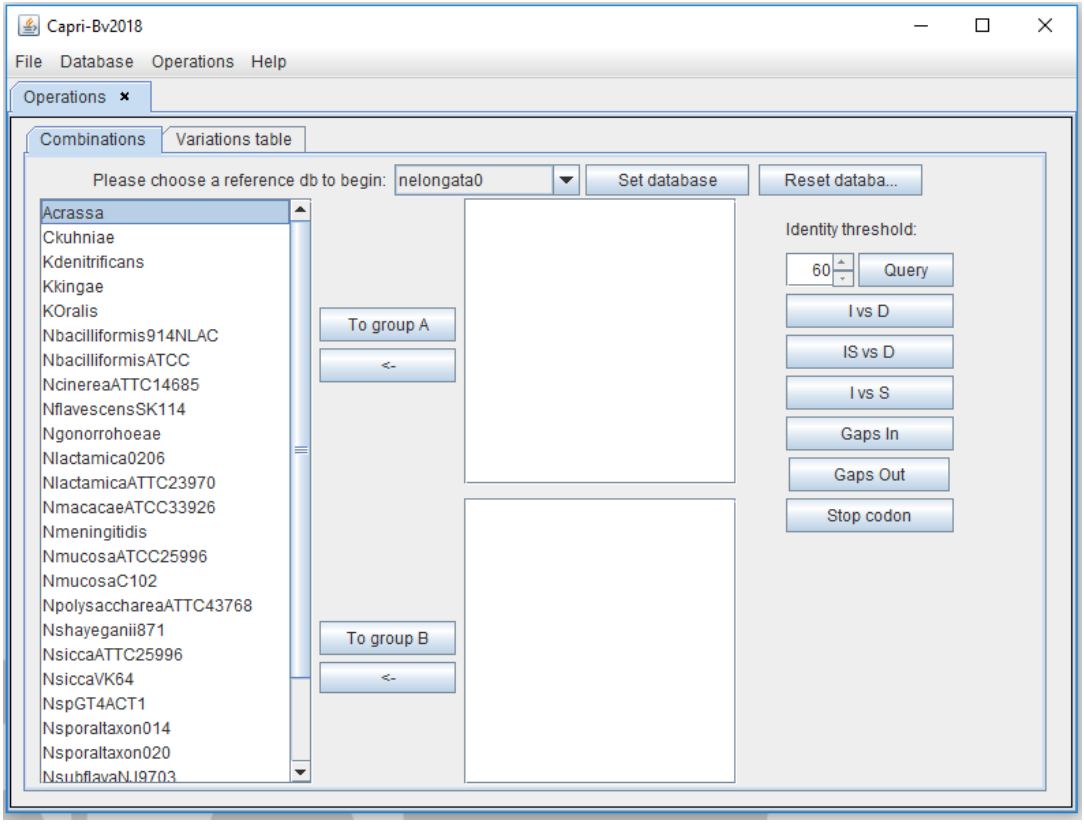
Operations menu

Subsequently to database feeding, there is two options. First, “combinations” to perform operations aiming to comparing two groups of organisms to obtain list of variations in conserved proteins. Second, that is dependant to achievement of the first one, is “CDD NCBI” in order to compare these lists to conserved domains in CDD databases in order to rank variations that could functionally affect proteins.

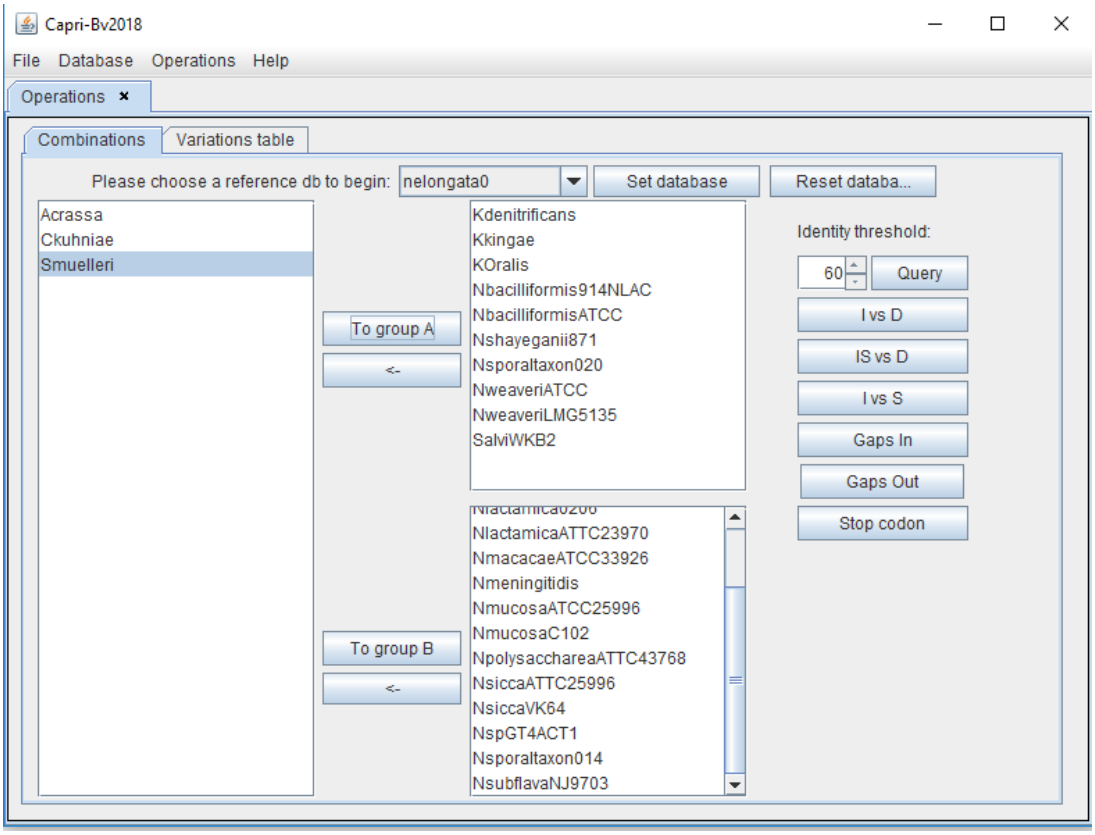


Combinations

After clicking combination, the user can set the database created previously that will be used (here in "nelongata0") for analyses. All the organisms in the database will be displayed as shown in the picture below.



The user will have to classify the organisms into two groups based on the phylogeny (ex: bacteria that diverge before/after the coccoid transition). The reference bacterium is always in group A. Therefore other organisms need to be separated in accordance. In the example, bacteria that diverge before the coccoid transition (bacilli) will be place in group A (as is the reference). The bacteria that diverge after this node (coccus) are placed in group B.



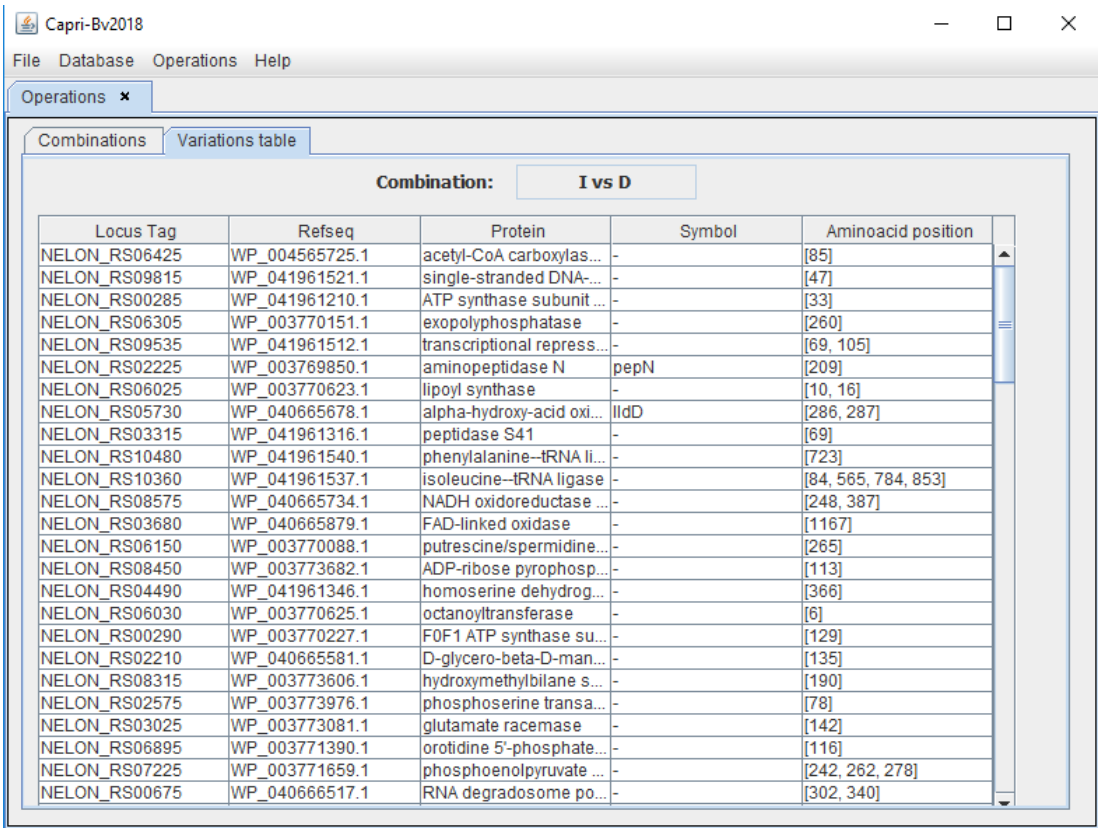
Next, the user can filter the proteins by percentage of identity, (herein at least 60% identical in all species) and then click "query". This will set the database as 60%.

Both groups can be compared based on 6 different analyses:

1. IvsD: Search for identical amino acid positions in group A but different in B
2. ISvsD: Search for amino acid positions identical or similar in group A but different in group B
3. IvsS: Search for identical amino acid positions in group A but similar in group B
4. Gap In: Search for the position of a gap that is in group A but not in group B
5. Gap Out: Search for the position of a gap that is in group B but not in group A
6. Stop codon: Search for the position of a stop codon in group B

Results

The user will have to save and name the results file in the “Results” folder. The extension ".CSV" need to be added. The results can be displayed on the screen for a preview, but the user can always open the file as a spreadsheet in Excel or Calc.

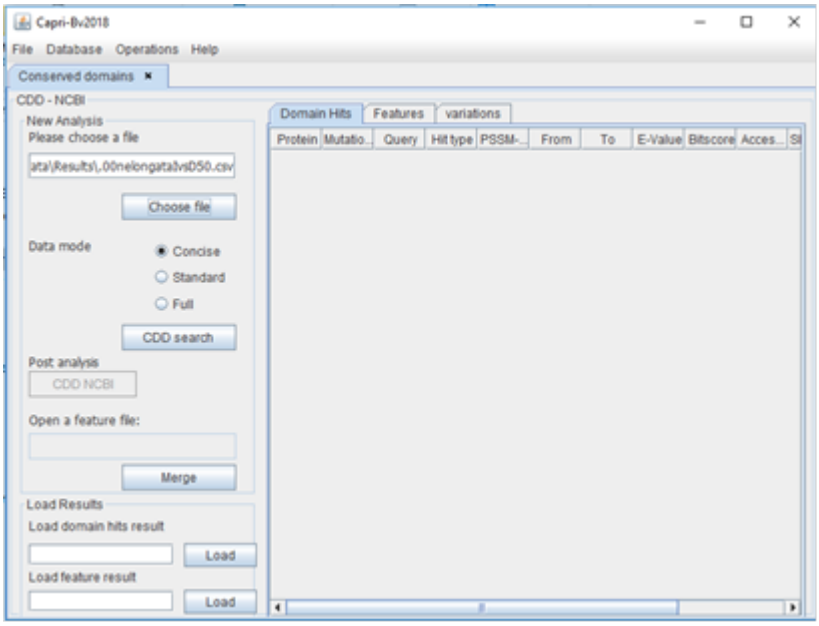


Locus Tag	Refseq	Protein	Symbol	Aminoacid position
NELON_RS06425	WP_004565725.1	acetyl-CoA carboxylas...	-	[85]
NELON_RS09815	WP_041961521.1	single-stranded DNA-...	-	[47]
NELON_RS00285	WP_041961210.1	ATP synthase subunit ...	-	[33]
NELON_RS06305	WP_003770151.1	exopolyphosphatase	-	[260]
NELON_RS09535	WP_041961512.1	transcriptional repress...	-	[69, 105]
NELON_RS02225	WP_003769850.1	aminopeptidase N	pepN	[209]
NELON_RS06025	WP_003770623.1	lipoyl synthase	-	[10, 16]
NELON_RS05730	WP_040665678.1	alpha-hydroxy-acid oxi...	lldD	[286, 287]
NELON_RS03315	WP_041961316.1	peptidase S41	-	[69]
NELON_RS10480	WP_041961540.1	phenylalanine--tRNA li...	-	[723]
NELON_RS10360	WP_041961537.1	isoleucine--tRNA ligase	-	[84, 565, 784, 853]
NELON_RS08575	WP_040665734.1	NADH oxidoreductase ...	-	[248, 387]
NELON_RS03680	WP_040665879.1	FAD-linked oxidase	-	[1167]
NELON_RS06150	WP_003770088.1	putrescine/spermidine...	-	[265]
NELON_RS08450	WP_003773682.1	ADP-ribose pyrophosp...	-	[113]
NELON_RS04490	WP_041961346.1	homoserine dehydrog...	-	[366]
NELON_RS06030	WP_003770625.1	octanoyltransferase	-	[6]
NELON_RS00290	WP_003770227.1	F0F1 ATP synthase su...	-	[129]
NELON_RS02210	WP_040665581.1	D-glycero-beta-D-man...	-	[135]
NELON_RS08315	WP_003773606.1	hydroxymethylbilane s...	-	[190]
NELON_RS02575	WP_003773976.1	phosphoserine transa...	-	[78]
NELON_RS03025	WP_003773081.1	glutamate racemase	-	[142]
NELON_RS06895	WP_003771390.1	orotidine 5'-phosphate...	-	[116]
NELON_RS07225	WP_003771659.1	phosphoenolpyruvate ...	-	[242, 262, 278]
NELON_RS00675	WP_040666517.1	RNA degradosome po...	-	[302, 340]

