INTRODUCTION

Dans le cadre du second semestre de Master 2 Biologie structurale, bioinformatique et biotechnologies, parcours biologie structurale intégrative et bioinformatique, mon stage s'est déroulé au sein de l'équipe CSTB (*Complex Systems and Translational Bioinformatics,* http://icube-cstb.unistra.fr/fr/index.php/Accueil) sous la tutelle d'Olivier Poch. Cette équipe étudie les systèmes et réseaux eucaryotiques (complexes macromoléculaires, voies de signalisation, organelles...) par des approches informatiques afin d'identifier tous les partenaires de ces systèmes complexes et d'étudier leurs interactions dynamiques pour modéliser l'impact de perturbations (classiquement des variations responsables de maladies génétiques) sur le comportement et la stabilité des systèmes.

A l'ère du haut débit, la quantité de données biologiques ne cesse d'augmenter, générant des besoins en outils de fouille et analyse susceptibles de manipuler et traiter de gros volumes de données (Greene *et al.*, 2014). Parmi les outils de fouille de données de séquences, on trouve le programme BLAST (*Basic Local Alignement Search Tool*) (Altschul *et al.*, 1990) qui est de loin le plus performant, rapide et utilisé. Schématiquement, BLAST permet d'identifier les régions similaires entre une séquence requête et des séquences cibles (protéiques ou nucléiques) présentes dans les bases de données de séquences protéiques/nucléiques (ex. Protein Data Bank, Swissprot, GenBank…) et cela, en s'appuyant sur des matrices de substitutions et sur des calculs statistiques basés sur des alignements deux à deux des séquences. BLAST est devenu un outil fondamental pour les études scientifiques, car il permet d'évaluer rapidement la proximité d'une séquence requête aux centaines de millions de séquences cibles disponibles à présent.

En terme d'applications, BLAST est utilisé afin d'identifier les gènes/protéines appartenant à une même famille ou de prédire des gènes/protéines. Il permet aussi d'aider à l'annotation de gènes/protéines requêtes de fonction inconnue en inférant les fonctions des gènes/protéines cibles similaires. De même, les études de recherche d'homologie (séquences descendant d'un ancêtre commun), de phylogénétique (reconstitution de l'histoire évolutive) ou de métagénomique (caractérisation des espèces présentes dans des échantillons complexes) s'appuient souvent sur l'exploitation des résultats de BLAST.

Dans le domaine de la recherche d'homologie, il existe de nombreux d'outils qui utilisent l'algorithme de BLAST (Dessimoz *et al.*, 2012), tels qu'OrthoInspector (Linard *et al.*, 2011). OrthoInspector est un logiciel de détection des relations d'orthologie entre espèces différentes exploitant la méthode de *Reciprocal Best Hit* (*RBH*) (Moreno-Hagelsieb *and* Latimer, 2008). Schématiquement, cette méthode se base sur la recherche du meilleur *hit* d'une espèce cible à partir d'une protéine requête. Puis, ce meilleur *hit* est utilisé pour exécuter un deuxième BLAST. Si le meilleur *hit* obtenu correspond à la protéine requête initiale, il est fort probable que les protéines découlent

d'un gène ancêtre commun (orthologues). Dans ce contexte, la génomique comparative, qui s'appuie souvent sur la notion d'orthologues et sur l'approche *RBH*, ambitionne d'exploiter les distributions des gènes/protéines orthologues entre espèces afin i) d'étudier l'évolution des génomes, leur organisation et leur conservation au fil du temps et ii) de mieux comprendre la distribution et l'évolution des fonctions biologiques.

Pour répondre aux divers types d'exploitations, plusieurs variantes du programme BLAST (PSI-BLAST, PHI-BLAST, BLAT, …) (Mount, 2007) ont été développées (voir le site : http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=*Web*&PAGE_TYPE=BlastHomeNew). En parallèle, une multitude d'outils d'analyse, d'exploitation et de visualisation des résultats issus de BLAST ont été développés. Dans ce cadre, on peut citer les outils : « BLAST 2 Sequences » (Tatusova *and* Madden, 1999), un outil permettant de comparer des séquences protéiques/nucléiques homologues d'espèces très proches, « BlastGrabber » (Neumann *et al.*, 2014) un outil téléchargeable offrant la possibilité de visualiser et de trier les résultats issus de BLAST ou « Kablammo » (Wintersinger *and* Wasmuth, 2015), un outil *web* permettant une visualisation interactive des résultats issus de BLAST.

Au regard de la puissance et de la diversité des programmes BLAST et de la multiplication des génomes complets séquencés par l'émergence des biotechnologies à haut débit, il est surprenant de constater que peu d'outils permettent d'exploiter, analyser et visualiser la totalité des fichiers BLAST issus d'un protéome requête. Après recherche bibliographique, outre les approches du type recherche d'orthologues qui mettent en jeu des protéomes complets, seul un outil est capable d'exploiter des résultats de BLAST de protéomes complets, à savoir : *Negative Proteome Database* (Reiter *et al.*, 2007). *Negative Proteome Database* est une ressource en ligne mise en place sous forme de base de données. Ce service peut être questionné pour chercher les protéines spécifiques à une espèce donnée et permet de comparer les protéomes de diverses espèces. La base de données contient les fichiers BLAST résultants de chaque protéine de plusieurs espèces, entre autres *Homo sapiens*, *Mus musculus* et *Arabidopsis thaliana*, et permet, après sélection de l'espèce requête, d'exclure certaines espèces des résultats de BLAST, c'est-à-dire d'obtenir des protéines pour lesquelles les espèces exclues n'ont pas obtenu de *hit*. L'utilisateur peut continuer son analyse sur cette série de protéines en choisissant cette fois les espèces qui ont obligatoirement obtenu un *hit.* Néanmoins, ce service se limite uniquement à quelques protéomes et n'autorise pas la réalisation, ni l'analyse de nouveaux protéomes.

C'est dans ce contexte que s'inscrit mon projet qui vise à créer une structure *web* souple et conviviale (http://lbgi.fr/blastome), nommée *BLASTome Dashboard* [*BLASTome* pour désigner l'ensemble des fichiers résultants de l'analyse par BLAST de toutes les protéines d'un protéome requête]. L'idée centrale est d'autoriser une analyse BLAST de protéomes requêtes *in extenso* depuis l'exécution des BLAST, le stockage des fichiers BLAST (fichiers texte contenant les

résultats de BLAST) jusqu'à leur exploitation dans un cadre permettant à l'utilisateur une libre définition des espèces cibles d'intérêt et la réalisation, à la volée, de filtres mimant une approche de génomique comparative. Ce site *web* a donc pour but, outre l'obtention du *BLASTome*, de pouvoir identifier des distributions des protéines d'espèces cibles similaires aux protéines du protéome requête et de réaliser des analyses préliminaires de génomique comparative simplifiées et rapides.

En plus, sont également implémentés, afin de faciliter l'analyse des séries de protéines obtenues après sélection, des outils permettant d'évaluer l'enrichissement en termes GO (*Gene Ontology*) (Ashburner *et al.*, 2000). *Gene Ontology* est un vocabulaire unifié décrivant les processus biologiques, les mécanismes moléculaires, ainsi que la localisation cellulaire des produits des gènes au sein de toute espèce. Au regard des quantités massives de données manipulées par *BLASTome Dashboard*, une attention toute particulière a été portée à la réalisation d'une visualisation synthétique et ergonomique des résultats.

