Étude de l'histoire évolutive des PI3K et des voies de signalisation associées

Héloïse Philippon

N° d'ordre NNT : 2016LYSE1099

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale E2M2 :
Évolution Écosystèmes Microbiologie Modélisation

Soutenue publiquement le 05/07/2016

par :

Héloïse PHILIPPON

Étude de l'histoire évolutive des PI3K et des voies de signalisation associées

Devant le jury composé de :

Dominique Mouchiroud, PU	Présidente
Emmanuel Douzery, PU	Rapporteur
Claudine Médigue, DR	Rapporteure
Jacques Van Helden, PU	Examinateur
Guy Perrière, DR	Directeur de thèse
Céline Brochier-Armanet, PU	Invitée

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Résumé

L'objectif principal de ma thèse était la caractérisation de l'histoire évolutive des voies de signalisation au travers d'une double approche : (i) l'analyse phylogénétique de leurs composés ; et (ii) l'identification et la caractérisation de leurs interactions par l'analyse des interactomes d'organismes modèles. Or, bien que de nombreux outils soient disponibles pour la reconstruction d'arbres de gènes individuels, peu de méthodes ont été développées pour l'étude évolutive d'un ensemble de protéines impliquées dans un même processus cellulaire. Pourtant, la plupart des protéines exercent leur fonction biologique en interagissant physiquement et directement avec d'autres protéines présentes dans la cellule.

Dans un premier temps, j'ai étudié l'histoire évolutive de la famille des PI3K (***P**hosphatidyl**i**nositol **3**-**k**inases*). Cette première analyse phylogénétique détaillée m'a permis de mettre en place une méthodologie applicable aux voies de signalisation. Un problème important rencontré dans cette étude a consisté en la sélection automatique de transcrits alternatifs et ceci m'a conduit à développer, dans un second temps, un logiciel dédié nommé BATfinder (***B**est **A**ligned **T**ranscript **finder***). Enfin, dans le but d'étudier la voie de signalisation AKT/mTOR, j'ai effectué l'implémentation de la méthodologie validée avec les PI3K. Cette implémentation a pris la forme d'un pipeline automatique nommé EPINe (***E**asy **P**hylogenetics for **I**nteraction **Ne**tworks*). Ce pipeline est théoriquement utilisable pour l'analyse phylogénétique de tout réseau métabolique eucaryote.

Abstract

The main goal of my thesis was the characterization of the evolutionary history of signalling pathways through a twofold approach: (i) the phylogenetic analysis of their components; and (ii) the identification and characterization of their interactions by the analysis of model organisms interactomes. While many tools are available for single gene trees reconstruction, only a few methods have been developed for the study of a set of proteins involved in the same cellular process. However, most of the proteins perform their biological function by physically and directly interacting with the other cellular proteins.

Initially, I studied the evolutionary history of the PI3K family (***P****hosphati-dyl****i****nositol* ***3****-****k****inases*). This first detailed phylogenetic analysis allowed me to set up a methodology suitable for signalling pathways. One of the important problems encountered in this study was the selection of alternative transcripts and this led me to develop a software called BATfinder (***B****est* ***A****ligned* ***T****ranscript* ***finder***). In order to study the AKT/mTOR signalling pathway, I have implemented the methodology previously validated with PI3Ks. This implementation was carried out as an automated pipeline called EPINe (***E****asy* ***P****hylogenetics for* ***I****nteraction* ***Ne****tworks*). This pipeline is theoretically usable for the phylogenetic analysis of any eukaryotic metabolic network.

Il y a des vides, que rien ne peut combler,
Ni l'amour d'un mari, ni les rires d'une enfant.
Malgré ta force, ton courage et ta volonté,
Cette foutue leucémie t'a éteinte lentement.

A jamais mon modèle, c'est bien évidemment,
A toi que je dédie cette thèse, ma chère maman.

Remerciements

J'aimerais commencer cette section par remercier mon directeur de thèse, Guy Perrière, qui m'a beaucoup encouragée et encadrée dans ce travail. Malgré les situations délicates auxquelles j'ai pu être confrontée, je le remercie de m'avoir toujours soutenue et aidée. Merci également à Céline Brochier-Armanet pour avoir autant participé à cette thèse malgré un nombre déjà important d'encadrements à son actif. Merci à tous les deux pour toutes ces discussions scientifiques qui m'ont appris tant de choses sur la phylogénie moléculaire et la bio-informatique. Merci également d'avoir eu la bonne idée d'écrire ce livre, un peu avant mon arrivée, qui m'a tant facilité la découverte de ce domaine passionnant !

Je voudrais ensuite remercier Emmanuel Douzery et Claudine Médigue pour avoir accepté d'être rapporteurs de cette thèse. Merci pour l'ensemble de vos commentaires constructifs qui ont permis d'améliorer ce manuscrit. Merci également à Jacques Van Helden et Dominique Mouchiroud d'avoir accepté de faire parti de mon jury.

Par ailleurs, cette thèse est ce qu'elle est grâce aux conseils avisés des membres de mon comité de pilotage. Merci donc à Christine Brun, Olivier Gandrillon, Christine Lasset et Marc Robinson-Rechavi qui m'ont permis de faire les bons choix. Je tiens également à remercier l'ARC santé 1 de la région Rhône-Alpes et France Génomique pour avoir financé mon travail.

Si ces années de doctorat ont été si enrichissantes c'est grâce à l'ensemble des personnes du LBBE mais aussi grâce aux trop souvent oubliés membres du PRABI que je remercie chaleureusement. Mentions spéciales à Magali, la copine de galère administrative ; Marie, la féministe avertie ; Fanny, la nouvelle mariée ; Aline, la miss Fitch et Amandine, la future maman. Merci à vous pour votre amitié, nos discussions, les midi piscine, les soirées spa, l'organisation de Sciences en Marche, les 80 km en vélo et toutes ces pauses café ! Dans cette dernière catégorie, je tiens à remercier Clothilde, Damien, Guillaume, Ghislain, Rémi, Michel, Cécile, Florent, Yann, Sylvain, Jean-François, Christine, Dominique, Philippe, Marc, Jös, Diamantis et Frédéric pour tous ces sujets plus ou moins loufoques toujours abordés de manière sérieuse et rigoureuse. Merci également aux *Célinettes* Anne, Monique et Najwa, en particulier pour leurs informations sur l'emploi du temps de leur chef. Merci à toi Laurent J., pour les conseils en statistiques et l'organisation des réunions

NGS. Je tiens ensuite à remercier tous mes co-bureaux, passés et présents : Thomas, Erika, Eugénie, Murray, Bérénice, Wandrille et Adil, pour leur soutient au jour le jour. Merci à Raquel de m'avoir permis de faire mes premiers enseignements, ces mémorables TP d'ASBIV. Bien sûr, je voudrais remercier tous les membres des pôles informatique et administratif pour leur aide précieuse et leur bonne humeur. Enfin, pour clôturer ce paragraphe labo, un merci spécial à toutes les personnes qui m'ont offert cette soirée hammam en fin de première année, la plus belle marque de soutient que j'ai jamais reçu.

J'aimerais ensuite dire un immense merci à Anaïs et Gwendoline, mes deux acolytes de l'INSA. Le voyage à Rome et tous ces week end à Rennes ont été des moments de joies précieux pendant cette thèse. Bien que souvent fatiguée, je garderai de bons souvenirs de ces passages à l'Amaryllis ! Merci aussi pour tous ces échanges de mails remplis d'encouragements et de conseils ainsi que pour toutes ces discussions *entre filles*. Merci enfin pour le séjour décompression à Uppsala après la rédaction de ce manuscrit.

Ah mon Emilie, ma conseillère stylistique et ma confidente, comment ne pas te remercier pour ta bonne humeur et ces fous rires depuis le collège. Merci pour toutes ces calories avalées dans les différents restaurants de Lyon pendant ces quatre années de labeurs mais aussi pour tout le reste.

Ce septième paragraphe ira naturellement à mon père, Eric, qui a éveillé mon intérêt pour les sciences dès mon plus jeune âge. Merci d'avoir été là pour nous malgré la peine que tu as pu ressentir dans ta vie. Cette thèse t'es bien évidemment également dédiée puisque sans toi je ne l'aurai sans doute pas entreprise. J'espère que nous aurons de nombreuses autres discussions philosophiques autours d'un café dans la cuisine. Puisses-tu encore éclairer mon chemin durant de longues années.

De ma famille j'aimerais ensuite remercier ma sœur Hélène, son mari Mickaël et leurs deux merveilleux enfants. Merci de m'avoir faite marraine la veille du début de cette thèse et de m'avoir offert tant de moments de détente au milieu des rires d'Anna et de Sacha. Merci également à Josiane, Gérard, Julie, Benjamin, Rachel et Rebecca pour toutes ces après-midi à Saint-Thurin qui m'ont permis de me ressourcer et de me reposer. J'aimerai te remercier particulièrement Josiane, pour tout ce que tu as fait pour moi, notamment tenter de combler ce vide immense. Merci également à mes beaux-parents, Annie et Michel, pour leur gentillesse et leur soutient ainsi que pour s'être intéressés à mon travail. Enfin, merci Françoise, pour

être entrée dans nos vies avec bienveillance et précaution. Merci à toi et papa pour ce pot de thèse qui restera sans doutes dans les annales du laboratoire !

Mais si cette thèse a été menée à terme, c'est essentiellement grâce à toi mon Sébastien, mon *Chep*, mon freluquet, mon PACSman. A mes côtés depuis 9 ans maintenant, je voudrais te remercier pour tout ce que tu as fait pour moi pendant ces années. Merci d'avoir cru en moi, de m'avoir encouragée, soutenue, fait tant rire et rendue si heureuse.

Enfin, merci à tous ceux dont le `ctrl+F` sur leur prénom n'a rien renvoyé, veuillez-m'excuser de vous avoir oublié...

Table des matières

Liste des abréviations

AIC	***A**kaike **I**nformation **C**riterion*
AP-MS	***A**ffinity **P**urification coupled to **M**ass **S**pectrometry*
AP2	***A**daptor **P**rotein **2** complex*
ATG	***A**u**T**opha**G**y related genes*
BaliBASE	***B**enchmark **Ali**gnment data**BASE***
BATfinder	***B**est **A**ligned **T**ranscript **finder***
BIC	***B**ayesian **I**nformation **C**riterion*
BLAST	***B**asic **L**ocal **A**lignment **S**earch **T**ool*
BLOSUM	***BLO**cks **SU**bstitution **M**atrix*
BMGE	***B**lock **M**apping and **G**athering with **E**ntropy*
CAT	***CAT**egories*
CME	***C**lathrin-**M**ediated **E**ndocytosis*
CS	***C**olumn **S**core*
CyTOR	***Cyt**okine **R**eceptor*
DDBJ	***D**NA **D**ata **B**ank of **J**apan*
DS	***D**istance **S**core*
EM	***E**xpectation-**M**aximization*
ENA	***E**uropean **N**ucleotide **A**rchive*
EPINe	***E**asy **P**hylogenetics for **I**nteraction **Ne**tworks*
FSA	***F**ast **S**tatistical **A**lignment*
GPCR	***G**-**P**rotein-**C**oupled **R**eceptor*
GUIDANCE	***GUID**e tree based **A**lig**N**ment **C**onfidenc**E***
HMM	***H**idden **M**arkov **M**odel*
HTP	***H**ight-**T**hrough**P**ut*
HUPO	***H**uman **P**roteome **O**rganization*
IGF1	***I**nsulin-like **G**rowth **F**actor **1***
JAK	***JA**nus **K**inase*
LECA	***L**ast **E**ukaryotic **C**ommon **A**ncestor*
LG	***L**e and **G**ascuel*
LUCA	***L**ast **U**niversal **C**ommon **A**ncestor*
MAFFT	***M**ultiple **A**lignment using **F**ast **F**ourier **T**ransform*
MACSE	***M**ultiple **A**lignment of **C**oding **SE**quences*
MCMC	***M**arkov **C**hain **M**onte-**C**arlo*
MCMCMC	***M**etropolis **C**oupling of **MCMC***
MIC	***M**etaoza, **I**chthyosporea et **C**hoanoflagellida*

MITab	***M**olecular **I**nteraction **Tab**ular format*
MSA	***M**ultiple **S**equences **A**lignment*
MUSCLE	***MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation*
NGS	***N**ext **G**eneration **S**equencing*
NNI	***N**earest **N**eighbor **I**nterchange*
norMD	***n**ew **o**bjective function for sco**r**ing using **M**ean **D**istance*
NR	***N**on **R**edundant*
OCG	***O**verlapping **C**luster **G**enerator*
PALO	***P**rotein **AL**ignment **O**ptimizer*
PAM	***P**oint **A**ccepted **M**utation*
PI3K	***P**hosphatidyl**i**nositol **3**-**k**inases*
PPI	***P**rotein-**P**rotein **I**nteraction*
PRANK	***PR**obabilistic **A**lig**N**ment **K**it*
PSI	***P**roteomics **S**tandards **I**nitiative*
PSI-MI	***P**roteomics **S**tandards **I**nitiative **M**olecular **I**nteraction* **XML**
PSICQUIC	***PSI** **C**ommon **QU**ery **I**nterfa**C**e*
RAS	***RA**t **S**arcoma*
RTK	***R**eceptor **T**yrosine **K**inases*
SGD	***S**accharomyces **G**enome **D**atabase*
SNP	***S**ingle **N**ucleotide **P**olymorphism*
SOCS	***S**uppressor **o**f **C**ytokine **S**ignaling*
SOLiD	***S**equencing by **O**ligonucleotide **Li**gation and **D**etection*
SP	***S**um-of-**P**airs*
SPC	***S**um-of-**P**airs **C**olumn score*
SPR	***S**ubtree **P**runing and **R**egrafting*
STAT	***S**ignal **T**ransducer and **A**ctivator of **T**ranscription*
T-COFFEE	***T**ree-based **Co**nsistency **O**bjective **F**unction **F**or alignment **E**valuation*
TrimmAL	***Trim**ming **AL**ignment*
UPGMA	***U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic mean*
VPS	***V**acuolar **P**rotein **S**orting **34***
WAG	***W**helan **A**nd **G**oldman*
WormPD	***Worm** **P**rotein **D**atabase*
WOT	***W**ith **O**ther **T**ranscripts*
Y2H	***Y**east two-**H**ybrid*
YPD	***Y**east **P**roteome **D**atabase*

Chapitre 1

Introduction biologique

Classer les organismes vivants, voici un défi auquel l'Homme s'est attelé très tôt. Dès l'antiquité le vivant fut classé selon divers critères de similarités, qu'ils soient physiques, physiologiques ou encore utilitaires (pour les plantes médicinales par exemple). Au fil des siècles, les organismes vivants ont été décrits, répertoriés et catégorisés par les naturalistes au gré de leurs voyages et de leurs observations. Cependant, ce n'est qu'au XVIIIème siècle qu'une classification globale du vivant ainsi qu'une uniformisation de la nomenclature furent proposées. Considéré comme le premier ouvrage de classification traditionnelle des espèces, *Systema Naturae* [1], publié en 1735 par Carl von Linné, divise le vivant en deux groupes majeurs : le règne animal et le règne végétal. Le système de regroupement en règnes, classes, ordres, genres et espèces proposé par Linné constitua dès lors la base de la classification actuelle. Néanmoins, cette codification fut rapidement modifiée pour arriver à sept niveaux que sont : le *règne*, l'*embranchement*, la *classe*, l'*ordre*, la *famille*, le *genre*, et l'*espèce*. De nos jours de nombreux rangs intermédiaires supplémentaires existent tels que les sous-classes et les super-familles [2].

Dans la vision des naturalistes du XVIIIème siècle, bien qu'ils soient regroupés en fonction de leurs similarités, les êtres vivants ne partagent pas de relations de parenté. Cette notion, bien qu'en partie déjà évoquée par les Lumières, n'est formellement introduite qu'en 1809 par Jean-Baptiste Lamarck dans son ouvrage intitulé *Philosophie Zoologique* [3]. La théorie de la transformation des espèces, avec notamment l'hypothèse de génération spontanée, que Lamarck présente sont néanmoins incomplètes et erronées. Ses interrogations sur la variabilité des individus mèneront Charles Darwin à proposer, un demi siècle plus tard, la théorie de la sélection naturelle. Dans son célèbre ouvrage *On the Origin of Species by Means of*

Natural Selection [4], dont la première version a été publiée en 1859, Darwin pose les fondements de la phylogénie moderne, avec l'idée que les espèces actuelles sont le résultat de la longue histoire du Vivant.

Pour matérialiser ces relations de parenté, Charles Darwin va introduire la représentation du vivant sous forme d'arbre, à l'image des arbres généalogiques. Sa première esquisse, très connue, provient de ses notes personnelles, datées de 1837 (Figure 1.1). Il est intéressant de noter que l'arbre de la vie présenté dans *On the Origin of Species by Means of Natural Selection* [4] illustre déjà le processus d'extinction des lignées.