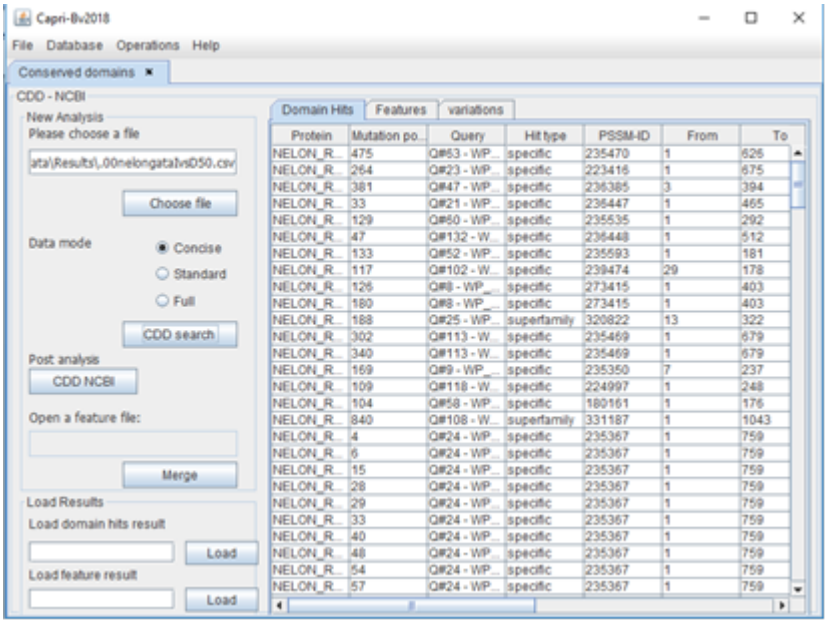
As example, the protein with locus tag NELON_RS0625 has a variation in the AA at position 85 This AA is identic in the group A but different in the group B. All these results can be subsequently ranked to sort those with high chance of functional impacts (see “CDD NCBI”).

CDD NCBI

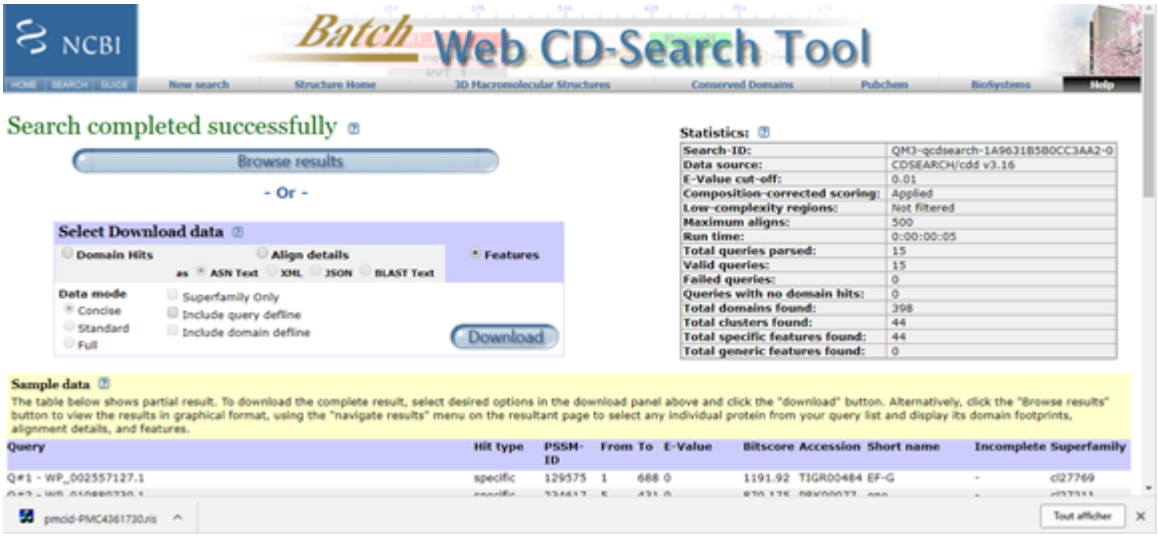
By clicking this option in operation, a window will appear. In the left panel, the user can add the results file obtained before. By clicking the “CDD search” button, CAPRIB compare the results obtained with the position of conserved domains in the CDD NCBI database.



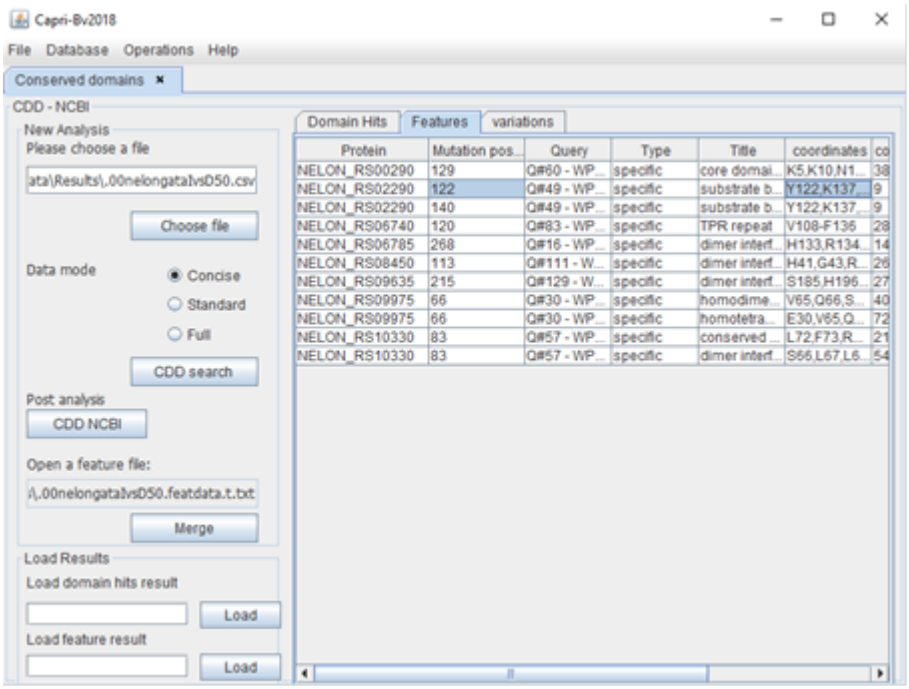
This will produce a merged file that shows only the proteins having variations into a conserved domain.



By clicking the CDD NCBI button the user will be send to the NCBI site where he could download the Features file for the concerned proteins and save it into our “Results” folder.



This file can be used in CAPRIB as shown bellow and by clicking the “Merge” button the program will generate a Feature merged file in the Results folder. This file contains only AA variations that are placed in a conserved CCD domain.



ANNEXE 3

RAPPORTS DE CAPRIB

Information filtrée du BLAST

L'information extraite du rapport BLAST est en format CSV, ce fichier montre les proteines qui surpassent le seuil de pourcentage d'identité avec d'autres informations comme on peut observer dans la **Erreur ! Source du renvoi introuvable.**Figure 29. Ces informations serviront à nourrir la base de données de l'organisme de référence. Il y a un fichier filtré par organisme comparé.

	A	B	C	D	E	F	G	H	I	J	K
1	Proteine	length Sbjct	Identity	similarity	e-value	AA_identi	AA_simila	AA_difere	gaps Quer	gaps Subje	Stop codo
2	NELON_RS10495	118	94	97	9,00E-60	1,2,3,4,5,6	15:Q/K,17	18:F/L,63:'			
3	NELON_RS02780	123	100	100	5,00E-79	1,2,3,4,5,6					
4	NELON_RS01050	143	98	99	6,00E-90	2,3,4,5,6,7	1:M/V,70:	104:A/T			
5	NELON_RS07710	58	91	98	4,00E-34	1,2,4,7,8,9	3:K/Q,5:F/	6:L/F			
6	NELON_RS10500	65	100	100	1,00E-38	1,2,3,4,5,6					
7	NELON_RS03780	76	100	100	2,00E-47	1,2,3,4,5,6					
8	NELON_RS02800	103	100	100	1,00E-65	1,2,3,4,5,6					
9	NELON_RS02310	51	100	100	1,00E-29	1,2,3,4,5,6					
10	NELON_RS06300	521	95	97	0.0	1,2,3,4,5,6		68:T/V,96:			
11	NELON_RS10505	134	99	99	6,00E-86	1,2,3,4,5,6	26:Y/F	116:S/Q			
12	NELON_RS02785	157	95	97	5,00E-95	1,2,3,4,5,6	32:I/V,72:I	51:E/A,54:	54,Q		
13	NELON_RS06395	558	99	99	0.0	1,2,3,4,5,6	158:E/D,3'	161:A/E,16			

Figure 29 Fichier filtré du BLAST.

On observe le Locus Tag de la protéine, la longueur du hit, le pourcentage d'identité, le pourcentage de similarité, la valeur e-value, les positions pour les acides aminés identiques, similaires, différents, gaps dans l'organisme référence, gaps dans l'organisme comparé et finalement codon-stop.

Rapports des opérations de CAPRIB

Les rapports obtenus sont en format CSV et pourront être visualisées en Excel (Figure 30). Ils contiennent des informations comme les organismes de chaque groupe, le nombre de protéines conservées au pourcentage d'identité testée. Le nombre de protéines candidate, c'est-à-dire les protéines avec des mutations qui répondent au critère choisi lors de la combinaison (IvsD, ISvsD, IvsS, GapIn, GapOut et stop codon), le nombre des mutations trouvées. La colonne Protein et Refseq correspondent au numéro d'accession et à la séquence de référence respectivement, à coté on trouve le nom et symbole de la protéine, Variant pos est la position ou les positions qui ont changé pour chaque protéine candidate, dans les autres colonnes on trouve la position de l'acide aminé qui a changé ainsi que la variation et sa distance de Grantham. Dans la partie inferieure on voit pour chaque organisme le nombre de fois qu'on a trouvé la variation dans chaque organisme du groupe B et la distance de Grantham de chaque substitution

Group A:	ecolik12, Kdenitrificans, Kkingae, KOralis, Nbacillif				
Group B:	NcinereaATTC14685, NflavescensSK114, Ngonorroh				
Common proteins:	156				
Proteins candidates:	13				
Mutations found:	25				
Protein	Refseq	Name	Symbol	Variant	NcinereaA N
NELON_RS03495	WP_00377	ornithine	-	[22]	{22=L/Y:36 {2
NELON_RS06795	WP_04066	malic enzy	-	[32, 117	{32=Q/S:6 {3
NELON_RS04540	WP_04066	30S ribosc	rpsA	[270]	{270=A/T:3 {2
NELON_RS01045	WP_00376	transcripti	-	[104]	{104=V/A: {1
NELON_RS02520	WP_00377	2,3,4,5-tet	-	[218]	{218=E/T:6 {2
NELON_RS07225	WP_00377	phosphoe	-	[242]	{242=H/F:3 {2
NELON_RS01310	WP_04066	30S ribosc	-	[42]	{42=R/L:10 {4
NELON_RS02290	WP_00376	molecular	-	[122, 20	{209=V/N: {2
NELON_RS06130	WP_00377	pyruvate d	aceE	[63, 444	{743=A/G: {1
NELON_RS06120	WP_04196	dihydrolip	-	[406]	{406=K/N: {4
NELON_RS07780	WP_04196	elongatio	-	[595]	{595=G/S:3 {5
NELON_RS06110	WP_00377	dihydrolip	-	[206, 37	{385=K/N: {3
NELON_RS09285	WP_00377	serine hyd	glyA	[324]	{324=K/M: {3
Stats: counting variations					
According to Grantham's distance, most similar amino acids are leucine a					
and the most distant are cysteine and tryptophan (value 215)					
NpolysaccharеaATTC	{P/S:74=2, V/A:64=3,			Y/V:55=1, L/Y:36=1, D/Q:61=1,	
NflavescensSK114	{P/S:74=3, V/A:64=2, Y/V:55=1, L/Y:36=1, D/Q:61=1,				

Figure 30 Rapport d'analyses fait par CAPRIB.