En accord avec les centres d'intérêt de l'équipe et pour permettre le développement du prototype de *BLASTome Dashboard*, 100 espèces cibles par défaut, regroupées en plusieurs clades, ont également été prédéfinies dans le cadre de l'étude du cil. Il s'agit de clades ayant des distributions d'espèces ciliées et non-ciliées intéressantes (*e.g. Fungi*, *Alveolata*, …) (Figure 1). En plus, seuls les *BLASTomes* de 11 espèces d'intérêt pour l'étude du cil (ou flagelle) ont été effectués . L'équipe CSTB s'intéresse étroitement à cette organelle qui est présente sous forme d'extension à la surface de nombreuses cellules eucaryotes. Chez l'homme, le cil peut atteindre jusqu'à 10 micromètres de long avec un diamètre d'environ 0.2 micromètre et se retrouve à la surface de la quasi-totalité des cellules. Les cils primaires et motiles ont une importance capitale dans le développement, le cycle cellulaire ou la reproduction humaine. Le cil primaire joue le rôle d'antenne au sein de la cellule étant capable de percevoir les signaux environnants (*e.g.* photorécepteurs). Le cil motile quant à lui est impliqué dans la motilité des cellules (*e.g.* les spermatozoïdes) ou le déplacement du milieu environnant (*e.g.* dans le cerveau). Comme en attestent sa distribution très étendue au sein des eucaryotes et son implication dans de nombreuses maladies génétiques (appelées ciliopathies), le cil représente un élément capital au sein du vivant.

Matériel et méthodes

➢ Infrastructure

• Serveurs

L'équipe CSTB dispose de sept serveurs DELL R720 configurés de la même façon sous Linux Ubuntu (ena, studio, esxi, bipbip, ouioui, milex, octave) assurant un environnement centralisé, le calcul et la gestion des données. Chacun des serveurs est équipé de mémoire vive

allant de 48 Go à 512 Go et d'environ 150 To cumulés de disques répartis entre les serveurs. Tous les disques sont directement accessibles à partir de toutes les machines par des montages NFS. Seul le serveur octave est accessible de l'extérieur. Ena, le serveur central utilisé pour le développement de *BLASTome Dashboard*, dispose de 96 To de disques et de 128 Go de mémoire vive.

En plus des sept serveurs, l'équipe a également accès aux trois machines de la plate-forme bioinformatique de Strasbourg BISTRO (*bioinfo-**bistro**.fr*). Cette plate-forme implique plusieurs équipes de sept laboratoires et instituts *(ICube, GMGM, IBMC, IBMP, IGBMC, LGM, IPHC)* provenant de divers domaines des sciences du vivant. L'objectif de la plate-forme est de développer et implémenter des outils et des ressources bioinformatiques destinés au *BioData Mining*.

- Grille de calculs

Exécuter les BLAST du protéome complet d'une espèce nécessite beaucoup de calculs (~20 000 BLAST pour *Homo Sapiens,* ~6000 pour *Saccharomyces cerevisiae*, …). Dans la mesure où la réalisation d'un seul *BLASTome* dure un à deux jours sur les machines de l'équipe en utilisant tous les processeurs, il a été nécessaire d'exécuter les calculs sur les grilles de calculs auxquelles l'équipe à accès par l'intermédiaire de l'*IPHC* et de la plate-forme BISTRO. Ces grilles de calculs, reparties sur des serveurs dispersés dans le monde, sont à disposition et permettent de diminuer drastiquement les temps de calcul en parallélisant les travaux (2-3 heures). Ainsi, les 11 *BLASTomes* effectués durant mon stage ont été exécutés sur ces grilles sous forme de paquets d'environ 200 séquences et en paramétrant les protocoles afin de sécuriser le bon déroulement de ceux-ci. Les résultats sont déposés sur l'espace de stockage et peuvent être récupérés.

- Banque de données de référence

Afin de faciliter l'analyse, j'ai utilisé des bases de données biologiques de références. Notamment, j'ai utilisé UniProt (http://www.uniprot.org/, UniProt Consortium, 2015), qui est une base de connaissances sur les protéines et *Gene Ontology* (http://geneontology.org/), qui concerne les termes GO. Les termes GO sont regroupés en trois sections : processus biologiques, composants cellulaires et fonctions moléculaires. Il s'agit d'un vocabulaire contrôlé et structuré, permettant de comparer les gènes et leurs produits plus facilement. Ces termes sont utilisés lors de l'analyse d'une série de protéines, afin de mettre en évidence un éventuel enrichissement en protéines impliquées par exemple, dans un processus, une voie métabolique ou un compartiment cellulaire.

En plus des termes GO, j'ai utilisé les termes GOSlim (http://geneontology.org/page/go-slim-and-subset-guide) qui, schématiquement, sont des sous-divisions de la totalité des termes GO. Dans ce contexte, les termes GOSlim donnent une vision globale de l'enrichissement en termes GO, s'affranchissant des termes GO trop détaillés. Une version générique des termes GOSlim de certaines espèces modèles est disponible sur le site de *Gene Ontology Consortium*. Néanmoins, ces

sous-divisions peuvent être déterminées par l'utilisateur en fonction de ses besoins, voire en fonction des espèces étudiées.

- Clades cibles et *BLASTomes* requêtes

Dans le cadre de mon stage, 17 groupes/clades regroupant 100 espèces cibles ont été prédéfinis pour faciliter l'analyse des résultats (Figure 1). De même, 11 protéomes requêtes ont été utilisés pour réaliser des *BLASTomes* (*Homo sapiens, Mus musculus, Arabidopsis thaliana, Drosophila melanogaster, Saccharomyces cerevisiae, Batrachochytrium dendrobatidis, Trichoplax adhaerens, Ailuropoda melanoleuca, Caenorhabditis elegans, Chlamydomonas reinhardtii, Tuber melanosporum*).

➢ Outils

Au cours de mon stage, j'ai utilisé Gscope, la plate-forme d'analyse développée au laboratoire pour la création et gestion de projets de bioanalyse (plus de 200 projets à ce jour). Cette plate-forme dispose de nombreuses fonctionnalités et outils de traitements automatiques à haut débit. Notamment, elle permet de créer des projets grâce à un système d'exécution en cascade de programmes, tels que l'exécution des BLAST pour chaque protéine du protéome d'une espèce et de générer et structurer automatiquement les données (séquences protéiques, termes GO associés à chaque protéine du protéome, *etc*.) en les distribuant dans des dossiers spécifiques. Dans le cadre du développement de *BLASTome Dashboard*, le protéome est fourni sous forme d'un fichier FASTA, c'est-à-dire un fichier texte contenant l'ensemble des séquences protéiques dans un format unifié [une ligne avec chevron devant les noms et/ou identifiants de la séquence suivie, à la ligne, de la séquence à proprement parlé (code à une lettre et 80 caractères par ligne)]. De plus, Gscope fournit une interface de recherche et de visualisation intuitive et est également accessible *via* le *web*, ce qui implique la possibilité d'accéder aux projets par un navigateur *web.*

Pour assurer une interopérabilité entre les différentes banques de données utilisées (UniProt, GO), j'ai utilisé le programme « *IDmapping* », mis en place par Luc Moulinier. Il s'agit d'un programme pouvant être appelé en ligne de commande afin de convertir les divers identifiants des gènes/protéines (*gene ID, gene name,* ...) vers/depuis leurs identifiants UniProt. Dans la mesure où j'ai dû jongler entre les différents identifiants, ce programme m'a été d'une grande aide.

Après obtention des *BLASTomes* (*via* l'exécution du programme BLAST sur la grille de calcul de l'*IPHC* pour la totalité des protéines d'une espèce requête), Gscope extrait de chaque fichier BLAST, le meilleur *hit* de chaque espèce cible, c'est-à-dire, théoriquement, la protéine de chaque espèce cible la plus similaire à une protéine requête donnée. Seuls sont considérés les meilleurs *hits* vérifiant un Expect inférieur à 10^{-3}. Cette information est ensuite stockée dans un fichier texte, nommé taxobla (Figure 2) que mon programme va exploiter pour identifier les séries

de protéines d'espèces cibles similaires au protéome d'une espèce requête et vérifiant une distribution phylogénétique (présence/absence) spécifiée par l'utilisateur.