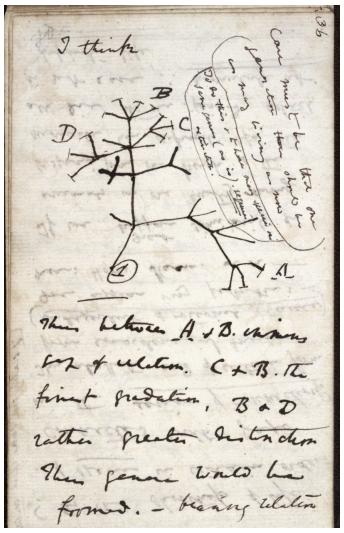

FIGURE *1.1* – ***Premier arbre phylogénétique théorique réalisé dans* First Notebook on Transmutation of species *par Darwin en 1837.***

Bien qu'au départ très controversée, la théorie de Darwin fut petit à petit adoptée par d'autres scientifiques. Les arbres phylogéniques se sont alors diversifiés, complexifiés et étoffés avec le but ultime d'obtenir l'arbre du Vivant complet. Cependant, ces phylogénies, puisque principalement basées sur des critères morphologiques, ne concernent alors que les animaux et les plantes. Les micro-organismes, bien que découverts deux siècles plus tôt par Antoni van Leeuwenhoek [5], ne sont en effet pas considérés dans ces études.

Au milieu du XXème siècle, la classification du vivant a connu un tournant considérable avec le développement de la méthodologie de la systématique phylogénétique développée par Willi Hennig [6]. En effet, l'obtention des premières séquences biologiques (protéines dans les années 1950, puis ADN au milieu des années 1970) a permis l'émergence d'une nouvelle classification fondée sur des similarités géniques et non plus morphologiques. La phylogénie moléculaire devient

alors la méthodologie privilégiée pour la classification des espèces, notamment pour les procaryotes et les micro-organismes pour lesquels les critères morphologiques ne sont pas pertinents. L'ensemble des classifications furent alors révisées et une nouvelle division du Vivant vit le jour. Dans leur article publié 1990 [7], Woese *et al.* proposent en effet une division en trois domaines que sont les *Archées*, les *Bactéries* et les *Eucaryotes*. Mon travail de thèse s'est focalisé sur ces derniers.

1.1 Les Eucaryotes

1.1.1 Diversité et phylogénie

Par opposition aux Procaryotes, les Eucaryotes se définissent essentiellement par la présence d'un noyau et d'organites cellulaires (mitochondries, plastes, etc.). Présents dans tous les milieux (terrestre et aquatique) ; uni- ou pluri-cellulaires, ce domaine regroupe des organismes à la diversité morphologique importante. Leur taille varie de quelques μm pour les Eucaryotes unicellulaires à environ 33 mètres pour la baleine bleue [2].

La classification actuelle divise les Eucaryotes en cinq groupes majeurs [8, 9] (Figure 1.2). Les animaux (ou Metazoa) et les champignons (ou Fungi) font ainsi partie du groupe des Opisthokonta tandis que les plantes constituent le groupe des Archaeplastida. Moins connus des non spécialistes, les Excavata, les Amoebozoa et les SAR (Stramenopiles, Alveolata et Rhizaria [8]) font également partie des Eucaryotes. Les représentants les plus étudiés de ces groupes sont respectivement *Leishmania donovani* (responsable de la leishmaniose), *Dictyostelium discoideum* et *Plasmodium falciparum* (responsable du paludisme). Enfin, divers protistes, tels que les Apusozoa ou les Haptophyta, font également partie du domaine des Eucaryotes mais ne sont pas classés parmi ces cinq groupe majeurs (en gris sur la Figure 1.2).

La phylogénie des Eucaryotes n'est pas complètement résolue et des zones d'ombres subsistent. Le placement de la racine des Eucaryotes ainsi que sa datation sont en particulier toujours débattues. Nommé LECA (***L**ast **E**ukaryotic **C**ommon **A**ncestor*) en référence à LUCA (***L**ast **U**niversal **C**ommon **A**ncestor*), la datation du dernier ancêtre commun des Eucaryotes a en effet fait l'objet de plusieurs articles contradictoires au début des années 2000 [10, 11, 12]. En 2004, deux études indépendantes ont ainsi daté son apparition sur Terre à respectivement 2 309 mil-

lions d'années [13] (avec un intervalle de confiance se situant entre 2115 et 2503 millions d'années) et 1085 millions d'années [14] (950-1259). Dans une étude plus récente, Eme *et al.* [15] se sont donc intéressés aux facteurs pouvant impacter ces estimations. De façon surprenante, l'étude a démontré que le placement de la racine n'a que peu d'effet sur les datations inférées tandis que les méthodes d'inférence testées ont abouti à des différences de l'ordre de 650 millions d'années. Au moyen de méthodes d'inférence bayésiennes plus pointues (voir section 2), leur étude utilisant 159 protéines provenant de 85 taxons différents et 19 fossiles, a daté l'apparition de LECA à 1007-1898 millions d'années, ne permettant donc pas d'améliorer la précision des précédentes estimations. Toutes ces analyses s'accordent néanmoins sur un point, l'hypothèse avancée en 2000 [16] selon laquelle la diversification des Eucaryotes s'est faite à la fois très tôt (moins de 300 millions d'années après l'apparition de LECA) et très rapidement [15].

En 2012, Adl *et al.* [8] ont publié une revue rassemblant et discutant les informations disponibles sur les relations phylogénétiques entre les différents grands groupes Eucaryotes. Ils ont ainsi proposé l'arbre de la Figure 1.2 qui m'a servi de référence lors de l'étude phylogénétique des protéines PI3K (voir chapitre 4).

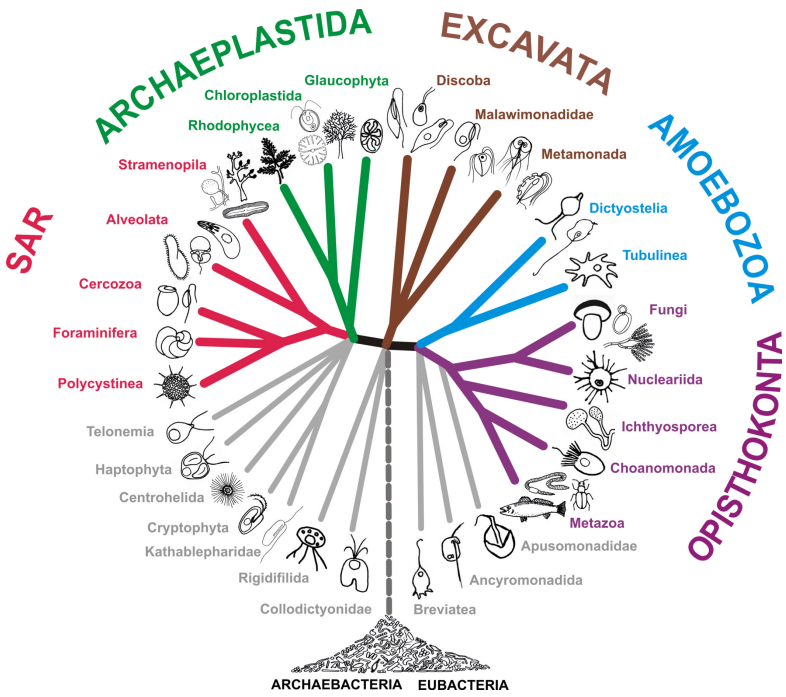

FIGURE *1.2* – ***Phylogénie des Eucaryotes proposée par Adl et al. [8].***

Concernant le placement de LECA dans l'arbre des Eucaryotes, l'hypothèse la plus communément admise de nos jours considère qu'il se situe à l'embran-

chement séparant les Unikonta (Amoebozoa et Opisthokonta) des Bikonta (SAR, Archaeplastida et Excavata) [17, 9]. Il s'agit de l'hypothèse que j'ai utilisée lors de l'enracinement et de l'interprétation des arbres phylogénétiques obtenus dans mes différentes analyses.

1.1.2 Spécificités génomiques

La cellule eucaryote présente une organisation complexe en différents compartiments (Figure 1.3). En plus du noyau, son cytoplasme est composé de deux autres compartiments primordiaux : l'appareil de Golgi et le réticulum endoplasmique qui permettent entre autres la production de protéines à partir de l'ADN cellulaire. Les fonctions énergétiques de la cellule eucaryote sont, quant à elles, assurées par des organites spécifiques : les mitochondries et les chloroplastes (chez les plantes).

Figure omise pour des raisons de droits d'auteur.

FIGURE *1.3* – ***Représentation de l'organisation d'une cellule eucaryote [18].***

Bien qu'il soit majoritairement contenu dans le noyau de la cellule, de l'ADN est également détecté dans les mitochondries et les chloroplastes. Support universel de l'information génétique, la molécule d'ADN évolue et présente plusieurs différences organisationnelles entre les trois grands domaines du Vivant. En effet, si dans tous les organismes vivants l'ADN est composé par l'enchaînement des quatre bases nucléotidiques, l'Adénine (A), la Thymine (T), la Cytosine (C) et la Guanine (G) ; il se présente le plus souvent sous forme circulaire chez les bactéries tandis qu'il est principalement compacté en chromosomes dans les cellules eucaryotes.

Par ailleurs, contrairement à leurs homologues procaryotes, les gènes eucaryotes sont constitués d'une alternance d'exons et d'introns. Si les premiers sont trans-

crits en ARN puis traduits en protéines, les introns sont, eux, éliminés au cours de l'*épissage*. Spécifique des organismes eucaryotes, ce mécanisme a lieu avant la phase de traduction et permet, à partir d'un même gène, de produire différents ARN messagers (ou transcrits alternatifs). Ces derniers sont ensuite traduits en différentes protéines pouvant êtres fonctionnelles ou non (Figure 1.4). Cette particularité génomique permet ainsi aux gènes eucaryotes de pouvoir générer plusieurs protéines à partir de la même séquence d'ADN.

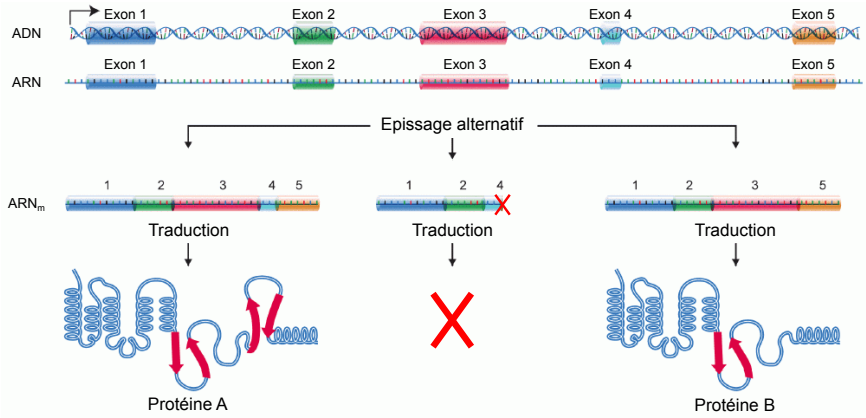

FIGURE *1.4* – ***Illustration de l'épissage alternatif (adapté de [19]).***

Le phénomène d'épissage alternatif est très courant chez les Eucaryotes. Environ 20% des gènes de plantes sont impactés [20] tandis que près de 90% des gènes humains produisent plus d'un ARNm [21]. On distingue classiquement cinq modèles majeurs d'épissage alternatifs (Figure 1.5) :

- L'*exon skipping* : un ou plusieurs exons sont entièrement exclus lors de la maturation de l'ARN pré-messager. Il s'agit du processus majoritaire chez l'Homme et la souris puisque environ 40% des transcrits alternatifs résultent de cet événement [22].
- La *rétention d'intron* : un ou plusieurs intron(s) ne sont pas éliminés lors de l'épissage. Leur traduction pouvant engendrer la présence prématurée d'un codon stop ou un décalage du cadre de lecture.
- L'*épissage alternatif en 3'* : l'exon peut « débuter » à différents endroits de la séquence d'ADN.
- L'*épissage alternatif en 5'* : l'exon peut se « terminer » à différents endroits de la séquence d'ADN.
- Les *exons mutuellement exclusifs* : lorsque deux exons sont mutuellement exclusifs, un seul des deux est retenu dans l'ARNm mature.

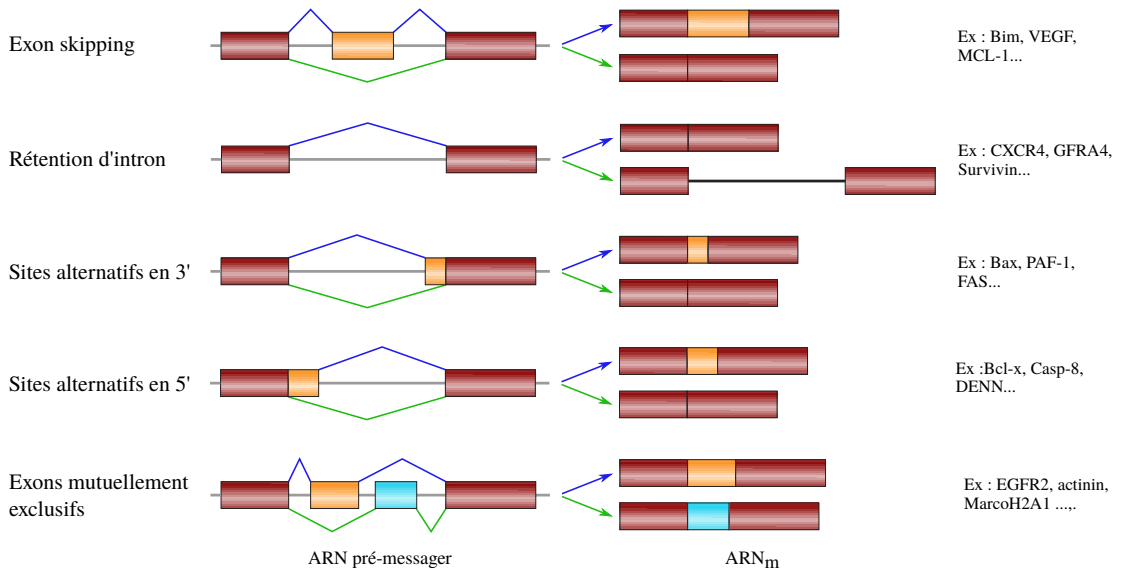

FIGURE *1.5* – ***Les cinq grandes catégories d'épissage alternatifs (d'après [23] et [24]).*** Les rectangles colorés et les lignes grisées représentent respectivement les exons et les introns du gène. Des exemples de gènes affectés par ces différents processus sont indiqués à droite de la figure.

Comme démontré dans le chapitre 3 de ce manuscrit, le mécanisme d'épissage est à l'origine de plusieurs biais lors de l'analyse phylogénétique de gènes eucaryotes. Cette particularité impose ainsi un tri supplémentaire afin de ne sélectionner qu'une seule séquence protéique par gène homologue considéré. Une partie de mon travail de thèse a consisté à automatiser cette sélection.

1.1.3 Données moléculaires disponibles

De nos jours, l'étude des relations entre les espèces s'effectue principalement à travers l'analyse de données moléculaires. Les séquences, *i.e.* un enchaînement orienté et ordonné de nucléotides (séquence nucléique) ou d'acides aminés (séquence protéique), constituent ainsi le matériau de base des phylogénéticiens moléculaires. Depuis une dizaine d'années le nombre de séquences disponibles a explosé grâce aux développement de techniques de séquençage pointues issues de plus de 70 ans de recherche et développement. Cette avalanche de données génomiques a rendu possible l'analyse phylogénétique plus précise de nombreux groupes taxonomiques dont la phylogénie n'était jusqu'alors peu résolue.

Si sa structure en double hélice fut résolue en 1953 par l'américain James Watson et l'anglais Francis Crick [25], la découverte de l'ADN est bien plus ancienne. En 1869, Johannes Friedrich Miescher, isole à partir de pus une substance qu'il nomme nucléine (qui vient du noyau) [26]. Le terme d'*acide nucléique* sera intro-

duit trente ans plus tard par le biologiste Richard Altman qui montre que cette molécule contient de l'acide phosphorique [27]. Les quatre bases azotées constituant l'ADN sont quant à elles caractérisées biochimiquement en 1919 par Phoebus Levene [28]. Ce dernier pense alors que l'ADN est une molécule assez courte et que l'enchaînement des quatre nucléotides est cyclique.

La première séquence biologique complète, obtenue en 1951 par Frederic Sanger et Hans Tuppy, fut celle de l'insuline bovine [29, 30]. Cependant, il faut attendre le début des années 1970 pour obtenir la première séquence d'ADN, composée de 17 nucléotides [31]. Fondée sur leur méthode du *Plus and Minus* de 1975 [32], Sanger *et al.* publient en 1977 la première longue séquence nucléique, correspondant au génome entier du bactériophage ϕX174, d'une taille de 5375 nucléotides [33]. Néanmoins, ce n'est qu'environ vingt ans plus tard que le premier génome complet non viral est publié. En effet, en 1995, Fleischmann *et al.* [34] obtiennent le génome bactérien d'*Haemophilus influenzae*, composé de près de 2 millions de bases. Quant au premier génome eucaryote entièrement séquencé, celui de la levure *Saccharomyces cerevisiae* ; il fut publié un an plus tard par Goffeau *et al.* [35] (voir Figure 1.6). Le premier génome eucaryote multi-cellulaire, celui du nématode *Caenorhabditis elegans*, est quant à lui publié en 1998 par le consortium dédié à ce projet [36].

Figure omise pour des raisons de droits d'auteur.

FIGURE *1.6* – ***Chronologie des analyses génomiques à grande échelle [37].***

Enfin, la première version (incomplète) du génome humain fut dévoilée en 2001 par Venter *et al.* [38] et McPherson *et al.* [39]. Si celle-ci nécessita une décennie de travaux et la collaboration d'une vingtaine d'équipes de recherche [37], l'appari-

tion du séquençage haut débit puis de la nouvelle génération de séquenceurs (***N**ext **G**eneration **S**equencing* ou NGS) dans les années 2000 a permis de réduire drastiquement le temps et le coût des séquençages génomiques. En effet, il est aujourd'hui possible d'obtenir un génome humain en seulement quelques jours.