Comparaison entre *N. elongata* avec d'autres organismes. L'analyse réalisée est IvsD donc on cherche les acides aminées identiques dans le Groupe A, mais différents dans le groupe B et un pourcentage d'identité minimal de 60%.

Rapport CDD combiné avec les données de CAPRIB

Le rapport obtenu en format TSV, est le résultat de combiner l'information obtenue voie API avec CDD et le rapport de l'analyse fait par CAPRIB. De manière générale CDD (Marchler-Bauer *et al.*, 2017) génère un rapport où par chaque protéine montre les domaines importants et CAPRIB combine l'information et montre seulement les domaines où se trouvent les variations candidates. Par exemple dans la Figure 31, on observe que la mutation à la position 270 de la protéine NELON_RS04540 se présente dans la région entre 1 et 557.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Protein	Mutation	Query	Hit type	PSSM-ID	From	To	E-Value	Bitscore	Accession	Short nam	Incomplet	Superfam
2	NELON_RS04540	270	Q#2 - WP_	specific	235775	1	557	0	908.773	PRK06299	rpsA	-	cl28253
3	NELON_RS02290	122	Q#3 - WP_	specific	236757	1	374	0	645.276	PRK10767	PRK10767	-	cl28246
4	NELON_RS02290	209	Q#3 - WP_	specific	236757	1	374	0	645.276	PRK10767	PRK10767	-	cl28246
5	NELON_RS09285	324	Q#4 - WP_	specific	234571	7	415	0	805.836	PRK00011	glyA	-	cl18945
6	NELON_RS06795	32	Q#5 - WP_	specific	235976	6	757	0	1404.07	PRK07232	PRK07232	-	cl27704
7	NELON_RS06795	117	Q#5 - WP_	specific	235976	6	757	0	1404.07	PRK07232	PRK07232	-	cl27704
8	NELON_RS06795	188	Q#5 - WP_	specific	235976	6	757	0	1404.07	PRK07232	PRK07232	-	cl27704
9	NELON_RS06795	205	Q#5 - WP_	specific	235976	6	757	0	1404.07	PRK07232	PRK07232	-	cl27704
10	NELON_RS06795	388	Q#5 - WP_	specific	235976	6	757	0	1404.07	PRK07232	PRK07232	-	cl27704
11	NELON_RS06795	396	Q#5 - WP_	specific	235976	6	757	0	1404.07	PRK07232	PRK07232	-	cl27704
12	NELON_RS06795	398	Q#5 - WP_	specific	235976	6	757	0	1404.07	PRK07232	PRK07232	-	cl27704
13	NELON_RS03495	22	Q#6 - WP_	specific	179366	1	329	0	626.913	PRK02102	PRK02102	-	cl27385
14	NELON_RS01045	104	Q#7 - WP_	specific	180161	1	176	5.25495e-	294.79	PRK05609	nusG	-	cl28343
15	NELON_RS06130	63	Q#8 - WP_	specific	236500	1	887	0	1880.23	PRK09405	aceE	-	cl27365
16	NELON_RS06130	444	Q#8 - WP_	specific	236500	1	887	0	1880.23	PRK09405	aceE	-	cl27365
17	NELON_RS06130	700	Q#8 - WP_	specific	236500	1	887	0	1880.23	PRK09405	aceE	-	cl27365
18	NELON_RS06130	743	Q#8 - WP_	specific	236500	1	887	0	1880.23	PRK09405	aceE	-	cl27365
19	NELON_RS02520	218	Q#9 - WP_	specific	236996	1	271	0	546.328	PRK11830	dapD	-	cl25958
20	NELON_RS07225	242	Q#10 - WP	specific	235809	1	795	0	1567.43	PRK06464	PRK06464	-	cl27021
21	NELON_RS01310	42	Q#11 - WP	specific	235532	3	85	5.8735e-3	122.185	PRK05610	rpsQ	-	cl00351
22	NELON_RS07780	595	Q#12 - WP	specific	235462	1	597	0	1290.38	PRK05433	PRK05433	-	cl27769
23	NELON_RS06120	406	Q#13 - WP	specific	237000	3	544	0	708.512	PRK11855	PRK11855	-	cl27351
24	NELON_RS06110	206	Q#14 - WP	superfami	332164	133	601	0	673.205	cl27343	Pyr_redox	-	-
25	NELON_RS06110	372	Q#14 - WP	superfami	332164	133	601	0	673.205	cl27343	Pyr_redox	-	-

Figure 31 Rapport combine CAPRIB-CDD.

ANNEXE 4

ALGORITHMES

Filtrer les informations du BLAST

ENTRÉE: seuil inferieur du pourcentage d'identité souhaité (%I), fichier de sortie (output_f), fichier du rapport BLAST à parcourir (input_f).

Imprimer dans output_f l'entête avec les noms des informations à extraire de l'input_f.

Pendant que l'input_f est ouvert, faire :

Pour chaque protéine qui soit égal ou plus grand au seuil %I, extraire:

- Nom de la proteine
- Longueur du hit
- Pourcentage d'identité
- Pourcentage de similarité
- E-value
- Liste de positions des acides aminés identiques.
- Liste de Positions et symboles des acides aminés similaires avec la notation chiffre « caractère /caractère » où le premier caractère est l'acide aminé dans l'organisme de référence et l'autre l'acide aminé dans l'organisme comparé.
- Liste de positions et symboles des acides aminés différents avec la même notation exposé précédemment
- Liste de positions où commence le gap dans la séquence de référence suivi des acides aminés qui ont été insérées dans l'organisme comparé.
- Liste de positions où commence le gap dans la séquence de l'organisme comparé suivi des acides aminés qui ont été perdus dans l'organisme de référence.
- Liste positions où il y a un stop codon dans les organismes comparés.

Imprimer ces informations dans le fichier output_f.

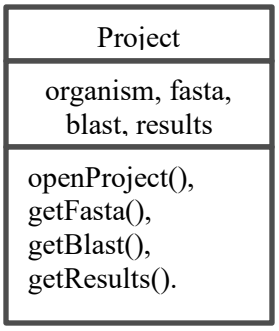
Fermer tous les fichiers.

Operations d'ensembles

Entrée : Listes d'organismes du groupe A et B, opération à réaliser (IvsD, ISvsD, IvsS, GapIn, GapOut ou StopCodon)

- Construire un dictionnaire avec les protéines qui soient présentes dans tous les organismes au seuil de pourcentage d'identité introduit par l'utilisateur.
- Pour chaque groupe on construit un ensemble selon l'opération, par exemple si c'est IvsD on fait un ensemble pour les positions des acides aminés identiques dans le groupe A et un autre ensemble pour les positions différents pour chaque protéine conservée.
- Réaliser l'intersection de deux ensembles

Imprimer le rapport avec les protéines et les mutations trouvées après l'opération d'intersection.



ANNEXE 5

Nouvelle version de l'article soumis à « Bioinformatics ».

OXFORD UNIVERSITY PRESS | Bioinformatics

Capri-B: A user-friendly tool to study amino acid changes and selection for the exploration of intra-genus evolution

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2019-2366
Category:	Original Paper
Date Submitted by the Author:	14-Nov-2019
Complete List of Authors:	Guerra-Maldonado, Juan F.; INRS-Institut Armand-Frappier, INRS-IAF Vincent, Antony; INRS-Institut Armand-Frappier, INRS-IAF Chenal, Martin; INRS-Institut Armand-Frappier, INRS-IAF Veyrier, Frédéric; INRS-Institut Armand-Frappier, INRS-IAF
Keywords:	Molecular evolution, Protein evolution, Microbiology

Capri-B: A user-friendly tool to study amino acid changes and selection for the exploration of intra-genus evolution

Juan F. Guerra Maldonado[1,#], Antony T. Vincent[1,#], Martin Chenal[1], Frederic J. Veyrier[1*]

1. INRS-Centre Armand-Frappier Santé-Biotechnologie, Bacterial Symbionts Evolution, Laval, Quebec, Canada

#. These authors contributed equally to this work.

*** Correspondence:**

Frédéric J. Veyrier, Ph.D.
Institut National de la Recherche Scientifique
(450) 687 5010 # 8831
frederic.veyrier@iaf.inrs.ca
531 Boul. des Prairies
Laval (Québec) H7V1B7 Canada

ABSTRACT

Motivations: Evolution of bacteria is shaped by different mechanisms such as mutations, gene deletion, duplication or insertion of foreign DNA among other things. These genetic changes can accumulate, as a result of natural selection, in the descendants. Using phylogeny and genome comparisons, evolutionary paths can be somehow retraced with recent events that are much easier to detect than older ones. For this reason, multiple tools are available to study the evolution within genomes of single species, such as gene composition, or subtler mutations such as SNPs. However, these tools are generally designed to compare similar genomes and require advanced skills in bioinformatics.

Results: We present CAPRIB, a unique tool developed in java that allows to determine the amino acid changes, at the genus level, that correlates with phenotypic differences between two groups of organisms. This tool has a user-friendly graphical interface and uses a database in SQL, making it easy to compare several genomes without the need for programming. We exemplify the utility of CAPRIB by extracting some of the amino acid changes that coincided with the emergence of slow-growing mycobacteria from fast-growing mycobacteria.

Availability and Implementation: The CAPRIB software and its documentation are freely available for Windows and UNIX-like operating systems at https://github.com/BactSymEvol/Caprib.

Supplementary information: Supplementary data are available at Bioinformatics online.

INTRODUCTION

Bacteria are ubiquitous in nearly any given environment because of their ability to quickly evolve and adapt. As for other living beings, bacteria evolve through genetic changes that allow them to increase their fitness in response to changing environments and hosts. Thanks to recent advances in sequencing technologies (reviewed in van Dijk et al., 2018; Vincent et al., 2017), we are now able to sequence bacterial genomes from diverse ecological niches (De Mandal and Panda 2015; Forde and O'Toole 2013), allowing us to unravel much more clearly than before the evolutionary mechanisms that shaped bacterial adaptation.

Gene deletions and insertions are two major genetic events that have been linked with bacterial evolution. Supporting this statement, it has been shown in the *Neisseriaceae* family that the deletion of a specific gene in a bacilli-shaped ancestor led to the transition to a coccoid shape, which is thought to help immune evasion in the human nasopharynx (Veyrier et al., 2015). On the other hand, horizontal gene transfer (HGT) is a major evolutionary force in prokaryotes through gene acquisition (Wiedenbeck and Cohan 2011). The *Mycobacterium* genus is a perfect evidence of beneficial insertions by HGT, with several genes acquired being necessary for the cell's metabolism and virulence (Becq et al., 2007; Jang et al., 2008; Panda et al., 2018; Veyrier et al., 2009). A recent study suggests that the unique cell envelope of bacteria members of the order *Corynebacteriales*, such as *Mycobacterium tuberculosis*, might be the result of a stepwise acquisition of multiple genes (Vincent et al., 2018), supporting the importance of these events in bacterial evolution. Given the increase of genomic studies, a considerable number of tools have been developed in recent years to investigate the evolution of the gene repertoire, such as MycoHIT (Veyrier et al., 2009), SaturnV (Freschi et al., 2019), GET_HOMOLOGUES (Contreras-Moreira and Vinuesa 2013) and Roary (Page et al., 2015).