➢ Site *web*

• Langages de programmation

Afin de mener à bien mon projet, j'ai combiné plusieurs types de langages de programmation et utilisé plusieurs librairies de ces langages.

Il m'a d'abord fallu mettre en place le programme clé de ce projet, « dashboard.py ». Ce programme est écrit en Python (version 2.7), langage de programmation orienté objet et multiplate-forme. Python présente divers avantages, notamment il est facile d'utilisation et dispose de nombreuses librairies destinées à la recherche. En plus de fournir des librairies, notamment *goatools* (voir plus bas), Python permet de créer et d'interroger des bases de données par SQLite et de créer des pages HTML.

• Bases de données : SQLite

Pour faciliter la manipulation et le stockage des données, j'utilise la librairie de Python SQLite3. SQLite est un moteur de bases de données relationnelles portable qui intègre la majorité des fonctionnalités du langage SQL (*Structured Query Language*). Les bases de données créées sont sous forme de simple fichier. Ce système me donne un accès rapide et facile aux données, telles que : les identifiants des protéines, les espèces cibles, *etc*. Un modèle de bases de données a été réalisé à l'aide de JMerise, un outil de modélisation permettant de générer les modèles conceptuels de données pour Merise.

• Librairie Python : *goatools*

Afin de calculer les enrichissements en termes GO, parmi les librairies proposées par Python, j'ai choisi la librairie *goatools*. Cette librairie, basée sur le test de Fischer, permet de calculer un enrichissement en prenant comme données d'entrées la liste de tous les gènes d'une espèce (G), les identifiants GO associés à chacun des gènes (totOrga_GO), la totalité des termes GO et identifiants associés présents et téléchargeables sur le site de « *Gene Ontology Consortium* » (http://geneontology.org/page/download-ontology) (tot_GO) et la liste des gènes sur lesquels le calcul doit être effectué (g). En ce qui concerne G et G_GO, *goatools* est en mesure de les récupérer directement sur le site du NCBI (*National Center for Biotechnology*). Dans la mesure où *goatools* se base directement sur les données fournies par le NCBI, il est nécessaire de lui fournir les *gene ID* (identifiants utilisés par le NCBI) des gènes. Dans ce cadre, les identifiants UniProt ont été convertis vers les *gene ID* par *IDmapping* en les extrayant de « *Entrez Gene* », la base de données

relative aux gènes du NCBI (Maglott *et al.*, 2011). Enfin, *goatools* permet également d'obtenir les GOSlim.

Comme le site du NCBI ne propose que les données relatives à quelques espèces modèles et pour permettre à *BLASTome Dashboard* d'analyser des génomes/protéomes récemment séquencés ou peu annotés, « dashboard.py » fourni à *goatools* les données nécessaires (G, totOrga_GO, tot_GO et g) au bon format. totOrga_GO est obtenu *via* Gscope qui récupère les termes GO associés à chaque gène dans la base de données du site *Gene Ontology* ainsi que G, la liste de toutes les protéines des espèces requêtes. Grâce à ces données, *goatools* peut calculer l'enrichissement des espèces requêtes non disponibles au NCBI.

- Visualisation et interaction : HTML, CSS et JavaScript

La visualisation et l'accès aux données et résultats se fait *via* un site *web* dédié, (lbgi.fr/blastome), d'où l'utilisation du langage HTML (*Hypertext Markup Language*) qui est un langage de balisage rendant la visualisation de données à travers une page *web* possible.

Combiné à HTML, j'ai également utilisé CSS (*Cascading Style Sheets)*, un langage informatique qui facilite la mise en forme du contenu d'une page *web* afin de rendre la visualisation plus agréable et intuitive. Toutes les informations relatives à la mise en forme des pages *web* (couleurs, formes, police, surimpressions …) peuvent être regroupées dans un seul fichier CSS appelé par le code HTML afin d'obtenir une représentation homogène des différents éléments d'un site *web*. De plus, il suffit de modifier le code CSS pour changer l'apparence de toutes les pages.

- JavaScript, AJAX et JQuery

Afin de rendre *BLASTome Dashboard* dynamique et interactif, j'ai également utilisé JavaScript (Figure 3), qui est un langage de programmation orienté objet particulièrement utilisé dans la création de pages *web.* Ce langage est exécuté côté client et non serveur, mais il peut communiquer avec le serveur afin de récupérer des informations. Il permet, par exemple, de vérifier les données rentrées par l'utilisateur, mais également de rendre la page *web* dynamique dans un but esthétique et/ou ergonomique.

L'emploi du protocole AJAX (Figure 3) permet à JavaScript d'interroger le serveur sans recharger la page *web*, autorisant une interaction plus fine entre le serveur et le client.

JQuery (Figure 3) est une librairie de JavaScript libre et multiplate-forme pour faciliter : le traitement d'événements (cliquer, sélectionner, …), les échanges client/serveur (AJAX), la mise en place d'effets visuels et d'animations et la manipulation des feuilles de style CSS. Cette libraire se présente sous forme de fichier téléchargeable sur le serveur ou directement accessible par internet (https://ajax.googleapis.com/ajax/libs/jquery/1.12.2/jquery.min.js).

• Navigateurs

Pour vérifier la viabilité de mon code, j'ai travaillé sur le navigateur *Mozilla Firefox,* version 46.0. Il est à noter que le code CSS est reconnu par tous les navigateurs, mais que certains éléments de CSS ne sont pas compatibles avec certains navigateurs, voire mal implémentés. Pour éviter ce problème, des tests ont également été effectués sur les navigateurs *Google Chrome* (version 50) et *Opera* (version 37).

RÉSULTATS

Schématiquement*, BLASTome Dashboard* est un outil *web* qui a pour objectifs :

i) de réaliser les BLAST d'un protéome requête complet (ensemble des protéines d'une espèce) et de stocker les fichiers BLAST en résultant (*BLASTome*),

i) de réaliser et rendre accessible une analyse de génomique comparative simplifiée et rapide des *BLASTomes* par le dénombrement et le filtrage, à la volée, de séries de protéines requête vérifiant la présence, ou l'absence, de protéines similaires chez des clades/espèces cibles (profil de distribution des protéines similaires),

i) de permettre l'évaluation fonctionnelle des séries de protéines requêtes vérifiant un profil spécifié de distribution des protéines similaires *via* le calcul des enrichissements en termes GO.

Pour atteindre ces objectifs, un cahier des charges des fonctionnalités requises a été défini afin d'aboutir à une structure *web* flexible et adaptable aux besoins des utilisateurs. Très tôt, notre stratégie de développement s'est orientée vers la création d'une structure de type client léger, c'est-à-dire une structure où l'utilisateur n'a besoin que d'un navigateur *web* pour accéder aux fonctionnalités. De même, des fonctionnalités essentielles ont été arrêtées qui recouvrent :

- la réalisation et la visualisation à la volée du *BLASTome* d'un protéome requête,
- la libre définition par l'utilisateur des espèces et clades cibles,
- le choix, la définition et la visualisation par l'utilisateur de « processus d'intérêt »,
- le calcul rapide et l'affichage synthétique des résultats,
- l'aide à l'analyse des séries de protéines vérifiant un profil de distribution des protéines similaires.