On appelle « séquenceurs » les appareils dédiés à la détermination de l'ordre d'enchaînement des résidus d'une séquence moléculaire. Les séquenceurs d'aujourd'hui sont fondés sur des méthodes différentes dont le principal dilemme est de trouver le bon compromis entre erreurs de séquençage et taille des fragments d'ADN ou d'ARN (*lectures* ou *reads*) obtenus. En effet, la fiabilité des méthodes actuelles diminue généralement avec la longueur de la séquence produite. D'un autre côté, si les séquences courtes sont théoriquement plus fiables, elles rendent l'étape d'*assemblage* du génome beuacoup plus compliquée. Brièvement, les premiers séquenceurs nouvelle génération commercialisés furent :

- Le 454 de Roche : développé dès 2005 [40], il utilise le principe de pyroséquençage. La nouvelle version, baptisée GS-FLX, permet d'obtenir des lectures d'une longueur maximale de 900-1000 paires de bases (ou pb) [41].
- Le Genome Analyzer et le Solexa d'Illumina : fondés sur la méthode de séquençage par synthèse, leur commercialisation débuta en 2006 et 2007. A chaque cycle, les quatre bases azotées marquées par fluorescence sont ajoutées au milieu. Lors de l'appariement de la base complémentaire au fragment d'ADN, la fluorescence correspondante est libérée et détectée grâce à une caméra haute résolution [42]. Les lectures produites par ces méthodes ont une longueur de l'ordre de 150-300 pb.
- Le SOLiD d'Applied Biosystems : mise au point dès 2005 [43], la méthodologie SOLiD (***S**equencing by **O**ligonucleotide **Li**gation and **D**etection*) ne fut commercialisée qu'à partir de 2007 [44]. Le *5500 Series Genetic Analysis Systems* produit aujourd'hui des lectures d'environ 70 pb [45].

Depuis, le développement de nouveaux séquenceurs s'est accéléré et de nombreux modèles sont commercialisés chaque année par ces trois grandes compagnies.

En phylogénie moléculaire, les NGS ont rendu possible le séquençage de centaines de nouveaux génomes complets, permettant ainsi d'augmenter l'échantillonnage taxonomique des études et la précision des conclusions effectuées.

1.2 Les voies de signalisation

1.2.1 Définitions

On désigne par *voie de signalisation* un ensemble de composés interagissant et permettant la réponse de la cellule à un signal extracellulaire. L'activation d'une voie de signalisation se fait généralement suite à la reconnaissance d'un signal par une protéine membranaire. Il peut s'agir de la libération d'un agent chimique dans le milieu extracellulaire ou de la modification de paramètres physiques tels que la tension, la concentration en sel, etc. Chez les organismes pluricellulaires ces voies permettent notamment la cohésion et la communication des cellules entre elles.

Le terme *transduction du signal* est apparu pour la première fois dans un article de Rensing en 1972 [46]. Néanmoins, les avancées majeures dans la compréhension du phénomène sont attribuées à Martin Rodbell, qui, dès 1970, s'intéressa à l'effet du glucagon sur la production d'AMP cyclique (AMPc) par les cellules de foie de rat [47, 48]. Ses études démontrèrent que la présence de glucagon dans le milieu extracellulaire augmente l'activité de l'adenylate cyclase, une protéine cytoplasmique. En 1980, Rodbell proposa un modèle générique dans lequel le signal est transmis grâce à un ensemble de trois composés : un récepteur (R), un élément régulateur (N) et un élément catalytique (C) (Figure 1.7). Dans le cas de la production d'AMPc, l'élément régulateur correspond aux protéines régulatrices nommées *protéines G* tandis que l'élément catalytique est l'adenylate cyclase. Le prix Nobel de physiologie lui fut attribué en 1994 pour l'ensemble de ses travaux sur la transduction du signal.

Figure omise pour des raisons de droits d'auteur.

FIGURE *1.7* – ***Modèle générique de transduction du signal proposé par Rodbell en 1980 [48].*** R : récepteur, N : élément régulateur et C : élément catalytique.

Introduit récemment par Sanchez *et al.* [49], le terme d'*interactome* désigne, de manière plus générale, l'ensemble des interactions moléculaires d'un organisme telles que les interactions protéine-ADN, protéine-ARN ou encore protéine-protéine (ou PPI pour ***P**rotein-**P**rotein **I**nteraction*). Dans ce travail je ne me suis exclusivement intéressée à ces dernières. Dans la suite de ce manuscrit, un interactome désignera donc uniquement l'ensemble des PPI d'un organisme.

D'un point de vue mathématique, il s'agit d'un *graphe*, pouvant être orienté, dans lequel les nœuds représentent des protéines et les arêtes traduisent l'existence d'une interaction entre les deux protéines qu'elles relient. Les homopolymères sont donc représentés par une boucle sur la protéine concernée (Figure 1.8).

Figure omise pour des raisons de droits d'auteur.

FIGURE *1.8* – ***Exemple d'interactome et de six sous-réseaux de celui-ci [50].*** La boucle sur le nœud TrmD traduit la capacité de la protéine tRNA (guanine-N(1)-)-methyltransferase à former un homodimère [51].

S'il existe différents types d'interactions et d'interacteurs biologiques, il existe également des différences dans les qualités des prédictions d'interaction. En effet, il existe de nombreuses méthodes de détection d'interaction qui possèdent chacune leurs caractéristiques et leur défauts (voir Table 1.1). Les plus utilisées sont :

- La méthode de double hybridation (ou Y2H pour ***Y**east two-**H**ybrid*) : décrite en 1989 par Fields et Song [52], cette méthode de génie moléculaire *in vivo* utilise la levure (*S. cerevisiae*) comme modèle. Fondée sur un système de proie et d'appât, l'expression de gène rapporteur ne s'effectue que si les deux protéines considérées interagissent. Cette technique fut appliquée à grande échelle pour établir les interactomes de la levure [53, 54], du nématode [55, 56],

de la drosophile [57] et de l'Homme [58, 59]. Néanmoins cette technique donne de nombreux faux positifs [60, 61, 62].

- La méthode de coimmmunoprécipitation couplée à la spectrométrie de masse (ou AP-MS pour ***A**ffinity **P**urification coupled to **M**ass **S**pectrometry*) : utilisée dès la fin des années 1990 [63] puis à large échelle au début des années 2000 [64, 65], cette technique est l'une des plus fiables. La première étape, la coimmmunoprécipitation, consiste à isoler et à immobiliser une protéine *cible* puis à la *présenter* aux protéines partenaires potentielles. Après plusieurs passages de purification, les complexes obtenus sont analysés et identifiés par spectrométrie de masse [65].

- Les méthodes *in silico* : leur principal avantage est la prédiction d'interactions à grande échelle et à moindre coût. Elles peuvent être fondées sur le contexte génomique, la coévolution ou encore la coexpression des gènes [66]. La première méthode utilisant l'homologie des protéines, baptisée méthode de la Pierre de Rosette, a été développée en 1999 par Marcotte *et al.* [67].

Approche	**Technique**	**Description / caractéristiques**
In vitro	Coimmmunoprécipitation en tandem couplée à la spectrométrie de masse (TAP-MS)	Double marquage du locus chromosomique de la protéine d'intérêt, suivi d'un procédé de purification en deux étapes puis d'une analyse par spectrométrie de masse.
	Chromatographie d'affinité	Méthode très sensible, peut détecter les interactions faibles.
	Coimmmunoprécipitation	Fonctionne à partir d'extraits cellulaires entiers ; *i.e.* dans lesquels les protéines sont sous leur forme native.
	Biopuces	Méthode qui permet l'analyse simultanée de milliers de paramètres en une seule expérience.
	Complémentation de fragments protéiques	Détecte les PPI entre protéines de poids moléculaires quelconques. Fonctionne également quand elles sont exprimées à niveau endogène.
	Phage display	Analogue à la méthode double hybride mais avec l'utilisation d'un bactériophage.
	Cristallographie par rayons X	Permet de visualiser la structure des protéines à l'échelle atomique. Améliore la compréhension de l'interaction et de la fonction des protéines.
	Spectroscopie RMN	Méthode détectant les très faibles interactions protéine-protéine.
In vivo	Double hybridation (Y2H)	Généralement utilisée pour détecter les interactions d'une protéine d'intérêt par criblage. Méthode haut débit.
	Létalité synthétique	Repose sur les interactions fonctionnelles plutôt que physiques.
In silico	Approche par orthologie	Fondée sur l'hypothèse de conservation d'interaction entre protéines orthologues.
	Approche par paire de domaines protéiques	Prédit une interaction entre deux protéines si elles possèdent des domaines protéiques connus pour interagir.
	Similarité de structure	Méthode fondée sur la comparaison de structure (primaire, secondaire ou tertiaire) des protéines.
	Fusion de gènes	Aussi appelée méthode de la Pierre de Rosette.
	Double hybride *in silico* (I2H)	Fondée sur l'hypothèse que les protéines interagissant doivent co-évoluer pour maintenir la fonction de la protéine.
	Arbre phylogénétique	Méthode fondée sur l'histoire évolutive des gènes.
	Expression des gènes	Des protéines groupées selon un même motif d'expression ont plus de chances d'interagir entre elles.

TABLE *1.1* – ***Principales méthodes de détection d'interactions protéine-protéine (traduit de [62]).***

Les méthodes de détection de PPI à grande échelle (ou méthodes HTP pour ***H**ight-**T**hrough**P**ut*) ont été développées dès le début des années 2000. Néanmoins, le peu d'interactomes de bonne qualité disponibles de nos jours ne concernent que les organismes modèles tels que l'Homme, la levure, le nématode, la drosophile ou encore la plante *Arabidopsis thaliana*. De plus, les méthodes de détection n'étant pas de qualités équivalentes, les interactomes publiés pour une même espèce ne sont pas toujours directement comparables, c'est pourquoi leur taille n'augmente pas forcément avec le temps. En effet, la confiance que l'on peut accorder à une interaction déterminée par approche *in silico* n'est pas la même qu'une interaction détectée par Y2H. Enfin, le type (protéine-protéines, protéines-ADN, etc.) et la nature (association physique, phosphorylation, etc.) des interactions peuvent varier d'une étude à l'autre, expliquant la variation dans les nombres d'interactions et de protéines reportés.

Le premier interactome disponible fut celui de *S. cerevisiae*, publié en 1997 par Fromont-Racine *et al.* [53], dont la version actualisée en 2000 par Schwikowski *et al.* [60] était composée de 2709 interactions impliquant 2039 protéines. Pour centraliser ces nouvelles données, la YPD (***Y**east **P**roteome **D**atabase*) fut spécialement créée en 1999 par Hodges *et al.* [68]. Concernant le nématode, l'interactome publié en 2004 par Li *et al.* [56], était composé de 5460 interactions impliquant 2898 protéines. Une version plus récente, publiée en 2009 par Simonis *et al.* [69], était quant à elle composée de 2528 protéines formant 3864 interactions dites de très bonne qualité. Analogue à la YPD, la base de donnée WormPD (***Worm** **P**rotein **D**atabase*) consacrée au données d'interaction du nématode fut publiée par le même groupe de chercheurs en 2000 [70]. Composée de 4679 protéines formant 4780 interactions, la première version de bonne qualité de l'interactome de *Drosophila melanogaster* fut pour sa part publiée en 2003 par Giot *et al.* [57]. Le seul interactome d'Eucaryote non Opisthokonta, celui de la plante *A. thaliana*, fut quant à lui publié en 2011 par le *Arabidopsis Interactome Mapping Consortium* [71]. Les auteurs avaient alors détecté environ 6200 interactions de haute qualité impliquant environ 2700 protéines. Pour finir, concernant l'Homme, l'estimation du nombre d'interactions protéine-protéine varie d'environ 130000 [72] à 650000 [73] selon les études, mais seulement environ 60000 interactions ont été expérimentalement identifiées [74]. Le *Human Interactome Project* [75] a ainsi publié en 2014 l'interactome de référence humain [76], composé d'environ 30000 interactions considérées comme très fiables. Sélectionnant 13000 protéines humaines, les auteurs ont testé par méthode Y2H l'ensemble des 84.5 million d'interactions binaires possibles et en ont détecté

exactement 13944 (impliquant 4303 protéines). Ainsi, bien que le nombre d'interactions soient estimés à 25000-30000 pour la levure, 200000 pour le nématode, 75000 pour la drosophile et à plus de 130000 pour l'Homme (Table 1.2), très peu de ces interactions ont été expérimentalement vérifiées et validées [73, 77, 78].

Espèce	**Jeu de données**	**Nœuds**	**Arrêtes**	$\hat{M}_{\mathcal{N}}$	**IC** 95%
S. cerevisiae	Uetz *et al.* [54]	1 328	1 389	28 472	26 650 - 30 460
	Ito *et al.* [79]	3 245	4 367	14 940	13 500 - 16 650
	Ho *et al.* [64]	871	694	33 234	31 750 - 34 810
	Gavin *et al.* [80]	726	367	25 391	23 280 - 27 710
	DIP	4 959	17 226	25 229	24 100 - 26 440
D. melanogaster	Giot *et al.* [57]	6 991	20 240	75 506	72 700 - 78 400
	Stanyon *et al.* [81]	362	1 611	2 505 545	2 192 900-2 843 800
	Fromstecher *et al.* [82]	1 200	1 657	211 877	180 419 - 248 640
	DIP	7 451	22 636	74 336	71 700 - 77 100
C. elegans	Li *et al.* [56]	2 622	3 955	242 578	221 850 - 265 700
	DIP	2 638	3 970	240 544	220 030 - 263 270
H. sapiens	Stelz *et al.* [59]	1 665	3 083	646 557	588 990 - 706 640
	Rual *et al.* [58]	1 527	2 529	631 646	564 460 - 703 830
	DIP	1 085	1 346	672 918	625 170 - 722 670

TABLE *1.2* – ***Propriétés des différents interactomes eucaryotes et estimations de leur taille totale $\hat{M}_{\mathcal{N}}$, par Stumpf et al. en 2008 [73].***

Si les premiers interactomes ont fait l'objet de publications dédiées, l'interactome d'un organisme d'intérêt est aujourd'hui obtenu en interrogeant les principales bases de données interactomiques (Table 1.3). En effet, les bases telles que IntAct [83], BioGrid [84] ou MINT [85] recensent les interactions détectées dans de nombreuses études indépendantes pour différents organismes.

Banque	**Caractéristiques**	**Références**
STRING	Prédit des associations fonctionnelles entre protéines	Szklarczyk *et al* [86], 2011
BioGRID	Interactions géniques et protéiques issues de la littérature	Chatr-Aryamontri *et al* [84], 2015
IntAct	Banque de données d'interactions moléculaires	Kerrien *et al* [87], 2012
MIPS	Banque de données d'interactions moléculaires (initialement pour la levure puis pour les mamifères et les plantes)	Mewes *et al* [88], 2011
MINT	Fondée sur les données de la littérature	Licata *et al* [85], 2012
HPRD	PPI et modifications post-traductionnelles	Goel *et al* [89], 2012
SKEMPI	Uniquement pour les complexes dont la structure 3D est résolue	Moal et Fernandez-Recio [90], 2012
3did	Classification d'interactions entre domaines protéiques PFAM	Mosca *et al* [91], 2014
IBIS	Inférence de sites d'interaction entre partenaire holmologues	Shoemaker *et al* [92], 2012
PRISM	Clustering des interfaces basé sur les similarité de structure et d'évolution	Baspinar *et al* [93], 2014
InterEvol	Interfaces avec structure 3D connue, y compris les interologues structurels et les alignements d'orthologues	Faure*et al* [94], 2012

TABLE *1.3* – ***Principales bases de données d'interaction protéine-protéine (adapté et traduit de [74]).***

Dans le but de normaliser les descriptions dans le domaine de la protéomique et pour faciliter la comparaison et la recherche de données, la PSI (***P****roteomics*

S**tandards **I**nitiative*) a été créée en 2002 par le groupe HUPO (H**uman **P**roteome **O**rganization*) [95]. Dès 2004, le consortium spécifie un standard pour la description des interactions protéines-protéines baptisé PSI-MI XML (***P**roteomics **S**tandards **I**nitiative **M**olecular **I**nteraction* **XML**) [96], qui est par la suite simplifié en format MITAB (***M**olecular **I**nteraction **Tab**ular format*) [97]. Un service web dédié à l'interrogation simultanée des 27 principales banques de données de PPI, nommé PSICQUIC (***PSI** Common **QU**ery **I**nterfa**Ce***), a quant à lui été développé en 2011 [98]. Ainsi, l'utilisation de PSICQUIC sur les bases de données IntAct [83], BioGrid [84], MINT [85] et DIP [99] pour cinq organismes modèles d'Opisthokonta conduit à l'obtention d'interactomes aux tailles très différentes. La Table 1.4 montre par exemple qu'en ne sélectionnant que les interactions dites *physiques* et *directes* (codes PSI-MI MI:0915/MI:0218 et MI:0407 respectivement) la taille de l'interactome de la levure varie entre 15 803 et 119 149 tandis que son ordre varie entre 4 142 et 5 846.