Although gene acquisition and deletion can reveal insights on some evolutionary processes in bacteria, they are not sufficient to completely explain bacterial adaptation. Other subtler genetic events, like single nucleotide substitution or single amino-acid (AA) changes, are necessary for evolution. These point mutations arise naturally from replication errors, and are therefore believed to be much more frequent than gene rearrangements (Foster et al., 2015). In the *Mycobacterium tuberculosis* complex, at the species scale, thousands of single-nucleotide polymorphisms (SNPs) have been identified in numerous virulence genes (Mikheecheva et al., 2017). These point mutations have helped establish the evolutionary relationships between several lineages, and were even suggested to be used as broad phylogenetic markers (Coll et al., 2014; Filliol et al., 2006; Foster et al., 2015; Homolka et al., 2012). In addition, multiple studies have also shown the importance of amino acid changes in the evolution of some members of the *M. tuberculosis* complex (such as in RskA (Said-Salim et al., 2006), PhoR (Gonzalo-Asensio et al., 2014) but also during the emergence of antibiotic resistant mutants (Pi et al., 2019; Spies et al., 2011)).

Compared to the study of gene flow, where several bioinformatics tools exist, studying the impact of small modifications is often perilous. Furthermore, the more ancestral these subtle modifications are (such as those that arise at the birth of a specific genus), the more difficult their detection will be due to time effects. This kind of investigation usually involves mapping reads or genomic sequences against a close reference and then looking at the impact on the coding sequences. Tools such as snippy (https://github.com/tseemann/snippy) and snpEff (Cingolani et al., 2012) (used by snippy) allows this analysis efficiently. Other tools, for example kSNP3.0 (Gardner et al., 2015) and the Harvest suite (Treangen et al., 2014), integrate an evolutionary approach involving a phylogenetic reconstruction. Finally, the Kover tool, which uses a machine

learning approach, was developed to find genomic biomarkers explaining different phenotypes without the need for a reference (Drouin et al., 2016). However, these tools make it difficult to quickly analyze different datasets and to infer the impact of the AA changes without advanced expertise in bioinformatics. More importantly, they are impractical to identify SNPs or single amino-acid polymorphisms (SAPs) associated with a specific phenotypic change that occurred within a bacterial genus.

Here, we describe a new bioinformatics tool, CAPRIB (Comparative Analyses of Proteins In Bacteria), which efficiently finds through an easy-to-use graphical interface amino-acid changes that are strictly associated to the emergence of a phenotype. By comparing two groups of species separated by a phenotypic switch, this program can pinpoint the evolutionary events, at the genus scale, that led to this transition. This tool, whose core is in Java, uses a relational database in SQL to store raw information, allowing users to quickly change the structure of groups and save projects. In order to help users decipher the impact of AA changes, CAPRIB integrates statistical values (Grantham's distance (Grantham 1974) and exchangeability score (Yampolsky and Stoltzfus 2005)) that predict the potential impact of one amino acid change for another. It is also possible to put in relation the detected AA changes and the Conserved Domain Database (CDD) of the NCBI to have an insight on the structural involvement of the permutations.

Herein, we are specifically exemplifying the application of CAPRIB using the *Mycobacterium* genus from the *Actinobacteria* phylum. This genus is an ideal candidate to study bacterial evolution through amino-acids changes because of its genotypic and phenotypic diversity. First of all, mycobacteria species can be pathogenic, commensal or saprophytic in a variety of hosts and ecological niches (Malone and Gordon 2017; Tortoli 2014). Also, this genus

was historically divided into two phenotypically-different categories, fast growers and slow growers (Stahl and Urbance 1990) and more recently a third pseudo-intermediate lineage was revealed (Tortoli et al., 2017). These different growth rates are intrinsically linked to the evolution of mycobacteria, and are the basis of their phylogenetic classification (Tortoli et al., 2017). We demonstrate that CAPRIB can detect AA changes associated with a given node of evolution linked to a potential phenotypic switch. As a side application, this tool can also help to question phylogenetic history at the genus scale using these changes as markers as it is usually done with DNA SNPs at the species scale.

MATERIALS AND METHODS

Development of CAPRIB

CAPRIB is developed in Perl, SQL and Java. Perl is used to filter the BLAST reports, but also to communicate with the CDD-NCBI and combine this result with the candidate protein files and ultimately get the report in TSV format. SQL is used to build, manage and access the database. Finally, the core of CAPRIB and its GUI is in Java, mainly using the javax.swing and java.awt graphical libraries. It is designed to work on macOS, GNU/Linux and windows on a standard computer.

CAPRIB uses the results of TBLASTN (which can be generated directly with CAPRIB) in order to determine the similarity links (identical, similar or different) between the amino acids for the homologous proteins of the given dataset. This information is used to generate an SQL database. Subsequently, this database can be queried to perform different operations (Table 1). The CAPRIB software and its documentation are freely available at https://github.com/BactSymEvol/Caprib.

Phylogeny of mycobacteria

A dataset of species of the *Mycobacterium* bacterial genus has been assembled to optimize the quality of genomes and to be representative of species diversity (Supplementary File 1). A phylogenetic tree was created using a bioinformatics protocol published elsewhere (Vincent et al., 2019). Briefly, the 56 genome sequences (including the outgroup) were annotated using Prokka version 1.13.7 (Seemann 2014). Homologous links between the translated coding sequences were found using the combination of the two algorithms COG (Kristensen et al., 2010) and OMCL (Li et al., 2003) through GET_HOMOLOGUES version 20190411 (Contreras-Moreira and Vinuesa 2013). The 958 gene sequences (excluding paralogs) corresponding to the softcore (sequences present in more than 95% of the genomes) were aligned by codons using mafft version 7.407 (Katoh and Standley 2013) through TranslatorX version 1.1 (Abascal et al., 2010). The resulting alignments were filtered using BMGE version 1.12 (Criscuolo and Gribaldo, 2010) and concatenated in a partitioned supermatrix using AMAS (Borowiec, 2016). The evaluation of the best-fit model of each partition and the maximum-likelihood phylogeny were done using IQ-TREE version 1.6.11 (Kalyaanamoorthy et al., 2017; Nguyen et al., 2015). The robustness of the tree was assessed by performing 10,000 ultrafast bootstraps (Hoang et al., 2018).

Analysis with CAPRIB

Two databases were constructed: one using *M. tuberculosis* H37Rv (slow growing) and the other with *M. gilvum* Spyr1 (fast growing) as the reference. The comparisons were generated using TBLASTN version 2.9.0+ (Altschul et al., 1997). The same 56 genomes utilized for the presented phylogeny were also used but *M. leprea* and *M. lepraemurium* were excluded since they have a specific well-described evolution by reductive genomics and gene decay with

respectively 41% (Cole et al., 2001) and 30% pseudogenes (Benjak et al., 2017). The identity threshold for considering two sequences as homologues was determined by calculating the 10th percentile median for all BLAST results for a given reference. This permitted determining a minimum of 40% for *M. tuberculosis* H37Rv and 35% for *M. gilvum* Spyr1. The functional categories for some proteins of interest were determined using the eggNOG 5.0.0 database (Huerta-Cepas et al., 2019).

RESULTS

Description of CAPRIB main functions:

In order to investigate the amino acid changes that may be involved in the evolution of a group of organisms and in the emergence of a new phenotype (see Figure 1A), we have created CAPRIB, a tool with an easy-to-use graphical interface (Figure 1B). To make CAPRIB accessible to a maximum number of users, it only requires some dependencies that are often already installed on biologist computers (java, MySQL, PERL and BLAST+). It has also been designed to work on most operating systems (Windows, macOS and GNU/Linux). The architecture of CAPRIB, relying on relational databases (Figure 1C), allows users to create projects and to change parameters quickly for a given database, without having to redo the BLAST searches. A summary of the results can be visualized directly through CAPRIB while complete detailed results are available in a CSV file, compatible with spreadsheet tools. To guide users through all of the AA changes found by CAPRIB, the latter integrates an interface allowing to check if the permutations are in conserved domains according to the NCBI CDD database. In addition, scores (Grantham's distance (Grantham 1974) and exchangeability score (Yampolsky and Stoltzfus 2005)) are associated with each of the AA changes which helps assessing the potential impact of the permutations on the structure of the protein.

The SQL database: The user is performing TBLASTN (with BLAST+ using the JAVA interface) of protein sequences from reference (herein *M. tuberculosis* H37Rv a slow-growing species and *M. gilvum* Spyr1 a rapid-growing bacteria) against genomic sequences of other species from the genus available (herein 51 mycobacteria species and the three outgroup bacteria). Subsequently, the user is using the results of these comparisons to feed the database. The database is structured as shown in Figure 1C with a table of general features from the reference organism (such as name of proteins, locus tags, accession numbers), a second table with the blast information (query and subject names) and a third table that comprises the information extracted from the BLAST file. For this latter, Perl is used to parse and extract results from BLAST report. The originality of this tool is that it uses the information contained in the BLAST results and classifies each compared AA as identical, similar, or different in regard to the reference in a third SQL table (see Figure 1C). CAPRIB also stores information of gaps (insertion of AA or deletion of AA) and AA that changed to stop codon. This unique property allows to store information of comparison of all proteins from the reference against all genomic sequences using minimal computational requirement and thus to use standard computers.

Comparisons: The user can query the database according to different evolutionary strategies (Table 1). The first, I VS D, identifies positions with identical AAs in group A, but different in group B. The main objective of this strategy is to find positions with high conservation pressure in group A, and with a relaxed pressure in group B. The 2nd strategy, IS VS D, is less stringent than the first one and allows including positions with similar AAs in group A. The 3rd strategy, I VS S, makes it possible to identify the slightly more subtle differences on the protein structure since it indicates the positions with identical AAs in group A and similar to group B. The 4th and 5th strategies allow finding the gaps preserved in group A (GapIn) or B

(GapOut), respectively. Finally, the sixth strategy (Stop codon) makes it possible to identify stop codons present exclusively in group A, and thus to infer the potential truncations of proteins.

Tools to infer severity of AA changes: Once the user has performed comparisons, some tools are integrated to facilitate the classification of AA change in function of their predicted potential effect. We have first used two scores that are predictors of the effect of the substitution of one AA to another. The Grantham distance (Grantham 1974) that takes into account three parameters of AA difference such as composition, polarity, and molecular volume to compare amino acids combines properties. This table of one-by-one AA comparison correlates with protein residue substitution frequencies (Grantham 1974). We have also used a study of Yampolsy et al. (2005) that has experimentally measured the Experimental Exchangeability (EX) of each AA by another one by replacing around 10,000 AA in 12 proteins (Yampolsky and Stoltzfus 2005) which also generated a table of scores for AA exchangeability. Of note, in this case the score is inversely proportional to the ease of exchangeability. They have also measured the overall scores of specific AA exchangeability (Exchangeability as a source: EXsrc or as destination: EXdest). In CAPRIB, these scores are indicated in the results file next to the AA change (ex: 75=A/H:86:301:EXsrc=312:EXdest=290, herein the AA number 75, which is a Alanine in group A, is replaced by a Histidine in this query species from group B. This change has a Grantham score of 86 and an Ex score of 301. The EXsrc value of the A and the EXdest of the H are also indicated).

We have also implemented an option that links the result file with CDD (Conserved Domaine Database) of NCBI (Marchler-Bauer et al., 2017) as described in the help of CDD (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml#BatchRPSBWebAPI_GETorPOST). We have used the class CddNcbi which starts the script ccd.pl, collects results and fuses them

with the list of AA changes to produce a TSV file which can be open in CAPRIB or directly online in CDD website. With this option the users can verify if permutated AAs are located in specific positions of the proteins, such as conserved domains.

Use of CAPRIB to monitor the phylogenetic landscape and assess putative abnormality.

Mycobacteria were historically separated into two large phylogenetic groups that were correlated with their growth rate (Stahl and Urbance 1990): the slow-growing and fast-growing lineages. Recently, a third group (*M. terrae* complex) has been revealed to be intermediate (in terms of phylogeny but also in terms of phenotypes) between the slow and fast-growing mycobacteria (Fedrizzi et al., 2017; Tortoli et al., 2017). The fact that mycobacteria can be separated phylogenetically based on a phenotype (growth rate) makes it a model of choice to be investigated with CAPRIB. Two databases were constructed, one using *M. tuberculosis* H37Rv (slow grower) and the other with *M. gilvum* Spyr1 (fast grower) as references, using 53 genomes of mycobacteria with genomes of good quality (Supplementary File 1). We present in Figure 2 the phylogeny of this dataset, which is in accordance with previous phylogenies done using other datasets (Fedrizzi et al., 2017; Tortoli et al., 2017).