Dans ce contexte, trois notions ont été définies qui sont regroupées dans un cartouche spécifique de la page d'accueil du site *web* (Figure 4A) :

➢ Protéomes requêtes

Un protéome requête regroupe l'ensemble des protéines d'une espèce qui seront utilisées pour exécuter les BLAST. *BLASTome Dashboard* offre la possibilité d'exécuter le *BLASTome* de

n'importe quelle espèce, à condition de disposer du protéome au format FASTA. Ce type de fichier est téléchargeable, par exemple, sur le site UniProt, dans la section protéomes (http://www.uniprot.org/proteomes/). Un bouton (« New *BLASTome* ») redirige l'utilisateur vers une nouvelle page (lbgi.fr:1664/wali/wali.rvt?do=Gstore), à partir de laquelle il peut soumettre le fichier FASTA de son choix, ou bien sélectionner une des espèces proposées dans une liste déroulante. Ensuite, il suffit de suivre les instructions pour créer le projet correspondant et exécuter le *BLASTome* sur les grilles de calculs en automatique. Un code couleur permet de connaître l'état des projets (blanc : non-disponible, vert : disponible, orange : en cours/incomplet).

Dans le cadre du développement du site et de notre étude des protéines ciliaires comme processus test (voir plus bas), plusieurs *BLASTomes* ont été pré-calculés et sont dès à présent disponibles. Ces protéomes ont été choisis pour couvrir des tailles diverses (~6.700 à ~47.000 protéines) et sont distribués dans différents clades ou groupes taxonomiques.

➢ Espèces et clades cibles

Les espèces et clades cibles sont les seuls à être considérés pour les analyses et visualisations fournies par *BLASTome Dashboard*. Ceux utilisés par défaut pour le développement du site et détaillés dans la section Matériel et Méthodes sont tous des eucaryotes dont les génomes et protéomes sont complets et de qualité (peu de contigs, bonne couverture, bonne définition des gènes…). Les listes des espèces et de clades peuvent être modifiées par l'utilisateur en fournissant un fichier texte dont un exemple est disponible sur le site.

➢ Processus d'intérêt

Le concept de 'processus d'intérêt' a été introduit pour désigner des mécanismes biologiques liés à un ensemble de gènes, un complexe macromoléculaire, un compartiment… qui possèdent une distribution connue au sein des clades/espèces cibles et qui sont potentiellement impliqués dans un même processus. Pour caractériser les processus d'intérêt, un système de visualisation par code couleur a été mis en place. Ainsi, la représentation des espèces qui possèdent le processus d'intérêt sera de couleur verte, celle des espèces qui ne possèdent pas le processus sera de couleur rouge et celle dont on ne sait pas si l'espèce possède ou non le processus sera bleu. Ce système prend en entrée un fichier texte fourni par l'utilisateur qui contient les informations sur la présence/absence du processus au sein des espèces cibles (un fichier type au format idoine est disponible et téléchargeable sur le site). Par défaut, le code couleur présenté sur le site correspond aux espèces pourvues ou dépourvues d'une structure ciliaire, sujet d'étude privilégié de l'équipe. Cependant, afin de vérifier le bon fonctionnement du système de code couleur mis en place, un second test de code couleur est accessible, basé sur des résultats préliminaires décrivant la présence/absence de la topoisomérase au sein des eucaryotes (Claudine Mayer, communication personnelle).

➢ Pré-calculs : Identification des meilleurs *hits*

Afin de pouvoir effectuer sur le site *web*, la réalisation et la visualisation, à la volée, d'analyses de génomique comparative simplifiées, plusieurs données exploitant les *BLASTomes* et les espèces cibles sont pré-calculées.

Le programme principal, « dashboard.py », écrit en python, lit tous les fichiers taxobla (contenant la liste des meilleurs *hits* de chaque espèce cible, Figure 2) relatifs à un *BLASTome* requête et stocke dans une base de données (Figure 5), les informations concernant chaque BLAST (identifiant de la protéine requête, espèces cibles trouvées dans le BLAST, …). A noter que seules les espèces et clades cibles prédéfinis par l'utilisateur sont considérés et qu'afin de minimiser les faux positifs, seules les protéines similaires ayant un Expect inférieur à 10^{-3} ont été retenues.

Après avoir stocké ces informations dans la base de données SQLite (une base de donnée par *BLASTome* et par ensemble de clades), « dashboard.py » compte le nombre total de fichiers BLAST du protéome requête vérifiant la présence d'un meilleur *hit* pour chaque espèce cible (NbTotal). Grâce à la base de données SQLite, l'accès à ces informations est rapide.

➢ Visuel et fonctionnalités de la page d'accueil

Le site *web* est structuré en deux parties (Figure 4). Comme décrit précédemment, la partie supérieure (Figure 4A) comprend le cartouche regroupant les notions et fonctionnalités d'espèce requête, de clades/espèces cibles et de processus étudié. La partie inférieure (Figure 4B) est essentiellement constituée de cadres individualisés représentant les clades avec en leur sein, les espèces (représentées par des disques surmontés de l'acronyme à 4 lettres du nom de l'espèce) (Tableau 1). Pour chaque espèce cible, le NbTotal est fourni en passant la souris au-dessus du disque correspondant et approximé par une intensité différente de la couleur remplissant chaque disque (petites valeurs : moins foncé, valeurs hautes : plus foncé). La partie supérieure de chaque cadre, portant le nom du clade et le nombre d'espèces, ainsi que les disques représentant les espèces sont associés à un fond de couleur lié à la distribution du processus étudié, en l'occurrence dans la version par défaut, vert : espèces/clades possédant le cil ; rouge : espèces/clades ne possédant pas de cil, orange : clades regroupant un mélange d'espèces avec et sans cil et bleu : espèces indéfinies.

Pour gérer les sélections de protéines requêtes vérifiant un profil *ad hoc* de présence/absence de protéines similaires au sein des espèces cibles, nous avons distingué un cadre surplombant les cadres clade et présentant deux boutons : le premier pour remettre à zéro les sélections (« *Clear selection* »), le second pour soumettre les sélections (« *Send* »). Les sélections d'espèces peuvent s'effectuer de 2 façons non exclusives : 1) en cliquant sur les disques correspondants et sélectionnant ainsi des espèces précises gérées par mes procédures JavaScript et JQuery et 2) en fournissant dans l'espace texte situé après la phase '*Nb of species*', un nombre d'espèces précis à

retenir pour la sélection ou un nombre situé dans un intervalle (sous le format : M-N). A titre d'exemple, en cochant un unique disque (par exemple, le *Fungus Batrachochytrium dendrobatidis, Bden)*, l'utilisateur obtient NbTotal (7962), c'est-à-dire la liste complète des protéines de l'espèce requête (l'homme) présentant au moins une protéine similaire chez l'espèce cible (*Bden*). Par contre, en fournissant un nombre ou un intervalle dans le cadre approprié, l'utilisateur obtient la liste complète des protéines requête dont les fichiers BLAST présentent une protéine similaire de n'importe quelle espèce du clade pour peu que le fichier vérifie strictement le nombre (ou l'intervalle) spécifié. Ainsi, en partant du *BLASTome* humain et en fournissant un 1 dans l'espace réservé dans le cadre *Fungi*, on obtient tous les fichiers BLAST humains (1114) qui vérifient la présence d'une unique espèce de *Fungi*, quel que soit l'espèce de *Fungi*.

Ce système combinant choix au sein des clades et possibilité de cocher les espèces désirées permet de sélectionner des protéines requêtes vérifiant une distribution de protéines similaires très précise. Ceci associé à l'étude d'un processus donné, octroie la possibilité de filtrer le protéome requête en fonction dudit processus.