Espèce	**BioGRID**		**IntAct**		**DIP**		**MINT**	
	PPI	Protéines	PPI	Protéines	PPI	Protéines	PPI	Protéines
H. sapiens	214 912	15 464	92 803	13 103	16 137	4 988	18 671	5 315
D. melanogaster	37 509	7 971	27 698	8 552	23 873	7 892	367	256
C. elegans	6 221	3 205	19 743	9 382	5 642	3 177	574	486
S. cerevisiae	119 149	5 846	15 803	4 142	31 529	5 154	39 279	5 170
A. thaliana	39 517	9 090	14 443	4 920	319	246	418	242

TABLE *1.4* – ***Nombre d'interactions binaires physiques et directes disponibles en Mars 2016 pour cinq espèces modèles dans les quatre principales banques de données PPI.*** Données obtenues grâce au web-service PSICQUIC View [98].

En terme de caractéristiques des interactomes (nombre moyen/médian d'interaction par protéine, protéine la plus connectée, etc.), la Table 1.5 met également en évidence une importante hétérogénéité. Le nombre moyen d'interacteurs par protéine variant entre 6.47 et 27.79 par exemple pour l'interactome humain.

Espèce	**BioGRID**			**IntAct**			**DIP**			**MINT**		
	Moy.	Méd.	Max.	Moy.	Méd.	Max.	Moy.	Méd.	Max.	Moy.	Méd.	Max.
H. sapiens	27.79	9	2399	14.16	4	956	6.47	2	305	7.02	3	366
D. melanogaster	9.41	4	186	6.47	3	178	6.05	3	178	2.87	2	40
C. elegans	3.88	2	181	4.20	1	333	3.55	1	233	2.36	1	153
S. cerevisiae	40.76	17	2672	7.63	3	600	12.23	5	307	15.19	4	549
A. thaliana	8.69	3	1414	6.15	2.5	275	2.59	1	128	3.45	2	21

TABLE *1.5* – ***Caractéristiques des interactomes décrits dans la Table 1.4.*** Max. : degré maximal, Méd. : degré médian et Moy. : degré moyen.

1.2.2 Évolution des voies de signalisation

a) Caractéristiques générales

Les interactions protéine-protéine ont été étudiées d'un point de vue évolutif depuis le début des années 2000. Le terme d'*interologue*, introduit en 2000 par le laboratoire de Marc Vidal [55] désigne ainsi des protéines dont l'interaction est conservée entre deux espèces. Soit A et B deux protéines interagissant chez l'organisme O_1 et A' et B' leurs homologues respectifs dans l'espèce O_2. On parlera d'interologues si l'interaction A-B est conservée chez O_2, *i.e.* si A' interagit avec B'. Bien évidement, l'existence de cette interaction chez O_2 est soumise à l'existence de A' et de B'. Ainsi, la non conservation d'une interaction peut être due à la perte de l'un des gènes dans l'une des deux lignées. Les événements de duplications de gènes influencent également la conservation de l'interaction (Figure 1.9).

Figure omise pour des raisons de droits d'auteur.

FIGURE *1.9* – ***Modification des interactions suite à la duplication du gène codant pour la protéine A dans l'organisme O_1.*** Adapté de [100]. Les protéines X' de l'organisme O_2 correspondent aux homologues respectifs des protéines X de l'organisme O_1.

Le taux de conservation des interactions entre la levure et le nématode a été estimé à 16-30% en 2001 [101] puis à 30-54% en 2004 [102]. De même, les estimations de la proportion d'interactions conservées entre la levure et l'Humain varient énormément. Une première étude publiée en 2006 montre en effet qu'il n'y a quasiment aucune interaction conservée [103] tandis qu'une étude de 2007 estime que 16% sont conservées [104]. Deux études, publiées en 2008 et 2011, estiment quant à elles qu'il y a respectivement 90% et 71% des interactions conservées entre ces deux espèces [105, 106]. Ces variations s'expliquent en partie par la différence de quantité de données disponibles au moment de l'analyse mais également par la méthodologie employée. En effet, l'estimation des deux dernières études ne se fondent que sur les cas où les deux partenaires de l'interaction possèdent un homologue dans l'autre organisme. Les pertes d'interaction dues à la perte d'un des deux gènes ne sont donc pas prises en compte dans le calcul de ce pourcentage.

Plusieurs études se sont intéressées à la dynamique d'évolution des interactomes. Ainsi en 2001, Wagner a estimé qu'en moyenne 50-100 nouvelles interactions sont ajoutées à l'interactome de la levure chaque millions d'années et que le taux de perte d'interaction par paire de protéine par million d'années est d'environ 2.2×10^{-3} [107]. Plus récemment, Beltrao *et al.* [108] ont estimé qu'environ 10^{-5} interactions par paires de protéines par million d'années sont modifiées, représentant de 100 à 1000 interactions modifiées chaque million d'années dans les interactomes eucaryotes. De plus, ces auteurs ont détecté une forte corrélation positive entre le taux de reconnexion (modifications des interactions) et le degré de connectivité de la protéine ($R^2 > 0.92$ pour les quatre interactomes eucaryotes étudiés). Il a également été montré que la connectivité des protéines est inversement proportionnelle à leur taux de mutations [109, 102, 110]. Ainsi, les protéines très connectées, appelées *hubs*, sont en général plus conservées et évoluent plus lentement que les protéines possédant peu d'interacteurs. Une étude plus récente a observé la perte des protéines possédant moins de quatre interactions lors de la création du réseau d'interologues commun à l'Homme, le rat, la souris, la drosophile, le nématode et la levure [104], confirmant l'hypothèse avancée par Fraser *et al.* dès 2002 [109].

De manière globale, Shou *et al.* ont déterminé que le taux de modification des interactions entre deux organismes varie avec le temps de divergence mais également selon la fonction des protéines considérées [111]. Ainsi, les interactions entre protéines impliquées dans des voies métaboliques sont en moyenne plus conservées que celles des protéines impliquées dans la régulation des facteurs de transcription.

De même, Vo *et al.* [100], ont estimé qu'entre l'Homme et *Schizosaccharomyces pombe* environ 45% des interactions entre protéines impliquées dans des modifications de protéines cellulaires sont conservées tandis que cette proportion est supérieure à 80% pour les protéines impliquées dans l'assemblage des complexes macromoléculaires (Figure 1.10). De manière surprenante, cette étude révèle que la proportion d'interactions conservés entre *S. pombe* et l'Homme est très supérieure à celle entre *S. pombe* et *S. cerevisiae* (65% et 40% respectivement).

Concernant l'inférence d'interaction entre paires d'orthologues, Yu *et al.* [102] considèrent qu'au delà de 80% d'identité de séquences et d'une *E-value* jointe inférieure à 10^{-70} on peut inférer une conservation d'interaction. Dans les cas où les séquences des domaines d'interaction entre les deux protéines sont connues, une étude plus précise de la conservation de ces domaines peut donner des indications quant à la conservation ou non de cette interaction.

Figure omise pour des raisons de droits d'auteur.

FIGURE *1.10* – ***Proportions d'interactions conservées entre protéines impliquées dans différentes fonctions biologiques [100].*** S.p. : *S. pombe*, S.c. : *S. cerevisiae* et H.s. : *Homo sapiens*.

b) Phylogénie et évolution des voies de signalisation

La plupart des études phylogénétiques concernent un gène ou une famille génique. Très peu d'études ont visé à reconstruire l'histoire évolutive d'une voie de signalisation ou d'un réseau d'interactions protéine-protéine de manière détaillée. On peut néanmoins évoquer trois analyses dont deux ont été publiées en 2016.

Le système CytoR/JAK/STAT/SOCS

La première étude de 2016 s'est intéressée aux protéines CytoR (***Cyt****okine* ***R****eceptor*), JAK (***JA****nus* ***K****inase*), STAT (***S****ignal* ***T****ransducer and* ***A****ctivator of* ***T****ranscription*) et SOCS (***S****uppressor* ***of*** ***C****ytokine* ***S****ignaling*). Ce système biologique permet la réponse des cellules à la présence de cytokine dans le milieu extracellulaire. Sur la base de l'étude des quatre arbres phylogénétiques correspondants, Liongue *et al.* [112] ont conclu que la mise en place de ce système s'est effectuée de manière incrémentale depuis l'ancêtre commun des Unikonta qui ne possédait que la protéine STAT. Ils ont daté l'émergence du système complet au dernier ancêtre commun aux Placozoa, Cnidaria et Bilateria (Figure 1.11). Dans cette analyse, la conservation de leurs interactions n'a pas été étudiée.

Figure omise pour des raisons de droits d'auteur.

FIGURE *1.11* – ***Mise en place du système CytoR/JAK/STAT/SOCS chez les Unikonta [112].*** Chaque gène est représenté par une couleur.

Le système d'endocytose assistée par clathrine

La seconde étude [113] s'est quant à elle focalisée sur la mise en place du système d'endocytose assistée par clathrine (ou **CME** pour ***C****lathrin-****M****ediated* ***E****ndocytosis*) chez les Eucaryotes. Parmi les 35 protéines considérées, les auteurs ont déterminé que 22 étaient déjà présentes chez LECA tandis que seulement quatre datent de l'émergence des Metazoa (Figure 1.12).

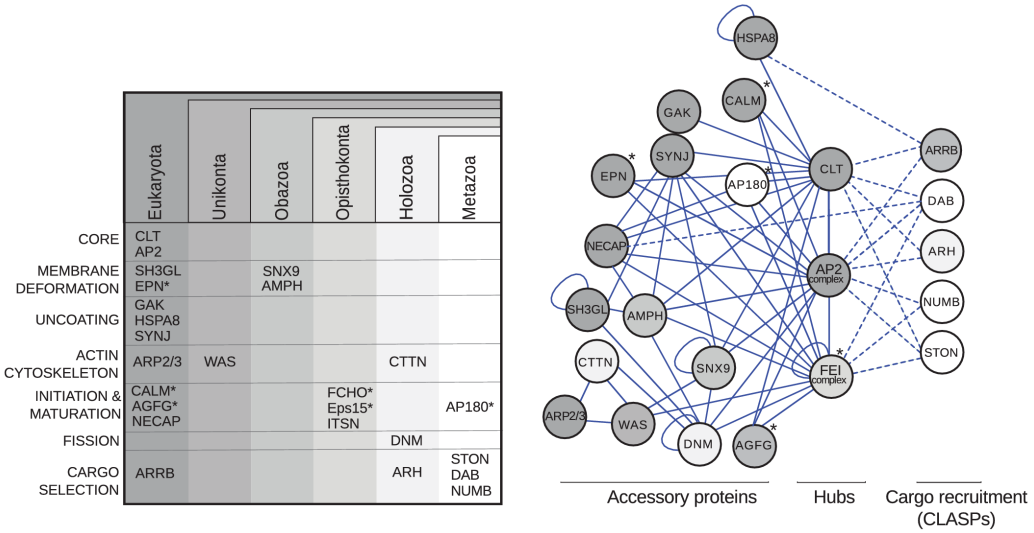

FIGURE *1.12* – ***Émergence des différentes protéines impliquées dans l'endocytose assistée par clathrine [113].***

Les auteurs n'ont par ailleurs pas détecté d'explosion de diversification mais une augmentation du nombre des paralogues chez les Vertebrata (dont 60% des protéines possèdent au moins un paralogue contre 12% pour les invertébrés). Contrairement à Liongue *et al.*, les auteurs de cette étude se sont intéressés à la conservation des interactions. Néanmoins ils n'ont étudié que l'interaction des sous-unités α et

μ de la protéine AP2 (***A**daptor **P**rotein **2** complex*) chez un mammifère (le rat), un champignon et un straménopile.

Le système des GPCR

Les GPCR (***G**-**P**rotein-**C**oupled **R**eceptor*) sont des protéines transmembranaires impliquées dans l'activation de nombreuses voies de signalisation chez les Eucaryotes. Cette famille de protéines, ainsi que les protéines G qu'elles régulent, ont constitué l'un des mécanismes de transduction du signal les plus étudiés [47, 114, 115]. En effet, chez les animaux ce système est au cœur des mécanismes transmettant des signaux intracellulaires en réponse à des signaux sensoriels tels que la lumière, les changements d'états physiologiques internes (neurotransmetteurs, hormones) ou encore la présence d'éléments lié à l'immunité (chimiokines) [116, 117]. Dans la continuité des précédentes études sur l'évolution de cette voie [118, 115, 119], De Mendoza *et al.* [120] ont inféré les arbres phylogénétiques d'une dizaine de protéines impliquées dans ce système (Figure 1.13).

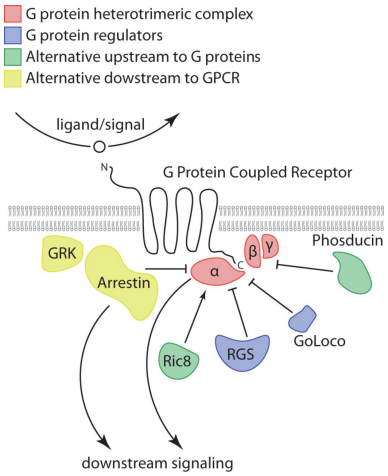

FIGURE *1.13* – ***Représentation schématique de la voie de signalisation des GPCR [120].***

Les auteurs ont étudié la présence, parmi 78 génomes eucaryotes, de protéines possédant l'un des sept domaines transmembranaires composant les GPCR et ont inféré les arbres de gènes correspondant aux autres protéines de la voie représentées dans la Figure 1.13. Bien qu'essentiellement détectées chez les Unikonta, la présence de certaines de ces protéines dans des génomes de Bikonta, permettent aux auteurs d'inférer la présence de la voie GPCR chez LECA. Ce scénario évolutif implique de nombreuses pertes indépendantes dans les différentes lignées eucaryotes, notamment chez les plantes (Figure 1.14). Les auteurs ont par ailleurs découvert que

certaines espèces possèdent les récepteurs mais pas les gènes codant les protéines G, et vice versa. Ils mettent ainsi en avant l'évolution indépendante des différents composants de la voie et sa modularité. Enfin, De Mendoza *et al.* ont également observé une explosion de la diversification des GPCR chez les métazoaires, qu'ils pensent liée au passage à la multicellularité.

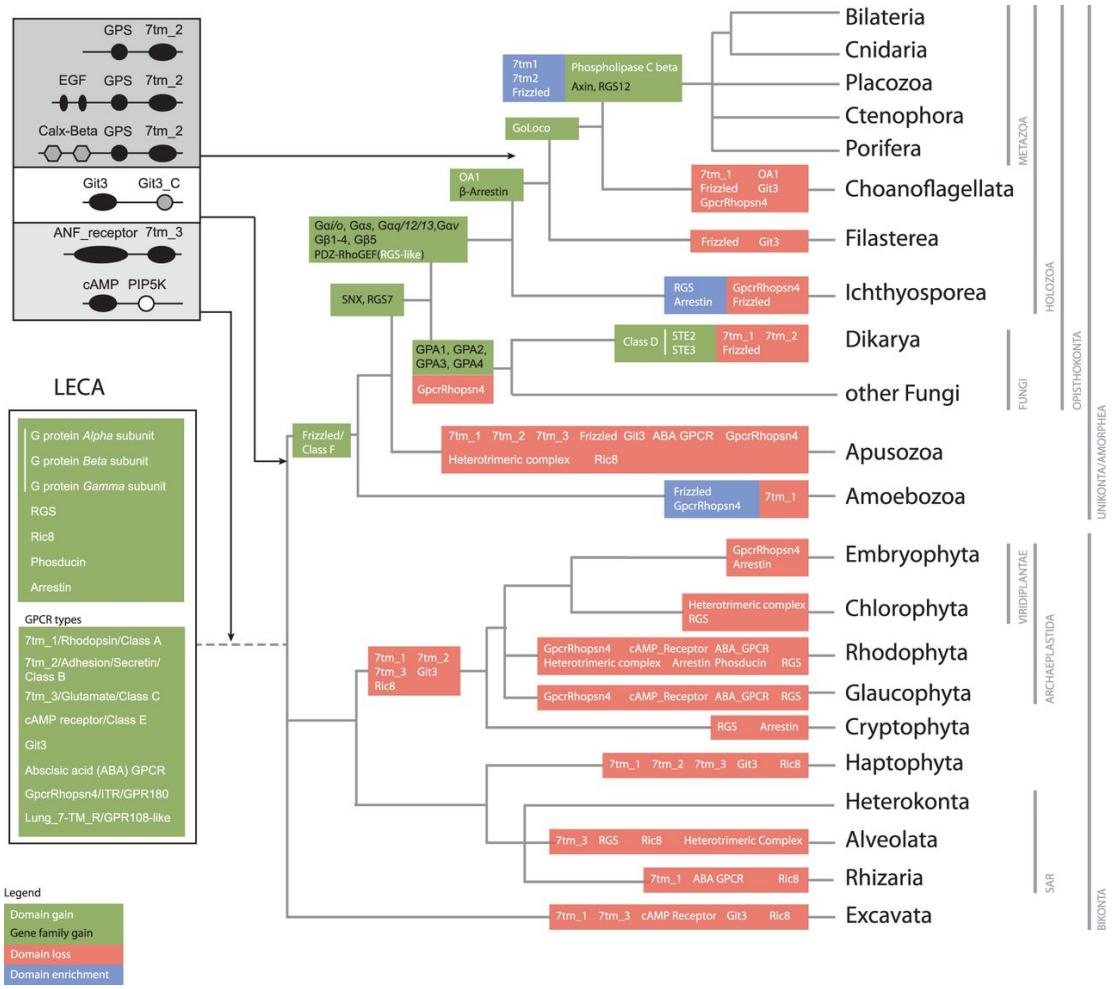

FIGURE *1.14* – ***Cladogramme résumant les événements évolutifs ayant affecté le système GPCR chez les Eucaryotes [120].***

Chapitre 2

Reconstruction phylogénétique

L'obtention des premières séquences nucléiques et protéiques dans les années 1950-1970 a permis l'essor d'une nouvelle discipline scientifique : la phylogénie moléculaire. Fondée sur le principe que les espèces portent leur histoire évolutive dans leur matériel génétique, la phylogénie moléculaire utilise les séquences d'ADN, d'ARN et de protéines. L'analyse par modélisation des principaux mécanismes évolutifs que sont la mutation, l'insertion et la délétion, a ainsi permis d'apporter des éléments de réponse à de nombreuses questions que la phylogénie classique ne parvenait pas à résoudre. Le développement de modèles de plus en plus complexes combinés à l'essor de l'informatique et à l'avalanche des données génomiques issues des méthodes de séquençage à haut débit ont contribué au succès et à l'essor de la phylogénie moléculaire.