Defining groups is crucial in order to achieve optimal results with CAPRIB. Of note, gold standard tools to infer phylogenetic reconstruction by maximum-likelihood, such IQ-TREE (Nguyen et al., 2015) and RAxML (Kozlov et al., 2019; Stamatakis 2014) could eventually bias grouping as they force strict bifurcating trees and do not allow polytomy. The fact that *M. terrae* clade species are considered as intermediates between slow and fast growers can cause a problem in the definition of groups and thus in the different AA changes found to explain the phenotype (i.e, growth difference). The molecular phylogeny carried out for the present study reveals the grouping of this clade among the slow growers, a result also found by several other studies

(Gupta et al., 2018; Rogall et al., 1990; Tortoli et al., 2017). However, the position of this clade has already been shown to be unstable by a seven genes multilocus-based phylogenetic study (Mignard and Flandrois 2008). Since CAPRIB can find markers specific to different groups, we challenged the phylogenetic position not only for this clade but also other nodes as seen in Figure 3. The number of markers was determined using CAPRIB for the three different possible topologies (Figure 3). Intriguingly, the topology of node 3 obtained by molecular phylogeny in figure 2, namely that the *M. terrae*-clade cluster with the slow growers, is the one with the least markers (366 conserved permutations)(Figure 3C). On the contrary, the dominant topology clusters the *M. terrae*-clade with the fast growers (872 conserved permutations). Since the expected number of AA changes is not in accordance with the molecular phylogeny at node 3, it was interesting to check the number of permutations supporting the next node, labelled 4, in Figure 2 (divergence between slow-growers and clade *M. terrae* Figure 3D). This time, the topology with the highest number of markers is that supporting slow-growers as being monophyletic and sharing a common ancestor with the *M. terrae* clade, as in the molecular phylogeny (Figure 2). The fact that the present study arrives at ambiguous results concerning the positioning of the *M. terrae*-clade at node 3, despite the use of complete genomes and rigorous methods using several markers, highlights the evolutionary complexity of mycobacteria.

To demonstrate that this issue was specific to node 3, we also challenged other nodes of evolution. For example, the rapid-growing *M. abscessus*-clade has the most basal position among mycobacteria. Similar to the investigation done for the intermediate clade of *M. terrae*, CAPRIB was used to find the number of markers according to different topologies involving the *M. abscessus* clade (Figure 3B). This time, CAPRIB corroborates the node 2 obtained with the molecular phylogeny, namely that the topology placing the *M. abscessus* clade basal to other

mycobacteria has the most markers (5602 conserved mutations). Other topologies of node 1 were investigated to assess the bifurcation of *M. abscessus*-clade with outgroup species composed of other genera from the *Corynebacteriales* order (Figure 3A). This analysis confirms that the outgroup is basal as compared to *M. abscessus*-clade with other mycobacteria (1744 conserved mutations) while alternative topologies have fewer markers (562 and 581 conserved permutations). Finally, we tested if CAPRIB AA markers can define the phylogenetic link between single species. To do this, we determined the optimal topology between *M. tuberculosis*, *M. marinum* and *M. kansasii* (node 5). Again the topology that shows that *M. marinum* and *M. kansasii* have a common ancestor after *M. tuberculosis* ancestor separation has the most markers which is also supported by molecular phylogeny (Figure 3E).

Specific example showing the macroevolution from fast to slow-growing mycobacteria

In the context of incertitude concerning the *M. terrae*'s clade evolution (node 3) and in the light of their intermediate growth phenotype, we excluded them from the analysis. Importantly, AA changes found by excluding this clade include all possible evolutionary scenarios. Also *M. leprea* and *M. lepraemurium* were excluded since their genomes harbor high number of pseudogenes (Benjak et al., 2017; Cole et al., 2001). Using a cutoff of 40% identity, 1773 genes common to all genomes were found using *M. tuberculosis* H37Rv as a reference (Supplementary File 2) and 2122 using *M. gilvum* Spyr1 with a cutoff of 35% identity (Supplementary File 3).

Using the slow-growing species *M. tuberculosis* H37Rv, CAPRIB has allowed us to identify 1462 conserved amino acids that are different in fast-growing species. These AA changes are distributed among 709 of the 1773 genes found in *M. tuberculosis* H37Rv (Supplementary File 2). Considering AA changes with a Grantham's distance greater than 100 and an exchangeability score of less than 250, only 185 (~ 12.6%) have a high chance of altering

protein function, the rest being more conservative changes (such as A/G, G/A, V/A, A/V that count for 15%). Similarly, using the fast-growing species *M. gilvum* Spyr1, 1092 AA changes were found to be distributed among 567 proteins. Of these changes, 129 (~11.8 %) have a Grantham's distance greater than 100 and an exchangeability score of less than 250.

The lists of mutations found using *M. tuberculosis* H37Rv and *M. gilvum* Spyr1, and having a Grantham's distance greater than 100 and an exchangeability score of less than 250, were crossed to find the different permutations between slow and fast growers, but that are identical within the groups, such as schematized in Figure 1. This highly stringent analysis permitted to find 30 mutations having a high conservative pressure in both groups but that are drastically different between groups (Figure 4A). We later investigated the biological functions of the proteins containing the mutations and the putative pathway that could link them through a STRING analyses (Figure 4B). It was interesting to note that this analysis can highlight some hotspots or pathways that could have evolved in the ancestor of slow growing mycobacteria after its divergence with rapid growing mycobacteria, such as the two proteins GlnE and GlnA1 encoded by neighbor genes or the PonA1 and PonA2 and WhiB4 proteins (with PonA2 and WhiB4 encoded by neighbor genes). In addition to these proteins, it was possible to shed light on three proteins (Rv2727c:MiaA; Rv1205:LOG; Rv2097c:PafA) involved in cytokinin production (Samanovic et al., 2015), a molecule related to phytohormone influencing plant growth and development (Werner et al., 2001). Again, these three genes are often found in the same locus (Naseem et al., 2015). Interestingly, using CDD link within CAPRIB, we realized that a recent study solved the structure of the LOG protein from *Corynebacterium glutamicum* and found that the equivalent residue is involved in AMP binding (Seo et al., 2016). The fact that some hotspots

of evolution could be revealed emphasizes the strategic fit of the methodology developed and that it could be applied in multiple evolutionary contexts.

DISCUSSION

The stepwise adaptation of bacteria occurs through genetic alterations, in which only permissive changes are selected at each of these steps. By deciphering the adaptive mechanisms or pathogenesis emergence of these bacteria, we will not only be able to find crucial information on the bacterial physiology, but also on approaches for the treatment and diagnosis of infections. The high availability of sequencing technologies now allows genomes to be investigated on an unprecedented scale. Numerous tools focus on the detection of genetic events just before the speciation of a pathogen, leaving behind step-wise ancestral events at different nodes of evolution (including the ones not directly linked to pathogens speciation) that have drastic consequences on the pathogens as we know them today (what could be called the “butterfly effect”). As an easy illustration of this concept, we could cite the mycomembrane, which has evolved long before the host-adaptation of mycobacteria that is playing, now, a key role in *M. tuberculosis* pathogenesis (Forrellad et al., 2013). The CAPRIB tool, described in this study, allows users to extract, at the genus scale, drastic AA changes that are concomitant to a phenotypic switch between two bacterial groups. Rich from this knowledge, biologists could now focus their attention on these key residues and on the impact of these changes using molecular microbiology. One of CAPRIB's goals is to make data analysis simple, so it integrates a user-friendly graphical interface and relies on an SQL relational database, which allows comparisons between genomes to be saved. In addition, to maximize accessibility to CAPRIB, it has been designed to work on the majority of operating systems (Windows and UNIX-like).

During our software optimization, we realized that several AA changes in many proteins were detected but the majority of them were conservative. This is why CAPRIB integrates two scores (Grantham's distance and exchangeability score) for each amino acid change, allowing the user to infer the potential impact of the detected changes. In addition, CAPRIB allows to connect the list of AA changes with the NCBI CDD database, making it possible to check whether the permutations are in conserved domains. However, no absolute rule can infer with certainty the impact that a mutation may have on the functionality of a protein and on the network of interaction with other biological molecules. Comprehensive bioinformatics analyses, such as protein modeling and molecular dynamics, can help to verify the impact of mutations with greater certainty.

The study of AA polymorphism profiles with CAPRIB requires some essential prerequisites in order to obtain optimal results. Among these, there is, of course, the quality of the dataset. The more complete a dataset is, which should be representative of the diversity of the bacterial groups studied, the more powerful the results will be. This obviously includes having a robust molecular phylogeny to properly assign the species studied to different groups. This is why a core-genome approach, with several phylogenetic markers, has been favored in this study compared to a 16S phylogeny, that is faster to realize, but often less accurate (Janda and Abbott 2007). Moreover, since CAPRIB makes it possible to analyze the AA changes finely, it goes without saying that it is essential to have sequences of good quality. Fortunately, sequencing errors are often random and the allocation of bases is getting better. Also, it is ideal to compare groups with similar patristic distances (same length of branches in phylogeny). Although not essential, this makes it possible to minimize the number of false positives. Since CAPRIB finds permutations that are identical in one group and different in another, it goes without saying that if

the organisms of the first group are very close to each other, few mutations will have occurred, so it will be difficult to determine positions with a real pressure of conservation.

In this study, we used CAPRIB to shed light on some of the genetic events that could have participated in the emergence of the slow-growing lineage that comprises the majority of pathogenic *Mycobacterium*. Of note, other events such as gene insertions or deletions may also have influenced this process, but this was not the scope of CAPRIB as this has already been investigated (Veyrier et al., 2009; Veyrier et al., 2011; Wang et al., 2015). In the course of our study, we also realized that CAPRIB could be used to question phylogeny reconstruction similarly than what has been done with DNA SNPs to resolve species phylogeny. Our results concerning AA changes suggest that the divergence of the slow, rapid and intermediate lineages is not clear as already mentioned in the literature (Fedrizzi et al., 2017; Tortoli et al., 2017). Even if the first aim of CAPRIB was not this, it can be used as a tool to challenge evolutionary scenarios that are often strictly bifurcating (which could be false due to polytomy with resultant daughter species equidistant from each other, organism hybridization, transfer of genetic material). We acknowledge that the divergence of the three clades could have occurred almost simultaneously (polytomy), as compared to sequentially, and/or that the ancestor of slow-growers has subsequently experienced a high-rate of AA changes. On the contrary, using CAPRIB, we could add strength to the published hypothesis that the last common ancestor of all mycobacteria was a rapid growing species that gave birth to the *M. abscessus* clade and another lineage that later split into rapid and slow growing mycobacteria. When the ratio of the AA changes obtained by comparing an alternative scenario with the phylogeny-based scenario is less than 0.5 (e.i. more AA changes detected), the user needs to question and investigate deeper the evolutionary scenario at this specific node.

On the other side, it is interesting to note that several AA changes in proteins involved in the bacterial membrane have been found. This is consistent with the fact that this organelle is at the boundary with the environment / host and is therefore an evolutionary hotspot. This is especially true in mycobacteria that have a unique cell envelope among bacteria, the genesis of which is still debated (Vincent et al., 2018). AA permutations in several regulators have also been observed. These proteins have the ability to change the level of expression of several genes and thus have the potential to greatly impact global protein networks, resulting in important phenotypic differences. Finally, we shed light on the evolution of several hotspots such as the mycobacterial cytokinin pathway, concomitant to the emergence of slow-growing mycobacteria. There are only few studies on the role of this family of molecules that is similar to a phytohormone influencing plant growth and development (Werner et al., 2001). Some have shown that it can influence signaling in *M. tuberculosis* (Samanovic et al., 2018) whereas others have discovered that cytokinin accumulation is conditionally deleterious as it can lead to an aldehyde breakdown product that kills mycobacteria in the presence of nitric oxide produced by macrophages (Samanovic et al., 2015). It remains to be tested if evolution has sacrificed, in a biological tradeoff, the growth speed of the slow-growing mycobacteria by altering, for example, the cytokinin pathway, in order to allow a better survival in macrophages or other type of macrophage-like cells such as amebae.

In conclusion, we created a new bioinformatics tool, CAPRIB, which can identify key amino acid changes strictly associated with a given node of evolution that correlate phenotypic divergence at the genus scale. This can be applied for numerous studies driven by an evolutionary approach such as antibiotic resistance acquisition, cell-shape change or pathogenesis emergence among others. To highlight the usefulness of this software, we performed a stringent analysis

(non-exhaustive) to pinpoint some of the AA changes that are concomitant with the slow-growing lineage divergence inside the *Mycobacterium* genus. This tool could also potentially identify certain key proteins in different biological processes whose functions could then be validated experimentally. As previously done with *M. tuberculosis*, the current approach to assign biological functions of proteins usually involves the generation of a transposon random insertion mutant library, followed by screening to identify genes, phenotypes and essentiality (Coulombe et al., 2009; DeJesus et al., 2017; Sassetti et al., 2001, 2003). Although this approach allows to determine the functions of proteins, it remains very laborious. As shown by the present study, CAPRIB provides a more targeted approach to identify certain candidate proteins or pathways that have been intensively reworked during evolution at the genus scale, and that would justify efforts for experimental studies.

ACKNOWLEDGEMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN-2016-04940 and by Institut Pasteur through grants PTR 30-2017 and PTR 73-2017. ATV received a Postdoctoral Fellowship from the NSERC. FJV is a research scholar of the Fonds de Recherche du Québec – Santé.

Figure 1. (A) Schematization of the evolutionary approach used by CAPRIB. (B) CAPRIB main interface showing the possibility of defining groups and changing parameters. (C) Diagram of CAPRIB workflow and SQL database structure.

Figure 2. A phylogenetic tree based on softcore sequences, as described in the "Material and Methods" section. Clades containing the majority of slow, fast and intermediate growth species are represented in red, blue and purple, respectively. The *abscessus*-clade is represented in orange. The different organisms were also classified according to a study that separated the genus *Mycobacterium* into five genera (Gupta et al., 2018). The different nodes specifically investigated by this study are indicated on the tree. The set of bootstraps values are at 100, with the exception of an internal node in the MTBC that is at 76.

Figure 3. Analyzes of different phylogenetic topologies of mycobacteria using CAPRIB. The column "Phylogeny-based" is established via Figure 2, while the other two columns represent alternative topologies. For each of the topologies, the number of core proteins, the number of proteins with AA changes, the number of AA changes, and finally the AA changes ratio of the phylogeny-based topology to that analyzed are given. Results in bold indicate the topology with the most markers for each group of trees.

Figure 4. (A) AA changes between the slow and fast-growing mycobacteria, but conserved in each group. (B) STRING analysis showing putative relationships between proteins from the “A” panel.

Table 1. Operations available in CAPRIB

Operation	Definition	Set[a]
I vs D	Identical amino acids in group A that changed to a different amino acid in group B	$I_A \cap D_B$
IS vs D	Identical or similar amino acids in group A that changed to a different amino acid in group B	$(I_A \cup S_A) \cap D_B$
I vs S	Identical amino acids in group A that changed to a similar amino acid in group B	$I_A \cap S_B$
GapIn	Gaps conserved in group A	$N_A \cap (S_B \cup D_A)$
GapOut	Gaps conserved in group B	$(I_A \cup S_A \cup D_A) \cap O_B$
Stop codon	Amino acids in group A replaced by a stop codon in group B	$(I_A \cup S_A \cup D_A) \cap P_B$

a : sets of amino acid positions that are Identical (I), similar (S), different (D), starting position of a gap in the reference organism (N), starting position of a gap in the compared organism (O) and the position of a stop codon (P).

REFERENCES

Abascal F., Zardoya R., Telford M.J. (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.*, **38**, W7-13.

Altschul S.F., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.

Becq J., et al. (2007) Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. *Mol Biol Evol.*, **24**, 1861-1871.

Benjak A., et al. (2017) Insights from the genome sequence of *Mycobacterium lepraemurium*: Massive gene decay and reductive evolution. *MBio*, **8**, e01283.

Cingolani P., et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80-92.

Cole S.T., et al. (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007-1011.

Coll F., et al. (2014) A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun.*, **5**, 4812-4812.

Contreras-Moreira B., Vinuesa P. (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol.*, **79**, 7696-7701.

Coulombe F., et al. (2009) Increased NOD2-mediated recognition of N-glycolyl muramyl dipeptide. *J Exp Med.*, **206**, 1709-1716.

De Mandal S., Panda A. (2015) Microbial ecology in the era of next generation sequencing. *Next Generat Sequenc & Applic.*, **S1**, 001

DeJesus M.A., et al. (2017) Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. *MBio*, **8**, e02133-16

Drouin A., et al. (2016) Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, **17**, 754.

Fedrizzi T., et al. (2017) Genomic characterization of nontuberculous mycobacteria. *Sci Rep.*, **7**, 45258.

Filliol I., et al. (2006) Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol.*, **188**, 759-772.

Forde B.M., O'Toole P.W. (2013) Next-generation sequencing technologies and their impact on microbial genomics. *Brief Funct Genomics.*, **12**, 440-453.

Forrellad M.A., et al. (2013) Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence*, **4**, 3-66.

Foster P.L., Lee H., Popodi E., Townes J.P., Tang H. (2015) Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. *Proc Natl Acad Sci U S A.*, **112**, E5990.

Freschi L., et al. (2019) The *Pseudomonas aeruginosa* pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. *Genome Biol Evol.*, **11**, 109-120.

Gardner S.N., Slezak T., Hall B.G. (2015) kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics*, **31**, 2877-2878.

Gonzalo-Asensio J., et al. (2014) Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc Natl Acad Sci U S A.*, **111**, 11491-11496.

Grantham R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862-864.

Gupta R.S., Lo B., Son J. (2018) Phylogenomics and comparative genomic studies robustly support division of the genus *Mycobacterium* into an emended genus *Mycobacterium* and four novel genera. *Front Microbiol.*, **9**, 67.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. (2018) UFBoot2: Improving the ultrafast bootstrap approximation. *Mol Biol Evol.*, **35**, 518-522.

Homolka S., et al. (2012) High resolution discrimination of clinical *Mycobacterium tuberculosis* complex strains based on single nucleotide polymorphisms. *PloS one*, **7**, e39855.

Huerta-Cepas J., et al. (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309-D314.

Janda J.M., Abbott S.L. (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol.*, **45**, 2761-2764.

Jang J., Becq J., Gicquel B., Deschavanne P., Neyrolles O. (2008) Horizontally acquired genomic islands in the tubercle bacilli. *Trends Microbiol.*, **16**, 303-308.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods.*, **14**, 587-589.

Katoh K., Standley D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.*, **30**, 772-780.

Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, **35**, 4453-4455.

Kristensen D.M., et al. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, **26**, 1481-1487.

Li L., Stoeckert C.J., Jr., Roos D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178-2189.

Malone K.M., Gordon S.V. (2017) *Mycobacterium tuberculosis* complex members adapted to wild and domestic animals. *Adv Exp Med Biol.*, **1019**, 135-154.

Marchler-Bauer A., et al. (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.*, **45**, D200-D203.

Mignard S., Flandrois J.P. (2008) A seven-gene, multilocus, genus-wide approach to the phylogeny of mycobacteria using supertrees. *Int J Syst Evol Microbiol.*, **58**, 1432-1441.

Mikheecheva N.E., Zaychikova M.V., Melerzanov A.V., Danilenko V.N. (2017) A nonsynonymous SNP catalog of *Mycobacterium tuberculosis* virulence genes and its use for detecting new potentially virulent sublineages. *Genome Biol Evol.*, **9**, 887-899.

Naseem M., Sarukhanyan E., Dandekar T. (2015) LONELY-GUY knocks every door: Crosskingdom microbial pathogenesis. *Trends Plant Sci.*, **20**, 781-783.

Nguyen L.T., Schmidt H.A., von Haeseler A., Minh B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.*, **32**, 268-274.

Page A.J., et al. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691-3693.

Panda A., Drancourt M., Tuller T., Pontarotti P. (2018) Genome-wide analysis of horizontally acquired genes in the genus *Mycobacterium*. *Sci Rep*., **8**, 14817-14817.

Pi R., Liu Q., Jiang Q., Gao Q. (2019) Characterization of linezolid-resistance-associated mutations in *Mycobacterium tuberculosis* through WGS. *J Antimicrob Chemother.*, **74**, 1795-1798.

Rogall T., Wolters J., Flohr T., Bottger E.C. (1990) Towards a phylogeny and definition of species at the molecular level within the genus *Mycobacterium*. *Int J Syst Bacteriol.*, **40**, 323-330.

Said-Salim B., Mostowy S., Kristof A.S., Behr M.A. (2006) Mutations in *Mycobacterium tuberculosis* Rv0444c, the gene encoding anti-SigK, explain high level expression of MPB70 and MPB83 in *Mycobacterium bovis*. *Mol Microbiol.*, **62**, 1251-1263.

Samanovic M.I., et al. (2018) Cytokinin signaling in *Mycobacterium tuberculosis*. *MBio*, **9**, e00989-18.

Samanovic M.I., et al. (2015) Proteasomal control of cytokinin synthesis protects *Mycobacterium tuberculosis* against nitric oxide. *Mol Cell.*, **57**, 984-994.

Sassetti C.M., Boyd D.H., Rubin E.J. (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc Natl Acad Sci U S A.*, **98**, 12712-12717.

Sassetti C.M., Boyd D.H., Rubin E.J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol.*, **48**, 77-84.

Seemann T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068-2069.

Seo H., et al. (2016) Structural basis for cytokinin production by LOG from *Corynebacterium glutamicum*. *Sci Rep.*, **6**, 31390.

Spies F.S., et al. (2011) Streptomycin resistance and lineage-specific polymorphisms in *Mycobacterium tuberculosis gidB* gene. *J Clin Microbiol.*, **49**, 2625-2630.

Stahl D.A., Urbance J.W. (1990) The division between fast- and slow-growing species corresponds to natural relationships among the mycobacteria. *J Bacteriol.*, **172**, 116-124.

Stamatakis A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312-1313.

Tortoli E. (2014) Microbiological features and clinical relevance of new species of the genus *Mycobacterium*. *Clin Microbiol Rev.*, **27**, 727-752.

Tortoli E., et al. (2017) The new phylogeny of the genus *Mycobacterium*: The old and the news. *Infect Genet Evol.*, **56**, 19-25.

Treangen T.J., Ondov B.D., Koren S., Phillippy A.M. (2014) The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.*, **15**, 524.

van Dijk E.L., Jaszczyszyn Y., Naquin D., Thermes C. (2018) The third revolution in sequencing technology. *Trends Genet.*, **34**, 666-681.

Veyrier F., Pletzer D., Turenne C., Behr M.A. (2009) Phylogenetic detection of horizontal gene transfer during the step-wise genesis of *Mycobacterium tuberculosis*. *BMC Evol Biol.*, **9**, 196.

Veyrier F.J., et al. (2015) Common cell shape evolution of two nasopharyngeal pathogens. *PLoS Genet.*, **11**, e1005338.

Veyrier F.J., Dufort A., Behr M.A. (2011) The rise and fall of the *Mycobacterium tuberculosis* genome. *Trends Microbiol.*, **19**, 156-161.

Vincent A.T., Derome N., Boyle B., Culley A.I., Charette S.J. (2017) Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. *J Microbiol Methods.*, **138**, 60-71.

Vincent A.T., et al. (2019) Investigation of the virulence and genomics of *Aeromonas salmonicida* strains isolated from human patients. *Infect Genet Evol.*, **68**, 1-9.

Vincent A.T., et al. (2018) The mycobacterial cell envelope: A relict from the past or the result of recent evolution? *Front Microbiol.*, **9**, 2341.