A titre d'exemple, une sélection a été réalisée en inscrivant 0 dans les zones texte des clades *Amoebozoa* et *Capsaspora* et sélectionnant les espèces *Clamydomonas reiinhardti* (*Crei,* une espèce unicellulaire utilisée en tant qu'espèce modèle pour l'étude du cil) et *Volvox carteri* (*Vcar*). Ces deux espèces sont des algues vertes ciliées appartenant au clade des *Archaeplastida*. Dès que les sélections sont finies, il suffit de cliquer sur le bouton « *Send* » pour lancer l'analyse. La page est actualisée avec les nouvelles données (Figure 6). Le cartouche, n'ayant plus son utilité, disparaît au profit des informations relatives à l'analyse (nombre de protéines requêtes filtrées, en l'occurrence 455, et les filtres utilisés) (Figure 6A). Ensuite, les fichiers BLAST des protéines requêtes filtrées sont de nouveau pris en charge par le programme python pour compter combien de fois chaque espèce cible individuellement obtient un *hit* (NbFiltre, accessible en passant la souris au-dessus du disque correspondant). Ceci permet d'obtenir la répartition par espèce, des protéines similaires aux protéines filtrées de l'espèce requête et vérifiant les conditions désirées tout en gardant la même représentation (un cadre par clade, espèces représentées par des disques, intensité liée aux valeurs NbFiltre) (Figure 6B). En dessous des cadres clades vient alors l'ensemble des boutons portant l'identifiant UniProt d'une protéine requête filtrée (un par protéine) (Figure 6C), puis, les enrichissements en termes GO dans un cadre individualisé (Figure 6D).

- Analyse

Dans le but de faciliter l'analyse et l'exploitation des résultats, « dashboard.py » donne un accès direct aux fichiers BLAST des protéines filtrées, en cliquant sur le bouton d'une protéine requête (cité précédemment) (Figure 6C). Ces boutons permettent d'ouvrir une nouvelle page contenant le fichier BLAST de la protéine sélectionnée, fichier incluant le code couleur processus.

En d'autres termes, toutes les lignes du fichier BLAST concernant les protéines des espèces cibles auront un arrière-plan coloré en fonction du processus (jaune : espèce requête, vert : espèce possédant le processus, rouge : espèce ne possédant pas le processus, bleu : espèce dont l'état du processus est inconnu).

Le lien direct vers le site UniProt de la protéine requête est également disponible pour avoir des informations supplémentaires.

Toujours dans l'optique de faciliter l'analyse des protéines filtrées, un système d'enrichissement en termes GO, calculé pour n'importe quel protéome requête, a été mis en place grâce à la librairie *goatools* proposée par python. Les informations relatives à son génome et aux termes GO sont disponibles et utilisables par *goatools* sur le site du NCBI (décrit dans la section « Matériel et méthodes »). Après calcul de l'enrichissement en termes GO, g*oatools* écrit les résultats dans un fichier texte que l'utilisateur peut télécharger (Figure 6D, « GO Text format »). L'enrichissement en termes GO est calculé avec des données extraites de Gscope (totalité des gènes, termes GO associés à ces gènes). De plus, les résultats sont affichés dans la page de résultats du site dans un cadre en bas de page (Figure 6D) fournissant dans un format spécifique les informations essentielles sur les identifiants (GO, *Gene ID*), les processus et termes GO, les p-values des enrichissements… (Figure 7).

Lorsque l'espèce requête n'est pas une espèce modèle, voire lorsqu'il s'agit d'une espèce récemment séquencée, son génome est peu/pas annoté, donc peu/pas de termes GO sont associés à ses gènes. Par conséquent, l'enrichissement en termes GO, s'il y en a un, est souvent très limité. C'est la raison pour laquelle j'ai mis en place un système d'enrichissement indirect basé sur les termes GOSlim, termes moins détaillés et plus génériques que les termes GO (voir Section Matériel et Méthodes). Comme le NCBI ne met à disposition de *goatools* que des données relatives à quelques espèces modèles, j'ai choisi 6 espèces cibles modèles bien documentées appartenant à des groupes taxonomiques différents : *Homo sapiens, Mus musculus, Drosophila melanogaster, Saccharomyces cerevisiae, Caenorhabditis elegans* et *Arabidopsis thaliana*. « Dashboard.py » lit les fichiers taxobla relatifs aux fichiers BLAST des protéines requêtes filtrées, puis il transmet à *goatools,* les *gene ID* des meilleurs *hits* des 6 espèces modèles. Ensuite, *goatools* calcule, pour chaque espèce modèle individuellement, l'enrichissement en termes GOSlim sur la base des séries de meilleurs *hits*. Il est à noter que pour une même liste de protéines requêtes, les séries et enrichissements des termes GOSlim peuvent varier d'une espèce modèle à une autre. Les résultats GOSlim sont ensuite affichés dans six cadres individualisés sous les résultats d'enrichissement en termes GO et peuvent être téléchargés au même format texte (non visible sur la figure 6, car *Homo sapiens* est une espèce bien documentée). Les *gene ID* des protéines de l'espèce requête pour lesquelles un *hit* de l'espèce modèle cible a été retenu pour le calcul des enrichissements sont également ajoutés en fin de chaque ligne. Au final, lorsque l'espèce requête n'est pas une espèce

modèle, l'utilisateur peut exploiter 7 enrichissements, un en termes GO du protéome requête et six en termes GOSlim des espèces modèles.

En plus d'avoir la liste de chaque gène/protéine associé à un terme GO/GOSlim au format texte, chaque ligne peut être sélectionnée. Le fait de cliquer sur une ligne permet de colorer en jaune les boutons des protéines requêtes filtrées associées (Figure 6C et 6D). Ainsi, il est possible de visualiser dynamiquement les protéines connues de l'espèce requête associées aux termes GO enrichis, mais également des protéines sans terme GO pour lesquelles une protéine similaire d'une espèce modèle a été annotée.

➢ Vérification des choix/sélections

Afin de contrôler les choix effectués par l'utilisateur et éviter d'éventuelles erreurs de sélection, une procédure JavaScript a été instaurée. Cette procédure vérifie le nombre ou l'intervalle donné par l'utilisateur afin d'éviter, par exemple que celui-ci ne soumette un nombre plus grand que le nombre total d'espèces d'un clade. Les bornes des intervalles sont contrôlées de la même façon.

➢ Exemples d'application

- *Homo sapiens* et *Fungi*

Afin de tester la robustesse de cet outil *web* et dans le cadre de l'étude du cil, j'ai d'abord visualisé le *BLASTome* de l'homme (espèce requête par défaut) et j'ai exploré les distributions de protéines similaires au sein des protéomes des espèces de *Fungi* (clade mixte pour le processus d'intérêt : présence de cil).

Dans un premier abord, il appert que les NbTotal (nombre de fois où une protéine d'une espèce cible possède un meilleur *hit* dans un fichier BLAST de l'espèce requête) des divers *Fungi* sont variables (de ~2000 à ~8000). Cependant, lorsque les NbTotal sont comparés à la taille des protéomes correspondants (Figure 8), on observe que les deux valeurs ne semblent pas liées puisque la taille du protéome est parfois supérieure à NbTotal (triangle bleu sur figure 8) et parfois, on observe l'inverse (cercle vert sur la figure 8). Ces deux situations reflètent deux phénomènes distincts mais pas forcément exclusifs à savoir : qu'une taille de protéome supérieure à NbTotal résulte souvent de la présence dans ce protéome de protéines ne possédant pas d'homologues chez l'homme. A l'inverse, si la taille du protéome est inférieure à NbTotal, cela implique qu'une même protéine du protéome cible est similaire à plusieurs protéines humaines et se retrouve comme meilleur *hit* dans plusieurs fichiers BLAST humains. Cet exemple illustre bien les précautions et la gymnastique intellectuelle qu'il convient d'avoir en mémoire lorsque l'on analyse des résultats de *BLASTome Dashboard* qui ne se base pas sur la notion d'orthologie (plus robuste et plus précise) mais plutôt sur la présence de régions similaires entre protéines identifiées par BLAST.