Une analyse phylogénétique permet d'émettre des hypothèses quant aux événements évolutifs que le gène ou le génome d'étude a subi. On définit ainsi deux gènes comme étant *homologues* s'ils dérivent d'un même gène ancestral [121]. De plus, parmi les gènes homologues, on distingue (Figure 2.1) :

- Les gènes paralogues : ils résultent d'une duplication ancestrale au sein d'une même espèce.
- Les gènes orthologues : ils sont issus d'une spéciation.
- Les gènes xénologues : ils résultent d'un événement de transfert horizontal.

Cette dernière catégorie concernant d'avantage les génomes de procaryotes, elle ne sera pas considérée dans ce travail.

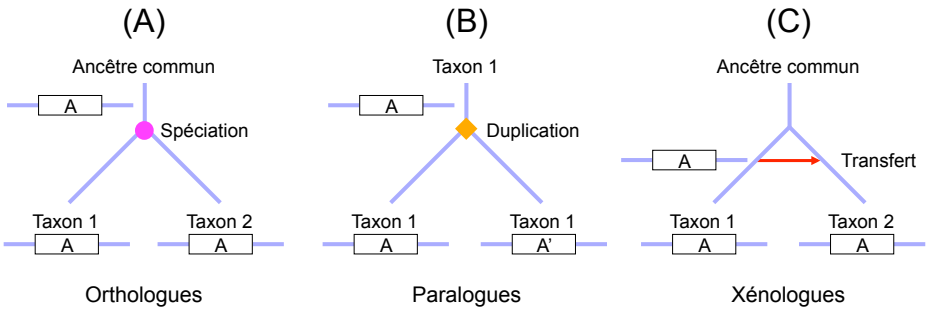

FIGURE *2.1* – ***Les différentes catégories de gènes homologues [121].***

Ce chapitre a pour objectif de présenter succinctement les trois grandes étapes d'une analyse phylogénétique classique : i) la recherche d'homologues, ii) l'alignement de séquences et iii) l'inférence de l'arbre phylogénétique proprement dite.

2.1 Constitution du jeu de données

2.1.1 Banques de données principales

Dès 1965, Dayhoff *et al.* publient l'*Atlas of Protein Sequence and Structure*, composé d'environ 70 protéines et considéré comme la première collection de séquences biologiques [122, 123]. Suite au développement des premières méthodes de séquençage rapide, la création d'espaces de stockage dédiés aux séquences biologiques est rapidement apparue comme une nécessité. De nos jours, il existe trois banques généralistes de séquences nucléiques : la DDBJ (***D****NA* ***D****ata* ***B****ank of* ***J****apan*) [124], l'ENA (***E****uropean* ***N****ucleotide* ***A****rchive*) [125] et GenBank [126]. En 2002, un consortium collaboratif a été créé de façon à normaliser les données contenues dans ces trois banques [127] . Concernant les protéines, la principale banque utilisée est UniProtKB [128] qui, en plus des séquences proprement dites, fournit d'autres informations telles que la fonction, la structure secondaire, les pathologies éventuellement associées à la protéine et les SNP (***S****ingle* ***N****ucleotide* ***P****olymorphism*) répertoriés.

De nos jours on dénombre plus de 1600 banques de données dédiées aux séquences biologiques [129], certaines n'étant plus maintenues et la plupart étant *spécifiques* (d'un organisme, d'un groupe taxonomique, d'une famille de gènes, etc). Afin de les inventorier, une édition annuelle spéciale du journal *Nucleic Acids Research* fut créée en 1993 [130]. L'édition 2015 de ce numéro répertorie ainsi 56

nouvelles banques publiées au cours de cette année ainsi que de nombreuses mises à jours de banques existantes [131]. De plus, avec l'arrivée des nouveaux séquenceurs, le nombre de nouvelles entrées dans ces banques a explosé dans les années 2000 (Figure 2.2). En dépit d'une vérification régulière, on trouve aujourd'hui de nombreuses redondances à l'intérieur des banques. Une étape de sélection des séquences et la création de nouveaux outils dédiés sont ainsi devenues une nécessité dans le cadre d'analyses phylogénétiques où cette redondance peut biaiser l'étape d'alignement et de sélection de sites conservés.

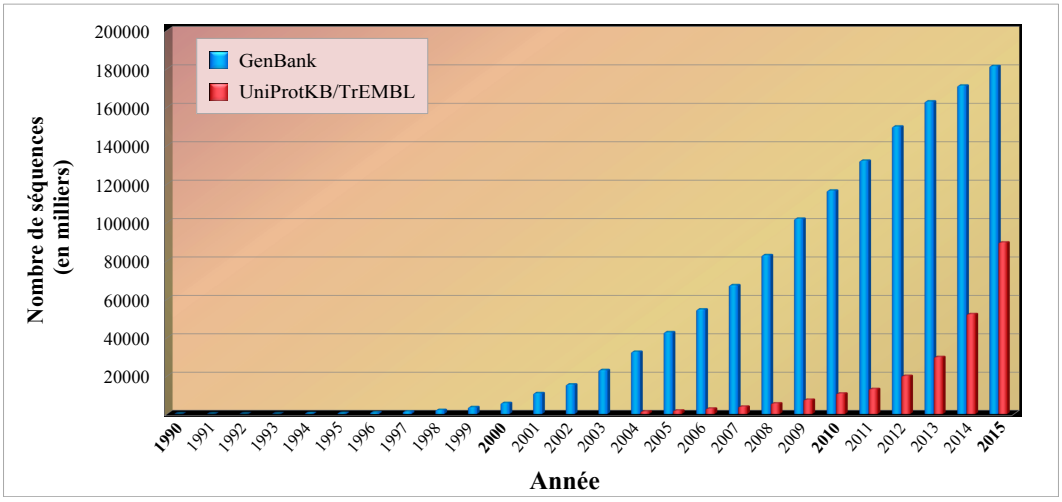

FIGURE *2.2* – ***Évolution du nombre de séquences contenues dans GenBank et UniProtKB/TrEMBL.*** D'après les données de [128, 132].

Malgré le nombre croissant de données génomiques générées, l'ensemble des génomes eucaryotes est loin d'être séquencé. On en dénombre seulement 1758 en 2014 [133] quand le nombre d'espèces de ce règne est estimé à environ 1.7 million [2]. De plus, il existe un biais de représentation taxonomique dans les banques de données. En effet, environ 85% des génomes eucaryotes disponibles en 2014 proviennent de Metazoa, Fungi et Embryophyta [133]. Et même au sein de ces trois groupes, il existe une hétérogénéité dans la couverture des espèces séquencées telles qu'une sous-représentation des invertébrés.

2.1.2 Recherche d'homologues

Dans la suite de ce manuscrit, un *résidu* désignera un nucléotide ou un acide aminé composant une séquence. Un *site* correspondra, quant à lui, à une la position (ou colonne) d'un alignement. De façon analogue aux gènes, deux résidus d'un

même site seront qualifiés d'homologues s'ils dérivent d'une position ancestrale commune.

a) Principe

Lorsque l'on s'intéresse à l'histoire évolutive d'un gène, la première étape consiste à identifier l'ensemble des copies de ce gène présentes dans les organismes vivants. Soit S la séquence étudiée (ou séquence requête) et B l'ensemble des séquences disponibles d'une banque ; la recherche d'homologues consiste à déterminer les séquences de B étant similaires à S car partageant un ancêtre commun avec cette dernière. Pour cela, une phase d'alignement et de détermination d'un score de similarité sont nécessaires. Le logiciel BLAST (***B****asic* ***L****ocal* ***A****lignment* ***S****earch* ***T****ool*) est depuis plus de vingt ans le principal outil utilisé pour la recherche d'homologues.

b) BLAST

Publié pour la première fois en 1990 par Altschul *et al.* [134], l'heuristique BLAST est composée de trois grandes étapes : i) le découpage de S en un ensemble de mots de longueur fixe la constituant ; ii) la recherche exacte parmi les séquences B de certains de ces mots (fragments aux propriétés biochimiques, physiques ou cinétiques particulières), et iii) l'élongation des mots trouvés dans les séquences B (Figure 2.3).

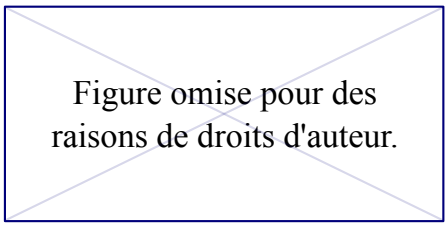

FIGURE *2.3* – ***Les trois étapes d'une recherche de séquences homologues avec l'algorithme BLAST de 1990 (adapté de [135]).***

Depuis sa création, l'algorithme a été partiellement modifié afin d'améliorer sa rapidité mais de conserver son efficacité et sa sensibilité. L'avancée majeure ayant été l'introduction des *gaps* dans les alignements, c'est-à-dire des trous correspondant à l'absence d'homologie à un site de l'alignement [136]. Représentés généralement par un tiret (–), la gestion des gaps est un problème récurrent en phylogénie moléculaire.

Par ailleurs, étant donné la taille actuelle des banques de données, une partie des séquences trouvées par BLAST ne seront, statistiquement, que le fruit du hasard. Pour traduire la significativité des résultats obtenus, Karlin et Altschul ont introduit l'utilisation de la *E(xpected)-value* [137]. Ce score, associé à chaque segment d'une séquence de B identifiée comme similaire à S, décrit le nombre de résultats attendus sous le modèle nul, *i.e.* dû au hasard. Plus le score de similarité entre S et la séquence de la banque sera élevé, plus la *E-value* associée sera faible [138]. Ainsi, une faible *E-value* traduira une forte probabilité d'origine commune entre ces deux séquences. Néanmoins, c'est à l'utilisateur de fixer le seuil en dessous duquel il considère la *E-value* comme significative. Pour les homologues « proches », il est possible de fixer ce seuil aux alentours de 10^{-30}, valeur que j'ai utilisée lors de l'étude des PI3K (voir chapitre 4).

Initialement dédié à la recherche de séquences de même nature (nucléiques ou protéiques), cinq algorithmes issus de la méthode de 1990 ont été développés. Ainsi, la recherche d'homologues d'une séquence protéique parmi une base de données nucléiques sera par exemple réalisée à l'aide du programme tBLASTn (Figure 2.4). Dans le cadre de mon travail, j'ai exclusivement utilisé le programme BLASTp permettant d'effectuer la recherche d'homologues protéiques à partir d'une séquence protéique.

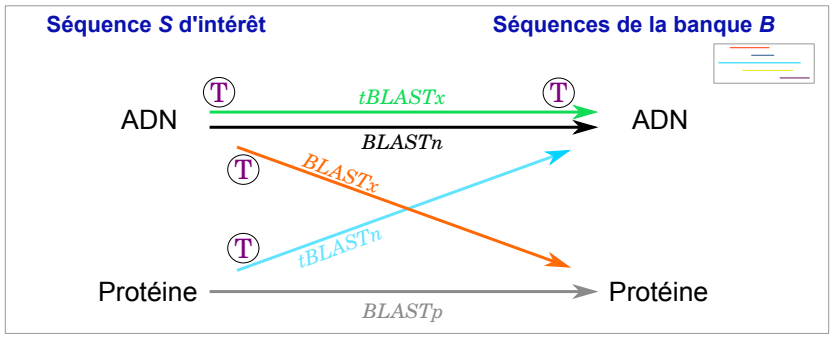

FIGURE 2.4 – ***Les cinq algorithmes BLAST selon la nature de la séquence S et de la banque B (L. Duret, comm. pers.).*** Le T violet entouré symbolise la traduction des séquences d'ADN dans les six phases de lectures.

c) Homologues lointains

Si les différents programmes BLAST présentés ci-dessus sont très performants, ils présentent des limites pour la détection d'homologues « lointains » ou ayant évolués rapidement. Pour pallier ce problème, Altschul *et al.* ont développé l'algorithme ***P**osition-**S**pecific **I**terated* (PSI)-BLAST, en 1997 [136]. La première itération d'une session PSI-BLAST est identique à une recherche BLASTp classique. A partir de l'alignement des résultats significatifs obtenus (*i.e.* en dessous d'une *E-value* fixée par l'utilisateur), une matrice de score position-spécifique est générée. Un score élevé sera attribué aux sites conservés de l'alignement, tandis que les positions plus variables seront associées à un score plus faible [139]. Cette matrice est ensuite utilisée comme requête pour une seconde recherche BLASTp, et ainsi de suite. Les résultats des différentes itérations dépendent donc fortement des séquences sélectionnées pour générer la matrice de score. La souplesse sur les sites variables apportée par cette approche permet de détecter des homologues plus lointains que l'approche classique. Le nombre d'itérations étant un paramètre fixé par l'utilisateur.

Une seconde solution pour la détection d'homologues lointains est la sélection et l'utilisation de certains homologues détectés par un premier BLAST comme requêtes pour effectuer une nouvelle recherche (séquences dites *seeds* ou « graines »). Soit une séquence S protéique humaine dont on recherche l'ensemble des homologues chez les Eucaryotes. Si, parmi les résultats significatifs, on retrouve des séquences de tous les Eucaryotes mais une seule séquence de plante par exemple, il est judicieux d'utiliser cette dernière comme graine pour une nouvelle recherche. En effet, cela permet de déterminer s'il existe d'autres séquences de plantes homologues à la séquence S humaine d'intérêt, qu'une seule itération de BLAST n'aurait pas détectées car trop dissemblables à S. Néanmoins, cette approche nécessite une connaissance des groupes taxonomiques considérés ainsi que de la composition des banques de données afin d'effectuer un choix perspicace de séquences graines. Il s'agit de la méthode employée dans l'étude de l'histoire évolutive des PI3K (voir chapitre 4).

2.1.3 Qualité du jeu de données

BLAST étant fondé sur la recherche de similarités locales, il est récurrent d'obtenir des séquences partielles parmi les homologues identifiés. Particulièrement courtes par rapport au reste du jeu de données, celles-ci sont à l'origine de biais

lors des phases d'alignement et de sélection de sites. Dans le cas d'études phylogénétiques concernant les organismes eucaryotes, il est également fortement probable que les jeux de données obtenus comprennent des séquences issues d'un même gène, *i.e.* des transcrits alternatifs. Malgré la nécessité évidente d'éliminer ces séquences, un seul logiciel dédié spécifiquement à cette problématique est disponible de nos jours (voir chapitre 3).

2.2 Alignement de séquences

2.2.1 Principe général

Au fil des générations une séquence nucléotidique subit diverses modifications appelées mutations. Trois événements évolutifs majeurs sont à l'origine des variabilités observées : la substitution, l'insertion et la délétion. Une mutation par substitution correspond à la modification ponctuelle d'un résidu de la molécule d'ADN. Une insertion, (respectivement une délétion), correspond à l'ajout, (respectivement la suppression), d'un ou plusieurs résidus. Ainsi, les séquences homologues étudiées peuvent présenter des différences de contenu mais également de longueurs. L'alignement de celles-ci vise à organiser les séquences de façon à ce que les sites dérivant d'un même résidu dans le gène ancestral apparaissent dans une même colonne. Cette étape est donc indispensable afin d'identifier et de faire correspondre les sites homologues entre eux.

a) Alignement de deux séquences

Soit deux séquences $A = [a_1 a_2 ... a_n]$ et $B = [b_1 b_2 ... b_m]$. Pour chaque résidu a_i $(1 \leq i \leq n)$ de A, il existe deux possibilités : soit ce résidu est homologue à un résidu b_j $(1 \leq j \leq m)$ de B, soit il ne l'est pas. Les *gaps* sont donc des brèches introduits dans A ou dans B lors de l'alignement afin de traduire une absence d'homologie due à une insertion ou à une délétion.