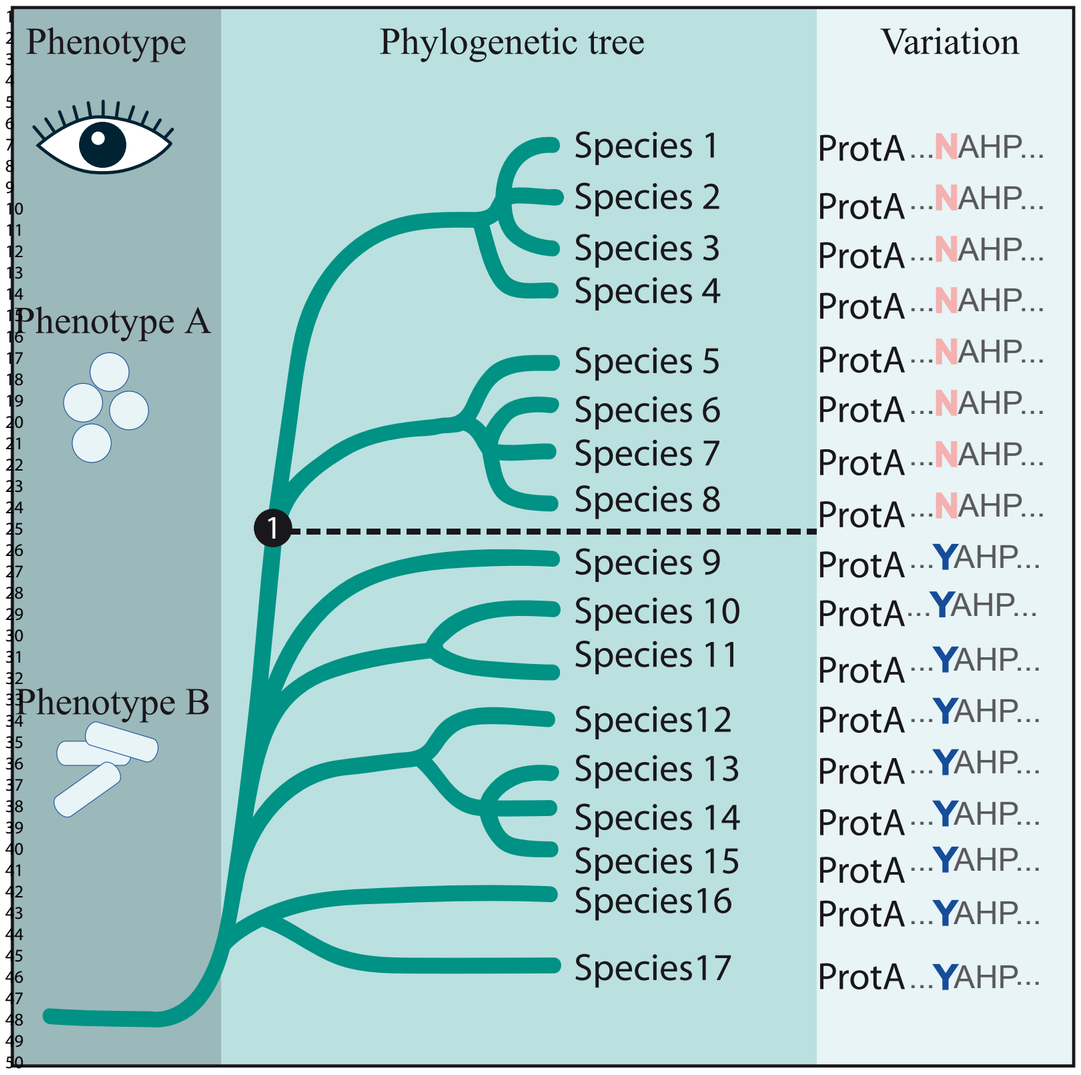
Wang J., et al. (2015) Insights on the emergence of *Mycobacterium tuberculosis* from the analysis of *Mycobacterium kansasii*. *Genome Biol Evol.*, **7**, 856-870.

Werner T., Motyka V., Strnad M., Schmulling T. (2001) Regulation of plant growth by cytokinin. *Proc Natl Acad Sci U S A.*, **98**, 10487-10492.

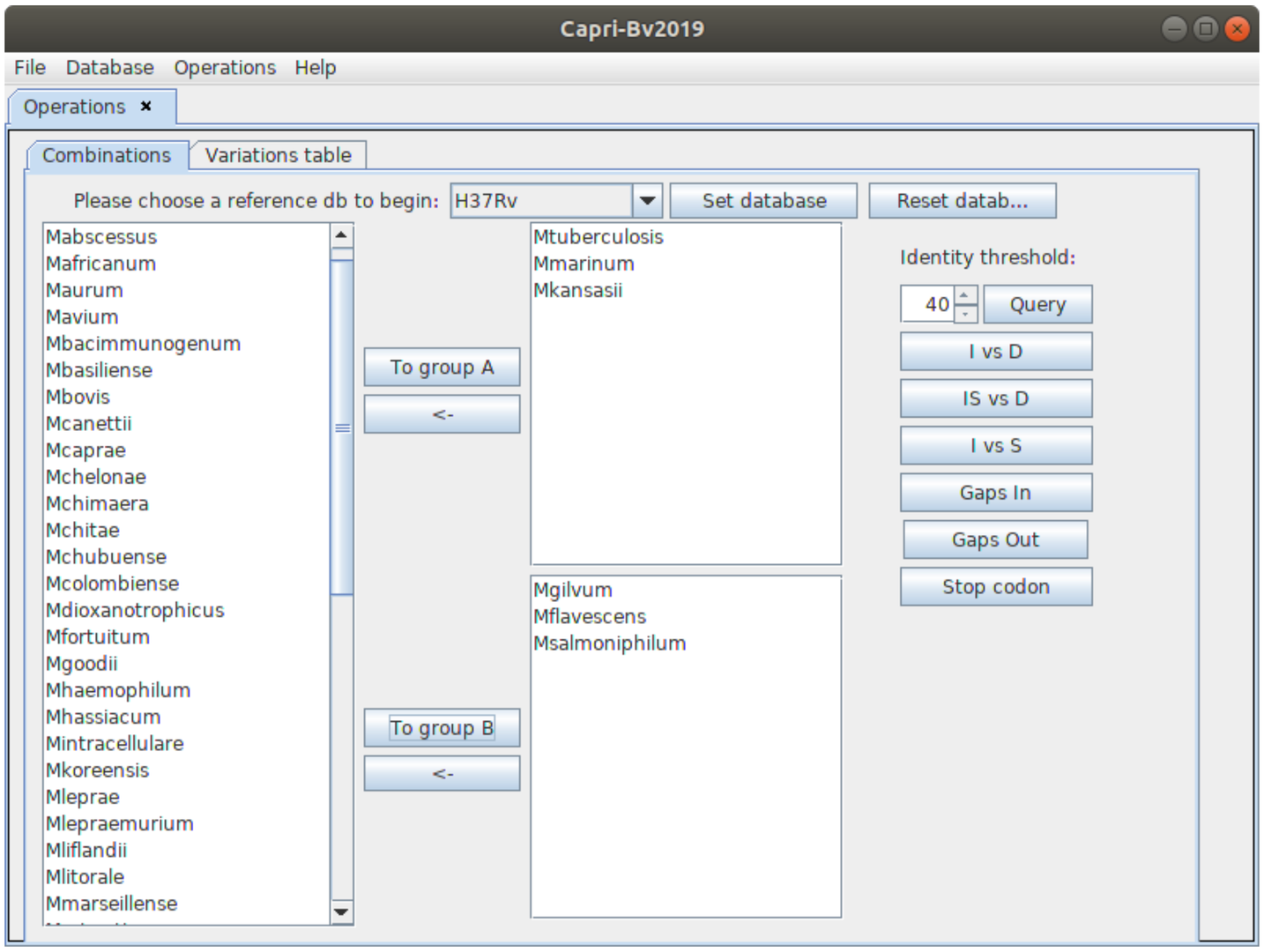
Wiedenbeck J., Cohan F.M. (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev.*, **35**, 957-976.

Yampolsky L.Y., Stoltzfus A. (2005) The exchangeability of amino acids in proteins. *Genetics*, **170**, 1459-1472.