Pour tester les diverses fonctionnalités proposées par *BLASTome Dashboard*, j'ai appliqué un filtre sur le clade des *Fungi*, toujours en utilisant le protéome humain comme espèce requête. Dans l'optique d'obtenir les protéines pour lesquelles les résultats de BLAST ne trouvent qu'un seul *Fungus* et aucune amibe (Amoebozoa : clade non-cilié), j'ai écrit 1 et 0 dans les zones texte respectives. J'ai ensuite systématiquement combiné cette sélection réalisée au niveau des clades avec la sélection d'une espèce particulière de *Fungi*. En d'autres termes, j'ai obtenu 16 séries de NbFiltre (une pour chaque espèce de *Fungi* correspondant aux protéines humaines ayant une similarité exclusive avec une seule espèce de *Fungi*, aucune similarité avec des protéines d'amibes et des similarités composites avec les protéines des autres espèces.

Batrachochytrium dendrobatidis (*Bden)* se trouve en tête de liste avec un NbFiltre de 297 fichiers BLAST (Figure 9). Si l'on analyse le visuel *BLASTome Dashboard* correspondant à cette sélection (Figure 10A), on constate que les disques des espèces ciliées présentes dans des clades mixtes (*Archaeplastida*, *Stramenopiles* et *Alveolata*) sont tous d'un vert foncé (au regard des intensités observées pour les espèces non-ciliées). Ceci semble indiquer un enrichissement en protéines spécifiques des espèces ciliées. Cette hypothèse est confirmée lorsque l'on analyse les enrichissements en termes GO, qui révèlent que les processus biologiques liés au cil (*microtubule cytoskeleton organization* : 35 protéines, *cell projection assembly* : 41 protéines, *organelle assembly* : 43 protéines, *intraciliary transport* : 15 protéines, ...) sont préférentiellement enrichis (Figure 10B). Ces résultats confirment que *Bden* est une espèce de choix pour l'étude du cil, car en tant que chytride (*Fungi* saprophites ou parasites) infectant les grenouilles (Refsnider *et al.*, 2015), il fait partie des rares *Fungi* ciliés. Le cil est un processus dans lequel beaucoup de protéines sont impliquées, ce qui explique pourquoi *Bden* se retrouve plus 'proche' de l'homme (NbFiltre le plus élevé) que ne le sont les autres *Fungi*.

Enfin, en analysant la Figure 9, on peut noter que, de façon inattendue, un deuxième *Fungi*, *Emericella nidulans* (*Enid*, *Fungi* non-cilié) se démarque du lot, avec un Nb_filtre de 151 protéines alors que la majorité des *Fungi* présente des valeurs de NbFiltre allant de 1 à 24. Les protéines communes à l'homme et à cette espèce sont majoritairement des protéases. De plus, les disques foncés (espèces cibles à NbFiltre élevé) se distribuent de façon aléatoire entre espèces ciliées (disques verts) et non-ciliées (disques rouges) (données non présentées). Ceci semble confirmer que ces familles de protéases ne sont sans doute pas liées aux fonctions ciliaires.

- *Batrachochytrium dendrobatidis* et protéines *Crinkler*

Dans le cadre de l'étude de la distribution des protéines ciliaires, de nombreux filtres ont été réalisés à partir du protéome requête de *Bden*, seul *Fungus* cilié. Ces tests ont révélé une situation originale avec une proximité surprenante entre *Bden* et *Phytophtora infestans* (*Pinf*), un *Stramenopiles*.

En effet, en choisissant le protéome de *Bden* comme protéome requête, nous avons réalisé un filtre uniquement à partir des clades et vérifiant *0 Amoebozoa, 0 Capsaspora* (clades non-ciliés), 1 *Fungi* et 1 *Stramenopiles*. Ce filtre aboutit à une série de 154 protéines de *Bden* pour lesquelles on observe dans les fichiers BLAST des *hits* uniquement chez un seul *Stramenopiles* et aucun *hit* chez des espèces d'*Amoebozoa* ou de *Capsaspora* (Figure 11). La distribution des NbFiltre au sein des *Stramenopiles* révèle que *Pinf* se démarque avec un NbFiltre de 100 protéines (disque vert foncé, encadré sur la Figure 11) alors que les NbFiltre des autres *Stramenopiles* varient entre 0 et 30 (couleur moins intense) (données non présentées). Ceci laisse supposer qu'il existe un processus commun entre *Pinf* et *Bden* impliquant une centaine de protéines. Afin de vérifier ce qui distingue *Pinf* des autres *Stramenopiles,* le filtre a été modifié en ajoutant une condition supplémentaire, à savoir la sélection de l'espèce *Pinf. BLASTome Dashboard* retourne une série de 100 protéines pour lesquelles *goatools* n'a pu calculer d'enrichissement (données non présentées). En regardant manuellement les résultats de chaque BLAST, il s'est avéré que 51 *hits* de *Pinf* sur 100 concernent des protéines Crinkler (CRN) ou assimilées (CRN-*like*), protéines impliquées dans les phénomènes d'infection. La présence de protéines CRN et CRN-*like* au sein de ces deux espèces a déjà été observée et commentée (Sun *et al.*, 2011 et Frades *and* Andreasson, 2016) et l'hypothèse retenue est que *Pinf* (Raffaele *et al.*, 2010) et *Bden* sont des espèces qui infectent d'autres espèces (les plantes et les grenouilles respectivement) et que les protéines CRN/CRN-*like* seraient liées aux processus d'infection et/ou de détournement de la machinerie de l'hôte. Enfin, on peut noter que dans la mesure où ces protéines CRN/CRN-*like* ne sont pas annotées au niveau du protéome de *Bden*, il est normal que le calcul d'enrichissement en termes GO ne trouve pas de résultats. De même, les protéines CRN/CRN-*like* étant spécifiques aux espèces infectieuses *Bden* et *Pinf*, un enrichissement en termes GOSlim ne peut pas être observé dans les espèces modèles choisies qui ne sont pas des espèces infectieuses et ne possèdent donc pas de protéines similaires.

En plus des 51 protéines CRN/CRN-*like*, quelques protéines (8) sont non-caractérisées et 16 correspondent à des protéines ciliaires. Enfin, les 25 dernières protéines de la série sont des protéines de *Bden* d'environ 450 acides aminés qui sont toutes similaires à une même protéine de *Pinf* (Identifiant Uniprot : D0N2M4_PHYIT) sans fonction définie et possédant un chromodomaine, un domaine structural impliqué dans la reconnaissance des acides nucléiques. Cependant, l'analyse des BLAST a révélé que la région similaire ne correspondait pas au chromodomaine. Le nombre important de protéines nous a incité à réaliser une étude détaillée des fichiers BLAST qui a révélé la présence récurrente et spécifique de protéines similaires sans fonction connue chez certaines espèces : *Nematostella vectensis* (*Cnidaria*), *Oikopleura dioica* et *Strongylocentrotus purpuratus* (*Deuterostomia*), *Tribolium castaneum* et *Acyrthosiphon pisum* (*Insecta*), *Caenhorabtidis elegans* et *Caenhorabtidis briggsae* (*Nematoda*). A notre connaissance, une telle distribution est tout à fait inattendue et totalement atypique, étant donné que ces espèces

sont phylogénétiquement éloignées. Cette famille de protéines originales demeure donc un mystère en ce qui concerne les causes d'une telle distribution (transposon ?) et les fonctions dans lesquelles cette famille pourrait être impliquée.