Le nombre d'alignements possibles entre deux séquences croissant de manière exponentielle avec le nombre de résidus des séquences [140], la principale difficulté est de déterminer quel est l'alignement optimal. Pour cela, un système de score a été mis en place. Tout d'abord, les insertions et les délétions étant des événe-

ments évolutifs moins fréquents que les substitutions [141], une pénalisation plus importante a été attribuée à la création ainsi qu'à l'expansion des gaps dans les alignements [142, 143]. Par ailleurs, initialement traitées de manière égale, il est apparu que les substitutions de résidus ne sont pas toutes équiprobables. En effet, bien qu'à partir d'un nucléotide il existe deux fois plus de *transversions* ($\mathrm{C, T} \longleftrightarrow \mathrm{A, G}$) possibles que de *transitions* ($\mathrm{C} \longleftrightarrow \mathrm{T}$ et $\mathrm{A} \longleftrightarrow \mathrm{G}$), ces dernières sont plus fréquemment observées [144, 145, 146, 121]. Ainsi, des matrices de substitutions ont été développées permettant d'attribuer des scores différents selon la nature des substitutions. Dans le cas des acides aminés, ces matrices étaient initialement fondées sur leurs propriétés physico-chimiques mais très vite, ce sont sur des critères phylogénétiques qu'elles ont été calculées. Ainsi, les matrices PAM (***P**oint **A**ccepted **M**utation*) [147] et BLOSUM (***BLO**cks **SU**bstitution **M**atrix*) [148] sont les matrices de substitutions protéiques les plus utilisées de nos jours. Parmi les différentes matrices BLOSUM, la matrice BLOSUM62 est la plus fréquemment utilisée par défaut dans les logiciels [136, 149, 150, 151, 152]. En effet, sur leur jeu de données test de 1992, Henikoff *et al.* ont déterminé qu'il s'agissait de la matrice produisant en moyenne les meilleurs résultats. Néanmoins, le choix de la matrice de substitution dépend des caractéristiques du jeu de données étudié et reste un point délicat dans le processus d'alignement.

Enfin, deux stratégies sont possibles lors de l'alignement de deux séquences : essayer d'aligner les séquences sur toute leur longueur ou bien rechercher les similarités locales (Figure 2.5).

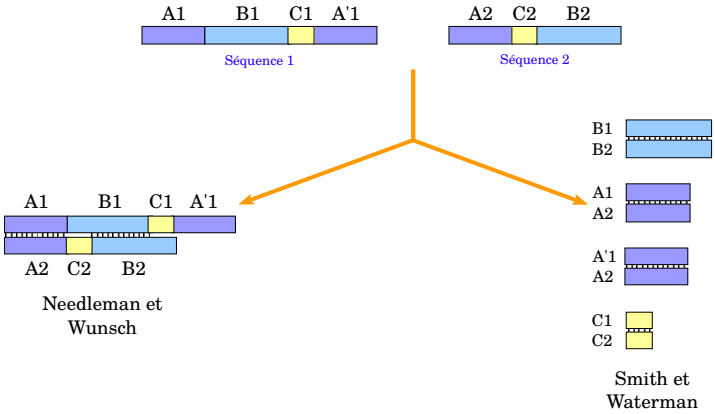

FIGURE *2.5* – ***Alignement global et alignement local [153].*** A_i, B_i et C_i représentent des fragments homologues.

La première méthode, dite d'*alignement global*, fut développée en 1970 par Needleman et Wunsch [154] et appliquée à l'alignement de deux séquences protéiques.

La seconde stratégie, dite d'*alignement local*, fut initialement publiée par Smith et Waterman en 1981 [155] et consiste à trouver les paires de segments de similarité maximale. L'alignement global conduira à un alignement de qualité uniquement si A et B sont de longueurs similaires et n'ont pas subit d'inversions ni d'insertions/délétions majeures.

b) Alignement multiple

Les alignements dits *multiples* (*i.e.* de plus de deux séquences), ou MSA (***M**ultiple **S**equences **A**lignment*), seront pratiquement toujours effectués à l'aide d'*heuristiques* produisant une approximation de l'alignement optimal. En effet, bien que théoriquement généralisable à l'alignement de plus de deux séquences, la complexité des deux algorithmes précédents implique une augmentation exponentielle du temps de calculs en fonction du nombre de séquences [121]. Ainsi, la méthodologie la plus couramment utilisée de nos jours est celle d'un alignement dit *progressif*.

Introduit en 1984 par Hogeweg et Hesper [156], la première étape consiste à inférer une matrice de distances à partir des alignements deux à deux des n séquences du jeu de données. Les séquences sont ensuite divisées en groupes selon leur similarité à l'aide d'une méthode de *clustering* [157, 158] ou d'un arbre guide [141, 159, 160]. A partir de l'alignement des deux séquences les plus similaires, l'algorithme consiste alors à incorporer progressivement dans l'alignement les $n-2$ séquences restantes (Figure 2.6).

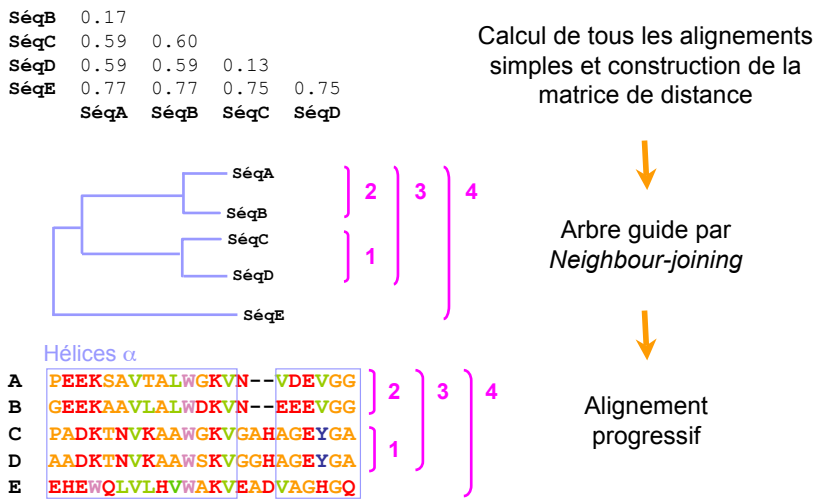

FIGURE 2.6 – ***Procédure d'alignement progressif utilisant un arbre guide calculé à partir de la matrice des scores d'alignements des paires de séquences [121].*** Méthodologie utilisée par le programme CLUSTALW [143].

2.2.2 Principaux logiciels d'alignements

CLUSTAL

Publié en 1988 par Higgins et Sharp [160], CLUSTAL est l'un des premiers logiciels à utiliser un arbre guide pour l'alignement conjoint de plusieurs séquences. La première étape d'alignement deux à deux de séquences est réalisée à l'aide de l'heuristique de Wilbur et Lipman [161], permettant un gain de calcul par rapport à l'algorithme exact de Needleman et Wunsch. La matrice des scores est ensuite définie par le nombre de résidus identiques entre les deux séquences moins la pénalité due aux gaps introduits. Dans la version originale du programme, l'incorporation progressif des séquences dans l'alignement multiple est effectué selon l'ordre du dendogramme obtenu par la méthode UPGMA (***U****nweighted* ***P****air* ***G****roup* ***M****ethod with* ***A****rithmetic mean*) [162]. Dès la version de CLUSTAL sortie en 1994, l'inférence de l'arbre guide est effectuée grâce à la méthode NJ (***N****eighbour-****J****oining*), permettant ainsi un gain de robustesse face à l'hétérogénéité des taux d'évolution sans perte de vitesse de calcul [143]. Dans le cas de l'ajout à l'alignement en cours d'un groupe de séquences et non d'une séquence unique, CLUSTAL construit préalablement une séquence consensus représentant ce sous-alignement, comme par exemple lors de l'ajout à l'alignement du groupe n° 2 dans la Figure 2.6.

La version actuelle de CLUSTAL, nommée CLUSTALO (pour CLUSTAL Ω), a été spécialement développée pour permettre l'alignement rapide de volumineux jeux de données. Ainsi, plusieurs milliers de séquences peuvent en théorie être alignées en quelques heures [163]. Parallélisé, l'algorithme de CLUSTALO utilise une approximation de la matrice des distances (nommée mBed [164]) afin d'éviter la phase chronophage d'alignement de chaque paire de séquences possibles. A partir de cette matrice, des sous-groupes sont formés à l'aide de l'algorithme K-means [165, 166] et des matrices de distances complètes sont calculées pour chacun de ces derniers. Enfin, les sous-arbres associés sont inférés par la méthode UPGMA [162]. L'alignement est ensuite réalisé de manière itérative en utilisant des profils de chaînes de Markov cachées (ou HMM pour ***H****idden* ***M****arkov* ***M****odel*).

T-COFFEE

Développé en 2000 par Notredame *et al.*, T-COFFEE (***T****ree-based* ***Co****nsistency* ***O****bjective* ***F****unction* ***F****or alignment* ***E****valuation*) est un logiciel permettant de combiner alignement local et alignement global [167]. En effet, lors de la première phase et contrairement aux autres logiciels disponibles à cette époque, T-COFFEE créé

une bibliothèque composée d'alignements globaux (obtenus avec l'algorithme de CLUSTALW) et d'alignements locaux (obtenus avec LALIGN [168]) de paires de séquences. A partir de ces derniers, le logiciel évalue les distances entre séquences en calculant un score position-spécifique, s'affranchissant ainsi de l'utilisation d'une matrice de substitutions. Malgré un gain certain sur la qualité des alignements, notamment pour les jeux de données avec moins de 30% de similarité [167], T-COFFEE reste un programme beaucoup plus lent que les autres [163]. Néanmoins, comme son nom l'indique, le point fort de T-COFFEE est l'introduction d'une contrainte de consistante qui n'existe pas dans les autres logiciels disponibles.

MAFFT

Le logiciel MAFFT (***M****ultiple* ***A****lignment using* ***F****ast* ***F****ourier* ***T****ransform*) est fondé sur l'hypothèse que les substitutions entre acides aminés aux propriétés physico-chimiques proches sont plus fréquentes [169]. Ainsi, la première version de MAFFT, publiée 2002, n'utilise pas de matrice de substitution mais repose sur le calcul de corrélations entre les séquences fondées sur la volumétrie et la polarité des acides aminés la composant [170]. Afin d'améliorer le temps de calcul, Katoh *et al.* se sont servi des transformées de Fourier des deux grandeurs précédentes ainsi que d'une matrice de similarité normalisée. L'arbre guide pour l'alignement progressif est inféré à l'aide de la matrice de substitutions JTT (**J**ones, **T**aylor and **T**horton) [171] et de l'algorithme UPGMA. En 2005, de nouveaux algorithmes basés sur une approximation des transformées de Fourier ou l'utilisation d'un alignement local ont été intégrés à la suite de programmes de MAFFT [172]. Par ailleurs, sa parallélisation en 2010 a permis un important gain en terme de temps de calcul [173]. Dans sa version actuelle, MAFFT propose une dizaine d'algorithmes différents ainsi que diverses options à choisir selon les caractéristiques du jeu de données à analyser [152].

MUSCLE

Publié par Edgar en 2004 sous forme de deux articles [174, 175], le logiciel MUSCLE (***MU****ltiple* ***S****equence* ***C****omparison by* ***L****og-****E****xpectation*) utilise un algorithme itératif divisé en trois grande étapes. La première est celle d'un alignement progressif classique : une matrice de distances fondée sur la fréquence de k-mers (*i.e.* des segments de longueur k) dans les séquences est calculée, un arbre guide est inféré par la méthode UPGMA et un alignement progressif est produit en suivant l'ordre de branchement de cet arbre. La deuxième étape consiste à améliorer l'alignement

obtenu en utilisant une nouvelle matrice fondée sur la distance de Kimura [176] au lieu des précédents k-mers. Enfin, lors de la troisième étape, MUSCLE scinde aléatoirement en deux l'arbre guide et réaligne les deux sous-alignements correspondants à l'aide de profils. Si l'alignement des deux blocs produit un meilleur score global, il est gardé, sinon il est rejeté (Figure 2.7).

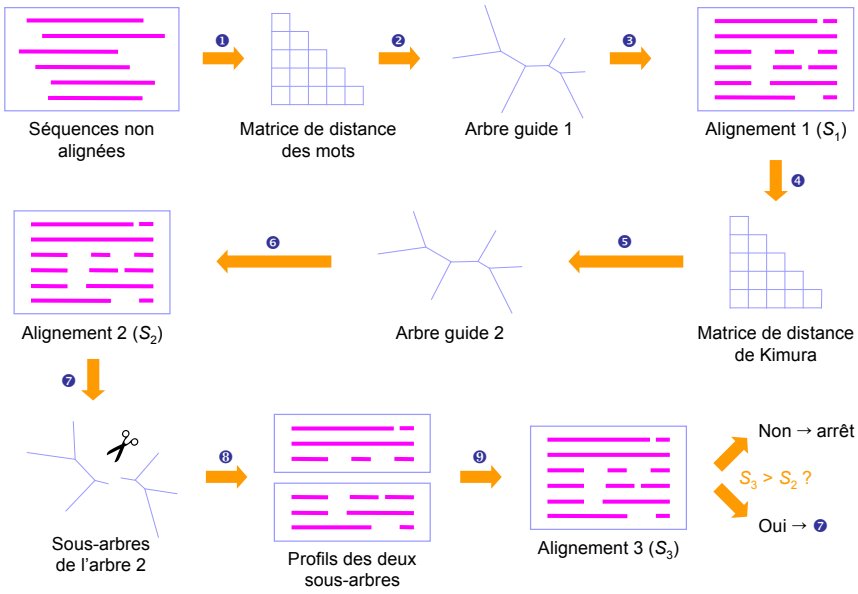

FIGURE 2.7 – ***Algorithme implémenté dans MUSCLE [175].***

PRANK

L'algorithme implémenté dans le logiciel PRANK (***PR**obabilistic **A**lig**N**ment **K**it*) a été développé afin de proposer des alignements non biaisés vers l'inférence systématique de délétions dans la séquence ancestrale [177]. En effet, lors de l'ajout d'une nouvelle séquence à l'alignement multiple en cours, les algorithmes progressifs précédents ont tendance à forcer l'insertion des gaps pré-existants dans cette séquence. Ce biais est à l'origine de l'inférence systématique de séquences ancestrales de grande taille. Afin d'équilibrer les probabilités d'occurrence d'une insertion et d'une délétion au long de l'évolution des génomes, le logiciel PRANK utilise un profil HMM dans lequel la probabilité d'ouverture d'un gap est la même pour les deux groupes de séquences à aligner. Par ailleurs, cette probabilité est exponentiellement proportionnelle à la distance évolutive qui sépare ces deux groupes de séquences. Ainsi PRANK est particulièrement utilisé pour l'inférence de séquences et génomes ancestraux [178, 179, 180].

FSA

Le programme probabiliste d'alignement FSA (***F**ast **S**tatistical **A**lignment*), développé en 2009 par Bradley *et al.* [181], permet quant à lui de ne pas utiliser d'arbre guide pour l'alignement multiple. En effet, ce logiciel repose sur le calcul des probabilités que deux résidus issus de deux séquences soient homologues entre eux. Celles-ci sont déterminées grâce à l'utilisation de profils HMM à trois ou cinq états : homologie, insertion ou délétion dans chacune des deux séquences considérées. A partir de ces probabilités, l'alignement multiple est ensuite effectué par approche itérative dite de *sequence annealing*. A noter que FSA propose un résultat graphique permettant à l'utilisateur de visualiser la fiabilité des positions alignées et ainsi vérifier manuellement l'alignement multiple obtenu.

MACSE

MACSE (***M**ultiple **A**lignment of **C**oding **SE**quences*) est un logiciel d'alignement développé en 2011 par Ranwez *et al.* [182] dans le but d'aligner des séquences nucléotidiques codantes en prenant en compte les changements de cadre de lecture lors d'insertion de gaps. Ce programme est divisé en trois grandes étapes : i) la traduction des séquences nucléotidiques en séquences protéiques, ii) l'alignement de ces séquences protéiques et iii) l'alignement des séquences nucléotidiques guidé par l'alignement protéique obtenu lors de la deuxième étape. Cette approche permet en particulier d'améliorer la qualité des alignements nucléotidiques lorsque des pseudogènes sont présents dans le jeu de données. MACSE peut également être utilisé sur des données de séquençage haut débit afin de détecter et de corriger les potentielles erreurs réalisées par les séquenceurs.

Comparaison

Chacun de ces logiciels ont été développés dans le but de résoudre un problème en particulier. Ainsi CLUSTALO a permis une réduction drastique du temps de calcul pour les gros jeux de données, PRANK a été construit de manière à prendre en compte les relations évolutives entre les séquences et T-COFFEE a introduit des contraintes de consistance dans le calcul des alignements. Ainsi, tout le problème de la comparaison de ces logiciels est de déterminer en quoi un alignement sera meilleur qu'un autre. En effet, le « vrai » alignement n'étant pas disponible, on ne peut pas déterminer quel est le meilleur logiciel tant au niveau du résultat qu'au niveau du temps de calcul. Dans le but d'avoir des alignements de référence, Thompson *et al.* [183] ont publié une banque de données nommée BAliBASE

(***B**enchmark **Ali**gnment data**BASE***) qui répertorie des alignements multiples corrigés manuellement et calibrés sur la structure 3D des protéines. La plupart des études comparatives ont été effectuées à l'aide de cette banque.