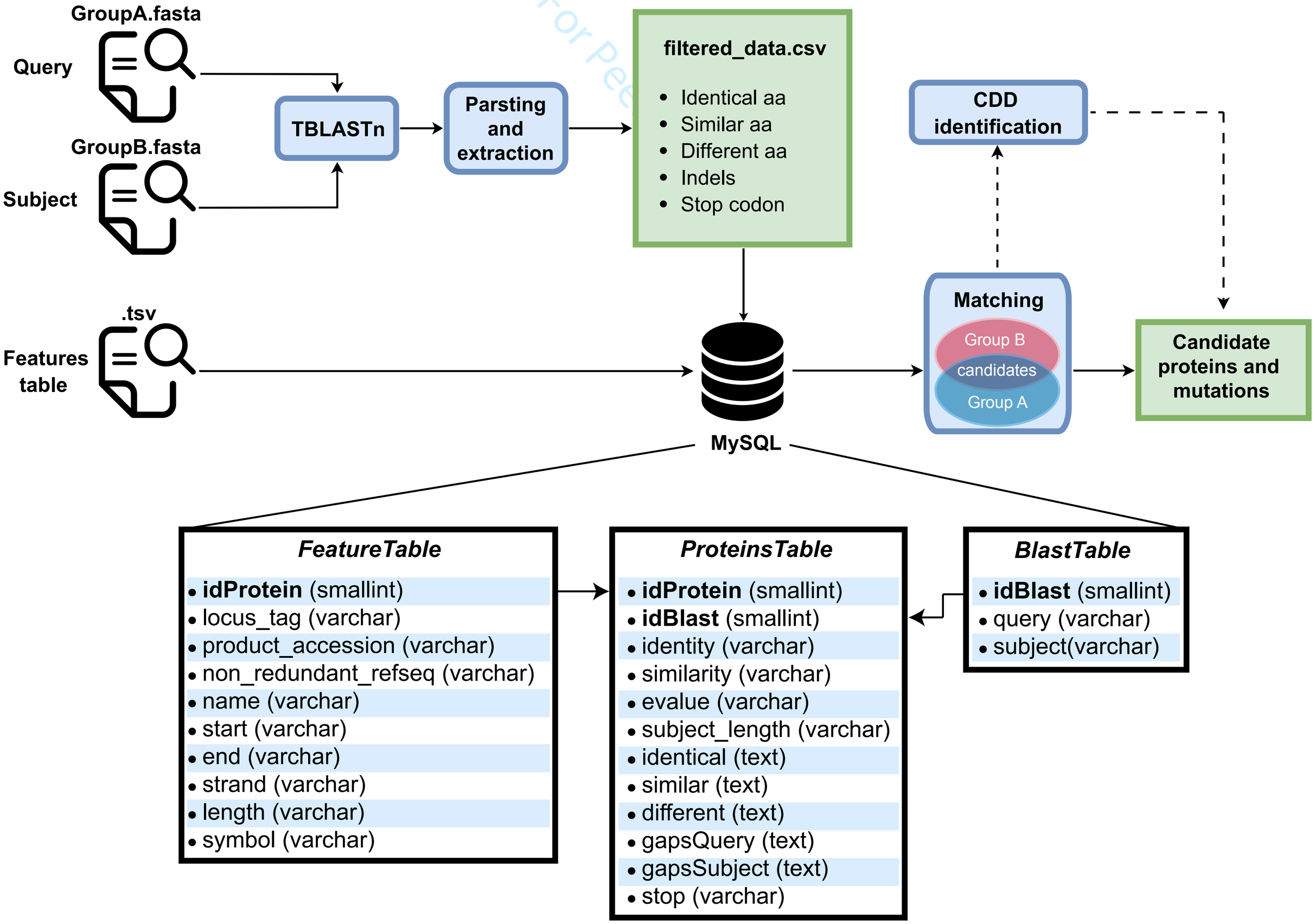
A

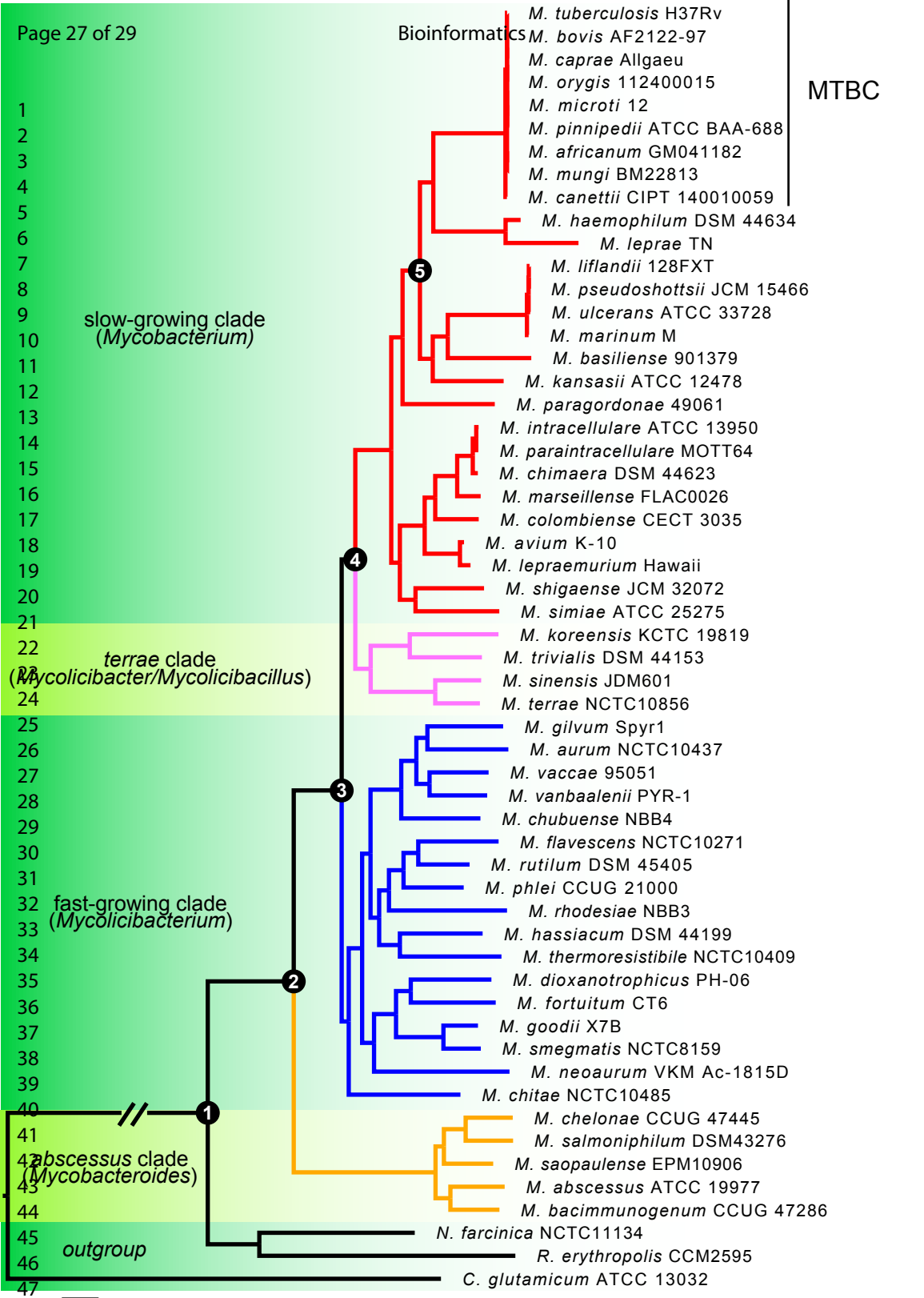

Phenotype
Phylogenetic tree
Variation
Phenotype A
Phenotype B
1
Species 1
Species 2
Species 3
Species 4
Species 5
Species 6
Species 7
Species 8
Species 9
Species 10
Species 11
Species12
Species 13
Species 14
Species 15
Species16
Species17
ProtA ...NAHP...
ProtA ...NAHP...
ProtA ...NAHP...
ProtA ...NAHP...
ProtA ...NAHP...
ProtA ...NAHP...
ProtA ...NAHP...
ProtA ...NAHP...
ProtA ...YAHP...
ProtA ...YAHP...
ProtA ...YAHP...
ProtA ...YAHP...
ProtA ...YAHP...
ProtA ...YAHP...
ProtA ...YAHP...
ProtA ...YAHP...
ProtA ...YAHP...

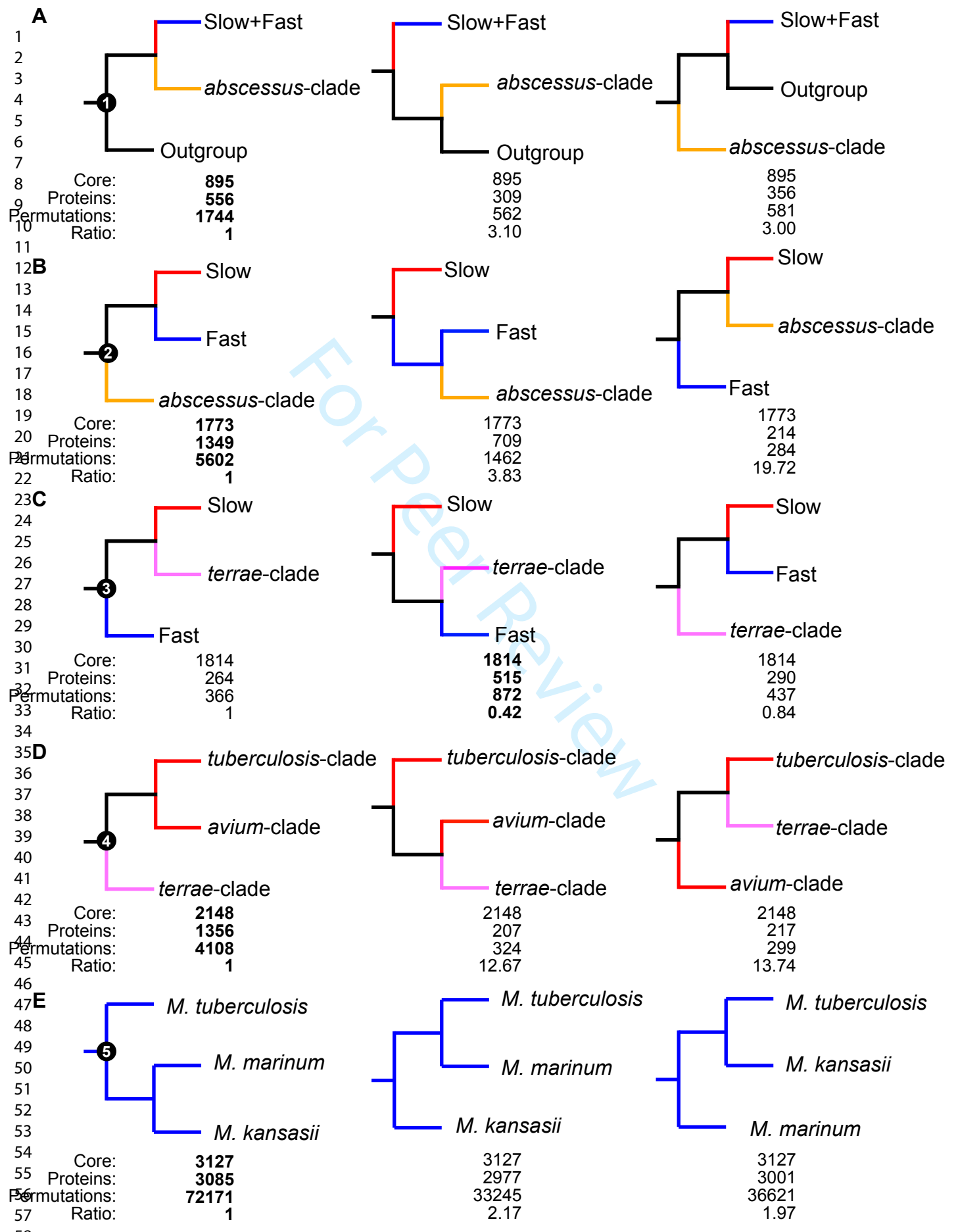
B


Capri-Bv2019
File
Database
Operations
Help
Operations
Combinations
Variations table
Please choose a reference db to begin:
H37Rv
Set database
Reset datab...
Mabscessus
Mafricanum
Maurum
Mavium
Mbacimmunogenum
Mbasiliense
Mbovis
Mcanettii
Mcaprae
Mchelonae
Mchimaera
Mchitae
Mchubuense
Mcolombiense
Mdioxanotrophicus
Mfortuitum
Mgoodii
Mhaemophilum
Mhassiacum
Mintracellulare
Mkoreensis
Mleprae
Mlepraemurium
Mliflandii
Mlitorale
Mmarseillense
To group A
<-
To group B
<-
Mtuberculosis
Mmarinum
Mkansasii
Mgilvum
Mflavescens
Msalmoniphilum
Identity threshold:
40
Query
I vs D
IS vs D
I vs S
Gaps In
Gaps Out
Stop codon

C


GroupA.fasta
Query
GroupB.fasta
Subject
TBLASTn
Parsting and extraction
filtered_data.csv
Identical aa
Similar aa
Different aa
Indels
Stop codon
CDD identification
.tsv
Features table
Matching
Group B
candidates
Group A
Candidate proteins and mutations
MySQL
FeatureTable
idProtein (smallint)
locus_tag (varchar)
product_accession (varchar)
non_redundant_refseq (varchar)
name (varchar)
start (varchar)
end (varchar)
strand (varchar)
length (varchar)
symbol (varchar)
ProteinsTable
idProtein (smallint)
idBlast (smallint)
identity (varchar)
similarity (varchar)
evalue (varchar)
subject_length (varchar)
identical (text)
similar (text)
different (text)
gapsQuery (text)
gapsSubject (text)
stop (varchar)
BlastTable
idBlast (smallint)
query (varchar)
subject(varchar)


M. tuberculosis H37Rv
M. bovis AF2122-97
M. caprae Allgaeu
M. orygis 112400015
M. microti 12
M. pinnipedii ATCC BAA-688
M. africanum GM041182
M. mungi BM22813
M. canettii CIPT 140010059
MTBC
M. haemophilum DSM 44634
M. leprae TN
M. liflandii 128FXT
M. pseudoshottsii JCM 15466
M. ulcerans ATCC 33728
M. marinum M
M. basiliense 901379
M. kansasii ATCC 12478
M. paragordonae 49061
M. intracellulare ATCC 13950
M. paraintracellulare MOTT64
M. chimaera DSM 44623
M. marseillense FLAC0026
M. colombiense CECT 3035
M. avium K-10
M. lepraemurium Hawaii
M. shigaense JCM 32072
M. simiae ATCC 25275
M. koreensis KCTC 19819
M. trivialis DSM 44153
M. sinensis JDM601
M. terrae NCTC10856
M. gilvum Spyr1
M. aurum NCTC10437
M. vaccae 95051
M. vanbaalenii PYR-1
M. chubuense NBB4
M. flavescens NCTC10271
M. rutilum DSM 45405
M. phlei CCUG 21000
M. rhodesiae NBB3
M. hassiacum DSM 44199
M. thermoresistibile NCTC10409
M. dioxanotrophicus PH-06
M. fortuitum CT6
M. goodii X7B
M. smegmatis NCTC8159
M. neoaurum VKM Ac-1815D
M. chitae NCTC10485
M. chelonae CCUG 47445
M. salmoniphilum DSM43276
M. saopaulense EPM10906
M. abscessus ATCC 19977
M. bacimmunogenum CCUG 47286
N. farcinica NCTC11134
R. erythropolis CCM2595
C. glutamicum ATCC 13032
slow-growing clade (Mycobacterium)
terrae clade (Mycolicibacter/Mycolicibacillus)
fast-growing clade (Mycolicibacterium)
abscessus clade (Mycobacteroides)
outgroup
1
2
3
4
5
0.05



A

Locus tag	Gene	Product	Cat.	Mutation	Grantham	Ex.
Rv2727c	*miaA*	tRNA delta(2)-isopentenylpyrophosphate transferase	J	197=C/W	215	139
Rv1205	*log*	LONELY GUY (LOG)	L	46=W/S	177	92
Rv2388c	*hemN*	oxygen-independent coproporphyrinogen III oxidase	H	107=W/S	177	92
Rv2788	*sirR*	transcriptional repressor SirR	K	155=W/S	177	92
Rv3253c	*Rv3253c*	cationic amino acid transport integral membrane protein	E	63=W/S	177	92
Rv1231c	*Rv1231c*	membrane protein	S	11=Y/D	160	87
Rv2221c	*glnE*	[glutamate--ammonia-ligase] adenylyltransferase	OT	553=Y/D	160	87
Rv1842c	*Rv1842c*	hypothetical protein	E	129=Y/N	143	129
Rv2509	*Rv2509*	short-chain type dehydrogenase/reductase	P	153=Y/N	143	129
Rv1485	*hemZ*	Ferrochelatase	H	317=L/G	138	201
Rv2969c	*Rv2969c*	hypothetical protein	O	28=G/L	138	193
Rv3667	*acs*	acetyl-CoAsynthetase	I	316=G/I	135	110
Rv2220	*glnA1*	glutamine synthetase	E	158=I/E	134	197
Rv0466	*Rv0466*	hypothetical protein	I	126=M/G	127	218
Rv0050	*ponA1*	bifunctional penicillin-insensitive transglycosylase	M	416=A/D	126	193
Rv3682	*ponA2*	bifunctional penicillin-insensitive transglycosylase	M	591=A/D	126	193
Rv2799	*Rv2799*	membrane protein	S	156=H/W	115	72
Rv2743c	*Rv2743c*	hypothetical protein	S	77=F/P	114	112
Rv0002	*dnaN*	DNA polymerase III subunit beta	L	286=L/R	102	185
Rv0502	*Rv0502*	hypothetical protein	I	117=R/L	102	242
Rv0702	*rplD*	50S ribosomal protein L4	J	48=R/L	102	242
Rv2164c	*Rv2164c*	hypothetical protein	D	175=R/L	102	242
Rv2220	*glnA1*	glutamine synthetase	E	101=L/R	102	185
Rv2917	*Rv2917*	hypothetical protein	L	169=L/R	102	185
Rv3522	*ltp4*	lipid transfer protein	I	126=R/L	102	242
Rv3681c	*whiB4*	transcriptional regulator WhiB4	K	36=L/R	102	185
Rv0501	*galE2*	UDP-glucose 4-epimerase GalE	GM	349=R/W	101	63
Rv2097c	*pafA*	proteasome accessory factor PafA	O	84=W/R	101	103
Rv2746c	*pgsA3*	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase	I	147=W/R	101	103
Rv3627c	*Rv3627c*	hypothetical protein	M	455=W/R	101	103

B