Discussion/Perspectives

Ce projet visait à fournir un outil facile d'utilisation et souple. C'est pourquoi tant de libertés ont été laissées à l'utilisateur en ce qui concerne le choix/la réalisation du *BLASTome* requête, le choix/la composition des clades cibles et le choix du processus. En effet, *BLASTome Dashboard* vise à aider les chercheurs en proposant d'ores et déjà, une méthode simple et rapide d'analyse de l'ensemble des résultats de BLAST découlant du protéome complet d'une espèce requête. Grâce aux diverses fonctionnalités (choix/dépôt du protéome, libre définition des clades, choix du processus d'intérêt), l'utilisateur garde une liberté d'action qu'il peut utiliser à bon escient pour son sujet d'étude et, grâce à l'exploitation d'un client léger, un simple navigateur *web* est nécessaire, ce qui évite les problèmes liés au téléchargement et à l'installation de logiciels. Enfin, un effort tout particulier a été fait pour obtenir une visualisation synthétique et ergonomique des résultats.

➢ Taille des protéomes

Il s'avère que la taille des protéomes (cibles et requêtes) influence grandement les résultats obtenus. En effet, plus le protéome des espèces cibles est grand, plus il y a de chance que son NbTotal soit grand. Cependant, les duplications (paralogues) et variants du protéome requête risquent de trouver toujours le même *hit* dans une espèce cible, ce qui augmente artificiellement le NbTotal. Dès lors, il est important de garder à l'esprit que le NbTotal ne reflète pas toujours le nombre de protéines similaires distinctes présentes chez les espèces requête et cible. Dans une future version de *BLASTome Dashboard*, un suivi des différences et des proportions entre protéome complet et NbTotal pourrait s'avérer intéressant pour mieux caractériser nos résultats. Ainsi, il serait envisageable d'identifier et caractériser non seulement des protéines requêtes, mais aussi les protéines d'espèces cibles qui ne sont pas similaires aux protéines requêtes, notamment lorsque la taille du protéome de l'espèce cible est largement supérieure au NbTotal [*e.g. Laccaria bicolor* (*Lbic)* et *Puccinia graminis* (*Pgra*) de la Figure 8]. Ces développements permettraient, en une seule sélection, de fournir des informations sur les nombreuses protéines de *Lbic* et *Pgra* qu'on ne retrouve pas chez l'homme et qui pourraient intéresser l'utilisateur.

➢ Choix des espèces/clades cibles

Le système de changement des espèces cibles a été mis en place pour permettre à l'utilisateur de choisir les espèces importantes pour son étude. En effet, en fonction du sujet d'étude, certaines espèces peuvent s'avérer plus intéressantes que d'autres. Même si, en l'état, *BLASTome Dashboard* ne propose que des espèces eucaryotiques par défaut, ce système est applicable à toutes les espèces du vivant, y compris les bactéries et les archées. Pour ces espèces procaryotes aux définitions taxonomiques moins précises, la liberté de choix des clades pourra s'avérer très utile. En effet, pour faciliter la visualisation et surtout, pour mettre en avant l'importance de l'histoire évolutive, les espèces peuvent être réparties en groupes d'espèces (clades) prédéfinis par l'utilisateur. Cela permet de retrouver plus facilement les clades/espèces intéressants et d'avoir une vision globale de ceux-ci.

Cependant, il ne faut pas oublier que seules les espèces à l'intérieur des clades cibles prédéfinies seront considérées lors de l'analyse. En effet, même en choisissant d'exclure un clade donné (*e.g* 0 *Amoebozoa*, 0 *Capsaspora*, ...), des espèces non inclues dans la définition des clades ne sont pas traitées et peuvent donc avoir des *hits* dans les fichiers BLAST sans que cela soit pris en compte. Un choix bien réfléchi des clades, et des espèces à l'intérieur des clades, joue donc un rôle prépondérant pour chaque étude.

Dans ce contexte, nous envisageons de mettre à la disposition de l'utilisateur un arbre de la vie avec les différents clades/espèces. Effectivement, pour l'instant, l'utilisateur doit fournir les clades/espèces dans un fichier au format texte. Afin de rendre le choix des clades/espèces plus aisé, un arbre de la vie sur lequel l'utilisateur pourrait cliquer sur des nœuds (clades ou groupes taxonomiques) et/ou sur les espèces cibles pourrait s'avérer plus pratique.

➢ Processus d'intérêt

Le fait de pouvoir rechercher des protéines similaires parmi certaines espèces, combiné à la visualisation de l'état d'une espèce par rapport à un processus (possède ou pas le processus), permet d'étendre les possibilités d'analyses et de développement. Nos différents tests ont montré qu'après un temps d'adaptation court, les utilisateurs s'appropriaient rapidement la notion et les fonctionnalités liées au processus d'intérêt.

➢ Filtre sur le nombre d'espèce par clade et intervalle

Afin de mimer les analyses de génomique comparative qui s'appuient sur la recherche des orthologues (Altenhoff *et al.*, 2016), mon outil octroie la possibilité de filtrer les protéines d'une espèce requête qui satisfont des distributions phylogénétiques des protéines similaires déduites des résultats issus de BLAST.

En effet, *BLASTome Dashboard* considère la présence d'au moins un *hit* (avec un Expect inférieur à 10^{-3}), ce qui en fait un outil moins fiable que les outils qui s'appuient sur la prédiction des orthologues. Cependant, malgré les imprécisions liées aux différences entre similitude et orthologie, ce site offre de nombreuses possibilités pour réaliser rapidement et facilement des analyses préliminaires de génomique comparative. A l'avenir, il serait envisageable de le coupler à OrthoInspector (Linard *et al.*, 2011), qui prédit les orthologues, ce qui aboutirait sans doute à une légère perte en terme de vitesse d'obtention des filtres, mais au profit de résultats plus fiables.

Dans ce contexte, nous avons pu vérifier qu'en offrant la possibilité de donner un intervalle, cela permet de rendre la sélection particulièrement souple et de limiter, dans une certaine mesure, les problèmes liés aux gènes mal annotés, non prédits, voire réellement absents. En effet, on peut noter que même si un groupe d'espèces possède un processus, chaque espèce ne va pas forcément posséder toutes les protéines impliquées dans ce processus. De même, pour les espèces cibles ayant subi une perte récente du processus, on observe souvent que seuls une partie des gènes impliqués dans celui-ci est préservée. La sélection par intervalle permet de prendre en compte ces problèmes et d'inclure également des espèces cibles dont on ne connaît pas le statut par rapport au processus.

- Goatools, GO et GOSlim

Le choix de *goatools* comme outil d'enrichissement en termes GO s'est révélé le plus adapté aux besoins de *BLASTome Dashboard.* De plus, j'ai mis en place un système permettant d'extraire les données (totalité des gènes, termes GO associés à ces gènes) nécessaires au calcul, par *goatools*, de l'enrichissement en termes GO. Ces développements se sont imposés face aux particularités des *BLASTome* d'espèces requêtes non modèles (peu de termes GO, mauvaises annotations…) et aux limites des données disponibles au NCBI (nombre limité d'espèces, redondance de certains gènes, multiples variants, …).

Les enrichissements en termes GO permettent d'avoir une idée générale des fonctions moléculaires, localisations cellulaires et processus biologiques présents au sein des protéines filtrées, à condition que des annotations de ces protéines en termes GO existent. Dans ce cadre, quand l'espèce requête n'est pas une espèce modèle, j'ai utilisé les GOSlim de 6 espèces modèles cibles prédéfinies et calculé l'enrichissement *via* les meilleurs *hits* de ces espèces modèles afin d'éviter des situations n'aboutissant à aucun résultat d'enrichissement en termes GO de l'espèce requête non-modèle. J'ai jugé préférable d'avoir des notions fonctionnelles 'approximatives' (*via* l'usage des GOSlim) car il m'a semblé dangereux d'inférer des conclusions fonctionnelles trop précises entre espèces modèles et non-modèles sur la base de *hit* BLAST qui n'implique pas toujours l'orthologie. En effet, au regard de leur éloignement phylogénétique, les fonctions des protéines (cibles et requêtes) peuvent varier d'une espèce à l'autre et même, d'une espèce modèle à

une autre. Ce phénomène a d'ailleurs été observé avec différents jeux de protéines requêtes filtrées qui ne révélaient pas les mêmes enrichissements en fonction de l'espèce modèle.