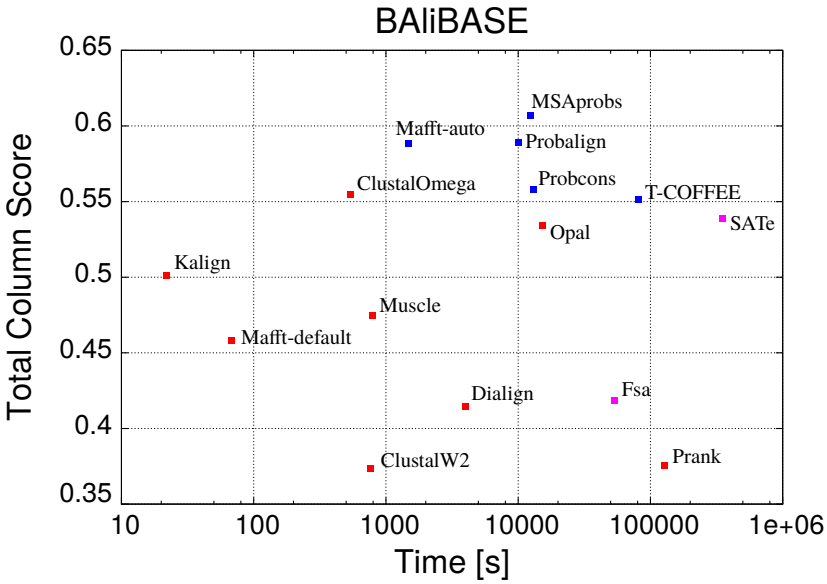

FIGURE *2.8* – ***Performances des principaux logiciels d'alignements en 2011 [163].*** Plus le score par colonne est élevé, meilleur est l'alignement.

Dans l'article consacré à CLUSTALO [163] MAFFT (avec le choix automatique de la méthode), ProbAlign [184], MSAprobs [185] et ProbCons [186] produisent de meilleurs alignements mais nécessitent des temps de calculs beaucoup plus longs que CLUSTALO(Figure 2.8). Dans cette étude, PRANK [187] et FSA [181] présentent les plus mauvais rapport résultat/temps de calcul. Dans un article de 2011, Thompson *et al.* [188] montrent que pour les jeux de séquences très divergents, T-COFFEE, ProbCons et MAFFT (avec l'option `L-INS-i`) produisent les meilleurs résultats mais que les deux premiers nécessitent 2.7 jours de calculs contre 1h12 pour le troisième. De manière générale, les auteurs observent une baisse des capacités des logiciels en fonction du degré de divergence entre les séquences. Plus récemment, une étude de 2014 recommande l'utilisation de ProbCons, T-COFFEE, ProbAlign et MAFFT tout en soulignant la rapidité de CLUSTALW et MUSCLE [189]. Enfin, la même année, Pervez *et al.* [190] désignent SATe [191] comme le logiciel possédant le meilleur rapport précision/temps de calcul parmi dix logiciels testés sur la version 3.0 de BAliBASE [192].

Sur la base des résultats obtenus dans ces études de comparaison mais également à la suite de tests sur mes jeux de données, j'ai principalement utilisé MAFFT dans le cadre de ce travail.

2.2.3 Sélection de sites

Malgré les améliorations apportées par ces différents logiciels, la fiabilité dans l'homologie inférée varie énormément d'une position à l'autre d'un alignement multiple. En effet, chaque site présente un degré de conservation ainsi qu'un nombre de gaps variables. Puisque des erreurs d'alignement peuvent entraîner des erreurs dans l'inférence des arbres, il est nécessaire d'éliminer les sites pour lesquels l'homologie est ambiguë ou le nombre de substitutions a atteint saturation. Si les premières études effectuaient cette sélection manuellement, de nombreux logiciels ont depuis été développés afin d'automatiser cette tâche.

La première approche a consisté à éliminer les sites évoluant trop rapidement afin d'éviter l'obtention de longues branches dans l'arbre inféré [193]. En 2000, Castresana présente ensuite Gblocks [194] qui est fondé sur un système de cinq seuils permettant de classer les sites en non conservés, moyennement conservés et très conservés. En effet, à chaque position ce logiciel compte le nombre de gaps présents et si, par exemple, il s'y trouve plus de 50% de gaps (valeur par défaut), le site est classé non conservé et sera en conséquence éliminé par le logiciel. Cette sélection de blocs permet d'éviter la conservation de positions isolées, situées aux alentours ou dans les *indels* (régions d'insertions/délétions) dont l'homologie est plus incertaine.

Fondé sur le même principe, le logiciel TrimAL (***Trim**ming **AL**ignment*) permet également une sélection des sites grâce au calcul d'un score mais celui-ci est pondéré par une matrice de substitution (BLOSUM62 par défaut) [150]. Si plusieurs alignements du même jeu de données lui sont fournis, TrimAL peut également effectuer la sélection en estimant la cohérence de ces alignements (score de consistance). L'idée selon laquelle une position d'un alignement est fiable si elle est conservée parmi plusieurs alignements est également la base du logiciel GUIDANCE (***GUID**e tree based **A**lig**N**ment **C**onfidenc**E***) [195]. Celui-ci génère par *bootstrap* des alignements perturbés du jeu de données et regarde la conservation des paires de résidus entre ces alignements perturbés et l'alignement de référence. Produisant de bons résultats, cette méthodologie nécessite toutefois des temps de calcul importants.

Une troisième méthode, basée sur le calcul de l'entropie, a été publiée par Criscuolo *et al.* en 2010 [151]. BMGE (***B**lock **M**apping and **G**athering with **E**ntropy*) attribue en effet un score à chacun des sites de l'alignement multiple grâce à une adaptation du calcul d'entropie de Von Neumann [196] qui est ensuite pondéré

par une matrice de substitution (BLOSUM62 par défaut). Les sites présentant une valeur d'entropie faible correspondent à des positions homogènes, donc sont les sites à conserver. Dans cet article, les auteurs comparent l'impact de la sélection de site sur l'inférence d'arbres phylogénétiques. A l'aide de données simulées et des données l'article de Gblocks [194], ils montrent que l'utilisation de logiciels de sélection de sites permettent l'obtention d'arbres plus précis et l'augmentation du support des branches « vraies » [151]. Cette justification d'élimination des sites ambigus est particulièrement valable pour les arbres inférés grâce à BioNJ. Enfin, les auteurs démontrent que BMGE fournit de meilleurs résultats que Gblocks [194], TrimAL [150] et Noisy [197] pour les séquences très divergentes mais que Gblocks est meilleur pour les séquences proches.

2.3 Inférence d'arbre de gènes

2.3.1 Définitions

A l'image des arbres généalogiques, un arbre phylogénétique est la représentation des liens de parentés existants entre des espèces (arbre d'espèces, Figure 2.9) ou des gènes homologues (arbre de gènes). Mathématiquement parlant, il s'agit d'un *graphe*, *i.e.* d'un ensemble de *nœuds* dont certains sont reliés par des *arêtes* (ou *branches*). A l'échelle d'un gène, les *nœuds externes* ou *feuilles* de l'arbre correspondent aux séquences du jeu de données, par opposition aux *nœuds internes* qui symbolisent les gènes ancestraux hypothétiques. Les arbres phylogénétiques racinés sont des graphes dits *binaires* et *orientés*.

En effet, dans le cadre d'un arbre de gène par exemple, un gène ne peut conduire théoriquement qu'à l'apparition de deux nouveaux gènes, que ce soit suite à un événement de spéciation ou de duplication. Ainsi, les nœuds d'un arbre phylogénétique sont uniquement de degré 1 (les feuilles) ou 3 (les nœuds internes). Néanmoins, dans la pratique, il est parfois difficile de déterminer avec précision l'ordre de succession de certains événements de spéciation. Dans ces cas, les arbres présenteront une multifurcation, c'est-à-dire qu'un nœud pourra donner naissance à plus de deux branches [121]. De la même façon, l'arbre de la Vie (c'est-à-dire l'arbre regroupant l'ensemble des espèces vivantes pour lesquelles des séquences sont disponibles) présente encore de nombreuses multifurcations car les relations de parentés entre

certain groupes d'espèces ne sont pas encore résolues avec certitude.

Par ailleurs, un arbre phylogénétique devient orienté grâce au placement d'une *racine* symbolisant le dernier ancêtre commun du jeu de données étudié. Dans cette configuration, la *racine* est de degré 2 et un ordre temporel entre les nœuds est établi : le temps s'écoule de la racine vers les feuilles. Par exemple, sur l'arbre raciné (B) de la Figure 2.9, le dernier ancêtre commun de l'Homme et du chat (nœud ❶) est plus ancien que celui du chat et du panda (nœud ❷). On peut noter que dans cette représentation, appelée *cladogramme*, les branches n'ont pas de longueurs, *i.e.* elles ne sont pas représentatives des distances évolutives séparant les différentes espèces. Par opposition, les *phylogrammes* sont des arbres phylogénétiques dans lesquels les branches indiquent le temps de divergence séparant deux nœuds. Ainsi, plus une branche est longue, plus la séquence a divergé de sa forme ancestrale.

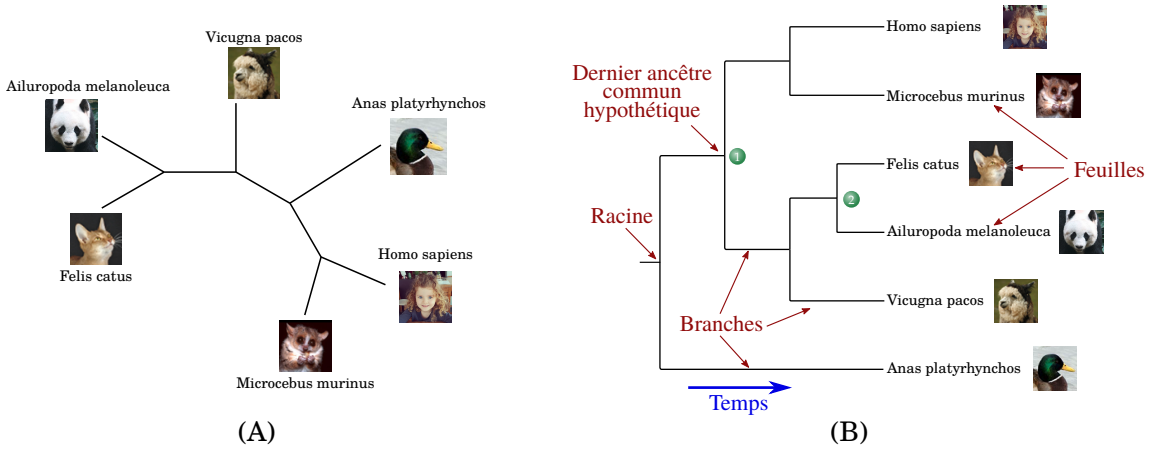

FIGURE *2.9* – ***Exemples d'arbres des espèces (A) non raciné et (B) raciné.*** D'après [198].

Il existe deux grandes méthodes d'enracinement d'arbre : l'enracinement aux *poids moyen* et l'utilisation d'un *groupe externe*. Ne nécessitant aucune information additionnelle, la première méthode consiste à placer la racine au milieu du chemin séparant les deux feuilles (ou groupes de feuilles) les plus éloignées. Néanmoins, ce raisonnement n'est valide que si les vitesses d'évolution sont égales le long de toutes les branches de l'arbre. Or cette hypothèse, plus connue sous le nom d'*horloge moléculaire* [199, 200], a été invalidée dans de nombreuses études [201, 202, 203, 204]. Ainsi, la seconde méthode d'enracinement consistant à inclure une ou plusieurs séquences provenant d'un groupe taxonomique extérieur aux séquences de l'étude, est privilégiée dans les analyses phylogénétiques. En effet, le dernier ancêtre commun au groupe d'étude et au groupe extérieur sera l'ancêtre le plus vieux de l'arbre, *i.e.* la racine. Cette stratégie nécessite donc une connaissance de l'arbre des espèces

étudiées ainsi que l'existence de séquences homologues dans ce groupe externe. Le principal enjeu est alors de choisir un groupe externe qui ne soit ni trop divergent ni trop proche du groupe d'étude afin d'éviter des biais de reconstruction [121].

Un seul arbre vrai

Soit un jeu de données composé de n séquences homologues, Cavalli-Sforza et Edwards ont montré que le nombre d'arbres possibles, racinés (N_{A_r}) et non racinés ($N_{A_{nr}}$) sont donnés par les formules [205] :

$$N_{A_r} = \frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{2.1}$$

$$N_{A_{nr}} = \frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{2.2}$$

Ainsi, il est possible de reconstruire plus de 34 millions d'arbres racinés à partir de seulement dix séquences homologues [121]. Bien sûr, seulement l'un d'entre eux représente la véritable histoire évolutive de ces dix séquences. Le principal enjeu de la reconstruction phylogénétique est donc de déterminer quel est l'arbre vrai parmi toutes ces topologies possibles.

Pour comparer les topologies de différents arbres, plusieurs grandeurs ont été développées dont la plus utilisée fut introduite en 1981 par Robinson et Foulds [206]. Soient deux arbres non racinés construits en utilisant le même jeu de séquences. Dans ce cas la distance de Robinson et Foulds est définie comme étant égale au double du nombre de branches internes pour lesquelles une bipartition différente est observée. Ainsi, deux arbres sont à une distance $d_T = 2$ s'ils n'ont qu'une bipartition de différence. Pour un arbre non raciné de n feuilles possédant $(n-3)$ branches internes, il existe donc $2(n-3)$ arbres à une distance $d_T = 2$. Une variante de cette mesure a par la suite été développée afin de prendre en compte les longueurs des branches de l'arbre [207].

2.3.2 Modèles d'évolution

Évaluer la distance évolutive d entre deux séquences est l'un des points clefs de la phylogénie moléculaire. Considérons deux séquences S_1 et S_2 ayant divergé d'une séquence ancestrale κ depuis un temps t (Figure 2.10). On définit la distance évolutive entre S_1 et S_2 comme étant le nombre moyen de substitutions par site ayant eu lieu depuis κ.

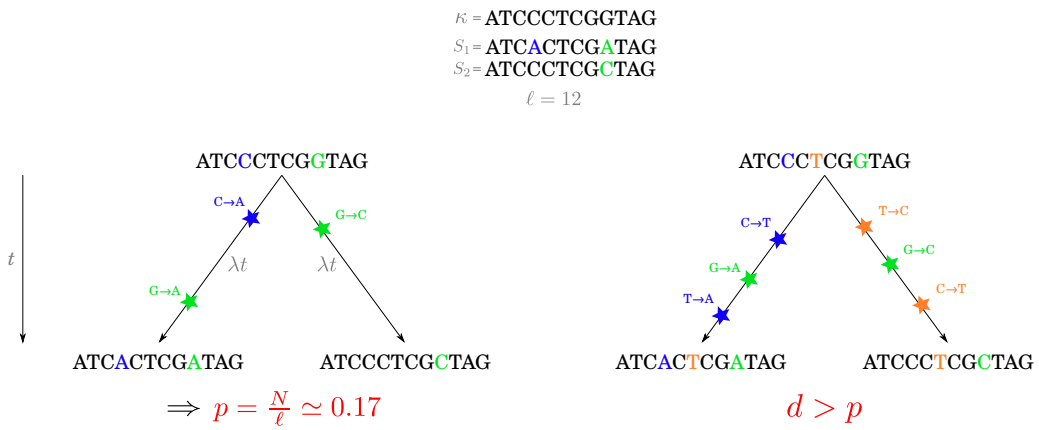

FIGURE *2.10* – ***Évolution d'une séquence κ en deux séquences S_1 et S_2 durant un temps t.*** A gauche le scénario le plus parcimonieux, à droite la véritable histoire évolutive.

Soient N le nombre de substitutions observées entre les séquences S_1 et S_2, et ℓ le nombre de sites homologues comparés. En première approximation on peut considérer que la distance évolutive entre ces deux séquences est égale au rapport N/ℓ, également appelée p-distance ou divergence observée. Dans ce modèle, la substitution est considérée comme un événement de Bernoulli, c'est-à-dire qu'à chaque site il y a ou il n'y a pas de substitution (respectivement succès et échec de l'épreuve de Bernoulli). L'ensemble des ℓ sites suit donc une loi Binomiale. Néanmoins, cette approximation sous-estime la distance évolutive réelle d entre les séquences, en particulier si elles ont divergé depuis longtemps. En effet, ce calcul ne prend pas en compte les substitutions intermédiaires qui ont pu avoir lieu dans chacune des séquences filles.

Modèles markoviens

Afin de prendre en compte les substitutions *multiples* ou *cachées*, la quasi-totalité des modèles utilisés en phylogénie moléculaire sont fondés sur une modélisation markovienne.

Comme dans tout processus de Markov de premier ordre en temps continu, l'état du caractère i à un instant $t + \mathrm{d}t$ ne dépend que de sont état en l'instant t. Ainsi, pour un site donné, l'apparition d'une substitution n'est pas dépendante des substitutions ayant pu avoir lieu avant l'instant t, et les réversions sont possibles. La plupart des modèles d'évolution supposent cinq grandes hypothèses : i) l'indépendance des sites, ii) l'uniformité du processus, iii) son homogénéité, iv) sa stationnarité et v) sa réversibilité. Ainsi, l'évolution de tous les sites de toutes les

séquences est modélisée par le même processus, les taux d'évolution sont constants au cours du temps et, à l'équilibre, la quantité de changements de l'état i vers j est égale à la quantité de changements de j vers i.

Dans le cas des séquences nucléotidiques, chaque nucléotide i est susceptible d'être substitué en un nucléotide j selon un taux q_{ij} (avec $i, j \in \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$). On définit la matrice $\mathbf{Q}$, des *taux de transitions instantanés* du processus de Markov, par :

$$\mathbf{Q} = \begin{bmatrix} -\lambda_A & q_{AT} & q_{AC} & q_{AG} \\ q_{TA} & -\lambda_T & q_{TC} & q_{TG} \\ q_{CA} & q_{CT} & -\lambda_C & q_{CG} \\ q_{GA} & q_{GT} & q_{GC} & -\lambda_G \end{bmatrix}$$

avec λ_j le taux d'évolution instantané du nucléotide j tel que $\lambda_j = \sum_{i, i \neq j} q_{ij}$ puisque, par définition, la somme des lignes d'une matrice des taux d'un processus de Markov vaut zéro.