Cependant, malgré ces limites et approximations, notre stratégie s'est avérée très fructueuse, notamment lors des analyses et sélections réalisées pour étudier les distributions des fonctions ciliaires, analyses qui ont souvent révélé des enrichissements connus et certains, inattendus.

➢ Visualisation

Pour arrêter le visuel de *BLASTome Dashboard*, plusieurs essais ont été effectués. En effet, il s'est avéré difficile d'obtenir une représentation ergonomique, synthétique et intuitive qui rende compte de la diversité et de la complexité de la phylogénie et des liens de similitude entre espèces cibles et requêtes. Un affichage en cadres clades et disques espèces s'est peu à peu imposé et semblait le plus judicieux. En effet, à la place d'une intensité variable code couleur qui fournit une approximation du NbTotal, j'avais d'abord pensé que la taille des disques pourrait être proportionnelle au NbTotal. Or, ceci s'est avéré être une mauvaise idée, compte tenu de la variabilité du NbTotal qui aboutissait à des présentations très hétérogènes et très difficiles à gérer par les programmes.

Cependant, certaines caractéristiques du visuel sont sans doute à améliorer. Malgré le nombre variable d'espèces à l'intérieur des clades, j'ai décidé de représenter les clades par des cadres à taille unique. Or, à l'aide de procédures Javascript ou JQuery, il serait sensé d'envisager de mettre en place des cadres qui s'adaptent mieux au nombre d'espèces dans le clade, ainsi qu'à la taille de l'écran.

➢ Applications

La force de ce client léger se base sur sa capacité à filtrer les protéines similaires entre espèce requête et espèces/clades cibles à la convenance de l'utilisateur. Le filtrage est réalisable à l'intérieur d'un clade, mais également entre clades. Ceci offre tout un champ d'exploitations, au sein desquelles on peut distinguer :

1. Trouver des protéines très spécifiques

Dans la mesure où les filtres exploitent les fichiers BLAST des protéines d'une espèce requête, ce service est donc en mesure de trouver toutes les protéines spécifiques à l'espèce requête (absence de similitude chez toutes les espèces), voire communes à l'espèce requête et à une ou plusieurs espèces cibles.

2. Rechercher des protéines impliquées dans un même processus

Il est connu que les protéines présentant des distributions phylogénétiques de protéines similaires et/ou des orthologues « atypiques » sont fréquemment impliquées dans un même processus. La force de ce client léger est d'offrir un survol rapide de ce type de phénomène.

3. Aider à l'annotation/caractérisation des gènes

Enfin, *BLASTome Dashboard* peut améliorer l'annotation et la caractérisation de groupes de gènes/protéines inconnus. En effet, l'utilisateur ayant accès aux fichiers BLAST du protéome, il peut se servir directement des alignements pour annoter des gènes/protéines inconnus en inférant la fonction trouvée chez les gènes/protéines similaires et cela, fonction uniquement fournie par *BLASTome Dashboard* à ma connaissance, dans un contexte de sélection de groupes de gènes potentiellement impliqués dans un même processus. Ce type d'utilisation pourrait s'avérer particulièrement performant pour les annotations et pourrait s'appliquer aussi bien, aux génomes/protéomes nouvellement séquencés dont l'annotation fait encore défaut qu'à des ré-annotations de protéomes.

➢ Perspectives

Pour faciliter la prise en main de cet outil, un guide d'utilisation et des exemples sont disponibles au cas par cas, cependant un tutoriel vidéo plus détaillé devra également être mis en place.

Dans un autre ordre d'idées, un système de protection des données est sans doute à mettre en place dans le cadre d'espèces récemment séquencées. En effet, de tels résultats doivent rester privés et inaccessibles tant que les analyses ne sont pas finalisées.

Enfin, bien que plusieurs fonctionnalités soient proposées (accès aux fichiers BLAST, lien vers UniProt, enrichissement en termes GO), d'autres aides à l'analyse sont envisageables, notamment celles couplant nos sélections et filtres à des analyses d'interactomes (ensembles des interactions connues ou prédites entre molécules au sein des espèces). L'interactomique est un domaine en plein essor (Mehta *and* Trinkle-Mulcahy, 2016) qui englobe tous types d'interactions entre les différents acteurs (protéines, gènes, lipides, ...) des processus biologiques (voies métaboliques, régulation, ...). Parmi ces types d'interactions, on retrouve les interactions protéines-protéines disponibles sur le site STRING (http://string-db.org/) qui est une banque de données de référence des interactions protéines-protéines et qui s'appuie sur la banque Uniprot (Szklarczyk *et al.*, 2015). Il serait donc envisageable et profitable de pouvoir coupler les résultats de nos sélections aux données d'interactions disponibles dans la banque STRING.

Bibliographie

Altenhoff, A.M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D.A., DeLuca, T., Forslund, K., Huerta-Cepas, J., Linard, B., Pereira, C., Pryszcz, L.P., *et al.* (2016). Standardized benchmarking in the quest for orthologs. Nat. Methods *13*, 425–430.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Dessimoz, C., Gabaldón, T., Roos, D.S., Sonnhammer, E.L.L., Herrero, J., Altenhoff, A., Apweiler, R., Ashburner, M., Blake, J., Boeckmann, B., *et al.* (2012). Toward community standards in the quest for orthologs. Bioinformatics *28*, 900–904.

Frades, I., and Andreasson, E. (2016). Phytophthora infestans specific phosphorylation patterns and new putative control targets. Fungal Biol *120*, 631–644.

Greene, C.S., Tan, J., Ung, M., Moore, J.H., and Cheng, C. (2014). Big data bioinformatics. J. Cell. Physiol. *229*, 1896–1900.

Linard, B., Thompson, J.D., Poch, O., and Lecompte, O. (2011). OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics *12*, 11.

Mehta, V., and Trinkle-Mulcahy, L. (2016). Recent advances in large-scale protein interactome mapping. F1000Research *5*.

Moreno-Hagelsieb, G., and Latimer, K. (2008). Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinforma. Oxf. Engl. *24*, 319–324.

Mount, D.W. (2007). Using the Basic Local Alignment Search Tool (BLAST). CSH Protoc 2007, pdb.top17.

Neumann, R.S., Kumar, S., Haverkamp, T.H.A., and Shalchian-Tabrizi, K. (2014). BLASTGrabber: a bioinformatic tool for visualization, analysis and sequence selection of massive BLAST data. BMC Bioinformatics *15*, 128.

Raffaele, S., Win, J., Cano, L.M., and Kamoun, S. (2010). Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of Phytophthora infestans. BMC Genomics *11*, 637.

Refsnider, J.M., Poorten, T.J., Langhammer, P.F., Burrowes, P.A., and Rosenblum, E.B. (2015). Genomic Correlates of Virulence Attenuation in the Deadly Amphibian Chytrid Fungus, Batrachochytrium dendrobatidis. G3 GenesGenomesGenetics *5*, 2291–2298.

Reiter, L.T., Do, L.H., Fischer, M.S., Hong, N.A., and Bier, E. (2007). Accentuate the negative: proteome comparisons using the negative proteome database. Fly (Austin) *1*, 164–171.

Sun, G., Yang, Z., Kosch, T., Summers, K., and Huang, J. (2011). Evidence for acquisition of virulence effectors in pathogenic chytrids. BMC Evol Biol *11*, 195.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., *et al.* (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447-452.

Tatusova, T.A., and Madden, T.L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. *174*, 247–250.

UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204-212.

Wintersinger, J.A., and Wasmuth, J.D. (2015). Kablammo: an interactive, *web*-based BLAST results visualizer. Bioinformatics *31*, 1305–1306.