A partir de cette matrice des taux instantanés, on peut définir la probabilité d'une substitution de i vers j pendant un temps t comme étant $p_{ij}(t + \mathrm{d}t) \simeq q_{ij}\mathrm{d}t$. Comme expliqué précédemment, la présence du nucléotide i à une position donnée à l'instant $t + \mathrm{d}t$ n'est conditionnée que par le nucléotide présent à cette position à l'instant t. Ainsi, deux scénarios sont possibles : i) à l'instant t, i était présent à cette position ou ii) un des trois autres nucléotides était présent. La probabilité de la présence du nucléotide i à l'instant $t + \mathrm{d}t$ est donc déterminée par l'équation :

$$\mathbb{P}_i(t + \mathrm{d}t) = (1 - \lambda_i \mathrm{d}t)\mathbb{P}_i(t) + \sum_{j \neq i} \mathbb{P}_j(t) q_{ji} \mathrm{d}t \tag{2.3}$$

Soit, sous forme matricielle :

$$\begin{aligned} \mathbf{P}(t + \mathrm{d}t) &= \mathbf{P}(t) + \mathbf{Q}\mathbf{P}(t)\mathrm{d}t \\ \Rightarrow \mathbf{P}(t) &= \mathrm{e}^{\mathbf{Q}t} \end{aligned} \tag{2.4}$$

avec $\mathbf{P}(t)$ la matrice dite *matrice de transition*, définie par :

$$\mathbf{P}(t) = \begin{bmatrix} p_{AA}(t) & p_{AT}(t) & p_{AC}(t) & p_{AG}(t) \\ p_{TA}(t) & p_{TT}(t) & p_{TC}(t) & p_{TG}(t) \\ p_{CA}(t) & p_{CT}(t) & p_{CC}(t) & p_{CG}(t) \\ p_{GA}(t) & p_{GT}(t) & p_{GC}(t) & p_{GG}(t) \end{bmatrix}$$

Les valeurs de $\mathbf{Q}$ peuvent être déterminées de façon empirique ou évaluées à

partir du jeu de données étudié. En effet, sous les hypothèses de stationnarité et de réversibilité, à l'équilibre la quantité d'échange $i \to j$ est égale à la quantité d'échange $j \to i$. Avec π_i la fréquence de la base i à l'équilibre, sous ces hypothèses, il vient que :

$$\begin{aligned} \pi_i p_{ij}(t) = \pi_j p_{ji}(t) \qquad & \forall i, j \in \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\} \\ \Rightarrow q_{ij} = \pi_j s_{ij} \qquad & i \neq j \end{aligned} \tag{2.5}$$

avec $s_{ij} = s_{ij}$ le paramètre d'*échangeabilité* entre i et j qu'il est possible de déterminer à partir de l'alignement multiples des séquences étudiées. A partir de (2.5) on en déduit l'expression de la matrice $\mathbf{Q}$ telle que :

$$\mathbf{Q} = \mathbf{S\Pi} = \begin{bmatrix} . & s_{AT} & s_{AC} & s_{AG} \\ s_{AT} & . & s_{CT} & s_{GT} \\ s_{AC} & s_{CT} & . & s_{CG} \\ s_{AG} & s_{GT} & s_{CG} & . \end{bmatrix} \times \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_T & 0 & 0 \\ 0 & 0 & \pi_C & 0 \\ 0 & 0 & 0 & \pi_G \end{bmatrix}$$

L'expression ci-dessus correspond au modèle GTR (*Generalized Time Reversible*) soit le modèle markovien standard comprenant le plus grand nombre de paramètres (six paramètres d'écheangeabilité et trois paramètres de fréquences à l'équilibre)[208]. L'ensemble des modèles classiques (dont certains sont décrits ci-dessous) sont des simplifications du GTR.

Pour déterminer la distance évolutive d séparant deux séquences, l'hypothèse de réversibilité du processus de Markov permet d'établir la relation :

$$d = 2 \sum_i \pi_i \lambda_i t \tag{2.6}$$

avec π_i et λ_i respectivement les fréquences à l'équilibre et les taux d'évolutions du nucléotide i. L'ensemble des ces équations permet de déterminer ensuite une relation entre la p-distance et la distance évolutive d selon les matrices $\mathbf{Q}$ envisagées.

a) Modèles nucléiques

Dans le premier modèle de substitution markovien publié, Jukes et Cantor [209] ont fixé tous les taux de substitutions instantanés comme égaux à α et toutes les fréquences à l'équilibre $\pi_i = 1/4$. De ce fait, les termes λ_i sont donc tous égaux à

3α. La matrice $\mathbf{Q}$ correspondante s'écrit :

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \alpha & \alpha & \alpha \\ \alpha & -\lambda & \alpha & \alpha \\ \alpha & \alpha & -\lambda & \alpha \\ \alpha & \alpha & \alpha & -\lambda \end{bmatrix}$$

La résolution de l'équation (2.4) conduit dans ce cas à des valeurs de la matrice $\mathbf{P}(t)$ égales à :

$$p_{ii}(t) = \frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4\alpha t} \quad \text{et} \quad p_{ij}(t) = \frac{1}{4} - \frac{1}{4}\mathrm{e}^{-4\alpha t} \tag{2.7}$$

avec $i \neq j$. Ainsi, à l'équilibre ($t \to \infty$), les fréquences de chaque base sont effectivement égales à 1/4. Soit p la p-distance entre deux séquences, le modèle de Jukes et Cantor estime que la distance évolutive d vaut :

$$\hat{d} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \tag{2.8}$$

Néanmoins, comme précisé auparavant, toutes les substitutions ne sont pas équivalentes. C'est pourquoi, en 1980, Kimura [210] a proposé un nouveau modèle dans lequel les taux de transition (α) sont différents des taux de transversion (β). Sous ce modèle, la distance évolutive entre deux séquences est donnée par :

$$\hat{d} = -\frac{1}{2}\ln(1 - 2r - v) - \frac{1}{4}\ln(1 - 2v) \tag{2.9}$$

avec r la fréquence des transitions et v celle des transversions ($p = r + v$).

De nombreux autres modèles ont ensuite été proposés pour améliorer la modélisation de l'évolution de séquences nucléiques. Parmi les plus répandus, citons F81 [211], HKY85 [212], T92 [213] et TN93 [214].

Par ailleurs, dès les années 1990, de nombreux modèles s'affranchissant des cinq hypothèses décrites précédemment (indépendance des sites, uniformité, homogénéité, stationnarité et réversibilité) ont été développés pour les séquences nucléiques. Concernant les modèles ne considérant pas l'indépendance des sites on peut citer ceux établis par Schöniger *et al.* [215] et Muse [216]. Un plus grand nombre de modèles non homogènes ont été publiés dont ceux de Yang et Roberts [217], de Foster [218] ou encore le modèle de Jayaswal *et al.* [219]. Concernant

les modèles non stationnaires, les plus couramment utilisés sont le LogDet [220] et celui de Blanquart et Lartillot [221]. Un des premiers modèles non réversibles est celui de Lobry [222], suivi de celui de Galtier et Gouy [223] (non homogène et non stationnaire).

b) Modèles protéiques

Si dans la modélisation de l'évolution des séquences nucléiques à l'aide de chaînes de Markov seuls quatre états de caractère sont à considérer, l'analyse des séquences protéiques nécessite l'utilisation d'une matrice $\mathbf{Q}$ de taille 20×20 et donc l'évaluation de 189 paramètres d'échangeabilité (matrice $\mathbf{S}$) et de 19 fréquences à l'équilibre (matrice $\mathbf{\Pi}$). Ainsi, pour les analyses protéiques, il est courant d'utiliser des valeurs empiriques déterminées sur des jeux de données de référence. Fondées sur des ensembles de séquences alignées, on distingue les matrices construites par des approches utilisant le maximum de parcimonie de celles inférées par maximum de vraisemblance.

PAM et JTT

PAM fut le premier modèle de substitution markovien construit sur la base d'un ensemble de 1300 séquences réparties en 71 familles « proches » (au moins 85% d'identité entre chaque paire possible à l'intérieur d'une famille) [147]. Pour chaque famille (alignement), un arbre est calculé et les séquences ancestrales sont inférées par maximum de parcimonie. Le nombre de substitutions entre paires de séquences est ensuite comptabilisé, et leurs fréquences relatives calculées. La matrice PAM1 est alors définie comme la matrice des taux de substitutions attendus si 1% des acides aminés sont mutés. A partir de cette matrice, les autres matrices PAM sont déduites par exponentiation de PAM1. Ainsi PAM250 correspond à PAM1^{250}. Publiée en 1992, la matrice JTT est basée sur une méthodologie très similaire à celle de PAM [171]. La différence majeure résidant dans le nombre de séquences étudiées : 16130 au lieu de 1300.

WAG et LG

En 2001, une nouvelle méthodologie pour le calcul des échangeabilités entre acides aminés fut proposée par Whelan et Goldman [224]. La matrice WAG (***W****helan* ***A****nd* ***G****oldman*) est en effet calculée par maximum de vraisemblance à partir de 182 alignements comptabilisant 3905 séquences. Pour chaque alignement, les distances

entre les paires de séquences sont calculées grâce à la matrice PAM, puis l'arbre correspondant est inféré par la méthode NJ. A partir de cet arbre considéré comme vrai, les longueurs des branches sont ré-estimées par maximum de vraisemblance en utilisant le modèle JTT. La matrice WAG est alors définie comme la matrice des taux maximisant la vraisemblance des données pour l'arbre considéré.

La matrice LG (pour ***L****e and* ***G****ascuel*), proposée en 2008 [225], est quant à elle fondée sur la même méthodologie que WAG mais autorise des taux de substitutions différents lors du recalcul des longueurs de branches (utilisation d'une loi Γ, voir ci-dessous). De plus, les auteurs ont décidé d'utiliser la matrice WAG et non JTT pour le second calcul ainsi que d'effectuer deux itérations supplémentaires de recalcul de topologies et de longueurs de branches afin d'améliorer la précision des taux inférés.

Modèles s'affranchissant des hypothèses classiques

Les années 2000 ont vu le développement de modèles protéiques essayant de s'affranchir de l'hypothèse d'homogénéité afin de prendre en compte les contraintes biologiques telles que les structures secondaires et tertiaires des protéines. En 2004, Lartillot et Philippe [226] présentent ainsi le modèle de mélange CAT (***CAT****egories*) dans lequel les sites sont divisés en catégories, chacune possédant une distribution de fréquences à l'équilibre spécifique (matrice $\mathbf{\Pi}$). Dans cette version du modèle, développée pour les inférences phylogénétiques bayésiennes, le nombre de catégories est déterminé en utilisant une distribution de Dirichlet et un processus de Monte-Carlo. La matrice d'échangeabilités $\mathbf{S}$ est, quant à elle, fixe pour toutes les catégories (il s'agit de JTT, WAG, Poisson ou encore MtREV [227]). En 2008, les auteurs proposent une version empirique du modèle [228], applicable aux reconstructions par maximum de vraisemblance. Ils déterminent en effet les profils des K catégories ($K \in \{10, 20, \ldots, 60\}$) à l'aide d'un algorithme EM (***E****xpectation-****M****aximization*) sur un ensemble de plus de 32000 alignements. Le modèle CAT20 ressort alors comme le meilleur compromis entre nombre de catégories et amélioration de l'arbre.

Enfin, la même année cette équipe présente également d'autres modèles de mélange estimés itérativement de façon supervisée (EXO, EX2 et EX3) et non supervisée (UL2 et UL3) [229] et ce sur un ensemble de 1771 alignements. Dans ces modèles, les sites sont divisés selon leur accessibilité au solvant, accessibilités calculées grâce à la structure tertiaire des protéines. EX2 et UL2 sont ainsi des modèles à deux matrices de taux (sites exposés et non exposés) tandis que EH0, EX3 et UL3 comprennent trois matrices différentes. Dans ce même article, les auteurs

comparent ces modèles ainsi que JTT et WAG au modèle LG (Figure 2.11).

Figure omise pour des raisons de droits d'auteur.

FIGURE *2.11* – ***Gain en AIC par rapport au modèle LG.*** Les valeurs négatives (respectivement positives) représentent des modèles moins bons (respectivement meilleurs) que LG. En blanc (resp. noir) les alignements avec un index de saturation inférieur (resp. supérieur) à deux. [229]

Les modèles comprenant une seule matrice de taux présentent systématiquement des valeurs au test AIC (***A****kaike* ***I****nformation* ***C****riterion*) [230] inférieures à celles obtenues par les modèles autorisant des taux différents entre les sites de l'alignement.

c) Correction par la loi Gamma

Sous l'hypothèse d'uniformité, le taux global de substitutions $\lambda = \sum_i \pi_i \lambda_i$ est le même pour tous les sites de l'alignement, ce qui constitue une hypothèse dont on sait qu'elle est erronée. Il a donc été proposé de moduler la valeur de ce taux par un facteur correctif r, ceci grâce à l'emploi d'une distribution Gamma. Cette distribution est caractérisée par deux paramètres : α (ou paramètre de *forme*) et β (ou paramètre d'*échelle*). Néanmoins, en phylogénie moléculaire, on fixe $\beta = 1/\alpha$ puisqu'on ne s'intéresse qu'à des taux d'évolution relatifs. Ainsi, l'allure de la distribution Gamma n'est déterminée que par la valeur de α. Plus α est grand, plus la variance de r diminue ; l'hypothèse d'uniformité correspondant au cas extrême où $\alpha \to \infty$. Plusieurs méthodes d'estimation de α ont été proposées mais c'est celle publiée par Yang [231] qui est la plus fréquemment utilisée.

2.3.3 Principales méthodes d'inférence d'arbre

L'inférence d'arbres phylogénétiques et la recherche de l'arbre vrai peut être effectuée selon quatre grandes familles de méthodes :

- le maximum de parcimonie,
- les méthodes des distances,
- le maximum de vraisemblance,
- l'inférence bayésienne.

Les sections suivantes ont pour but de présenter brièvement les principes généraux ainsi que les caractéristiques de chacune de ces méthodes. Pour une présentation plus détaillée, se référer au livre *Concepts et Méthodes en Phylogénie Moléculaire* publié en 2010 par Guy Perrière et Céline Brochier-Armanet [121].

a) Maximum de Parcimonie

La méthode du maximum de parcimonie [232] repose sur le principe du *rasoir d'Occam*, c'est-à-dire que la solution la plus simple est la plus vraisemblable. Selon cette méthode l'arbre retenu correspondra donc à la topologie impliquant le moins d'événements de substitutions possibles. Le premier algorithme récursif pour cette méthode fut formalisé en 1971 par Fitch [233].

Pour un arbre non raciné donné, la première étape du maximum de parcimonie consiste à enraciner aléatoirement cet arbre puis, pour chaque site i de l'alignement, on calcule $N_c^{(i)}$ le nombre minimal de changements depuis les feuilles jusqu'à la racine. La longueur L de l'arbre est ensuite obtenue en sommant l'ensemble des changements calculés à partir des ℓ sites considérés :

$$L = \sum_{i=1}^{\ell} N_c^{(i)} \tag{2.10}$$

Cette opération est répétée pour l'ensemble des topologies possibles et l'arbre ayant la longueur L minimale est retenu, c'est-à-dire l'arbre 1 dans l'exemple de la Figure 2.12.

Lors d'une reconstruction par maximum de parcimonie, seuls certains sites de l'alignement sont discriminants. Un site est ainsi dit *informatif* s'il présente au moins deux états trouvés au moins deux fois. D'un autre côté, un site *invariant* tel que le troisième site de la Figure 2.12 n'est pas informatif puisqu'il n'a aucun impact sur la longueur des différentes topologies possibles. De même, le cinquième

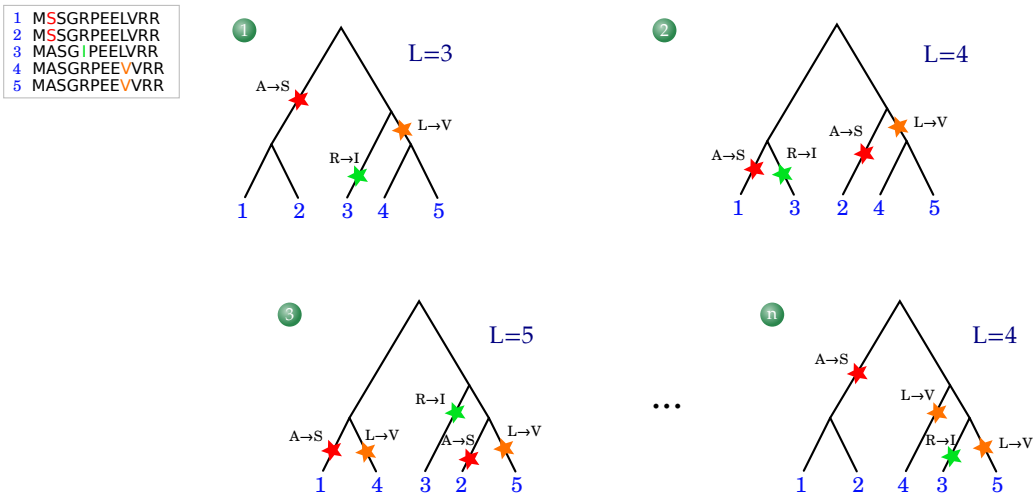

FIGURE *2.12* – ***Illustration du maximum de parcimonie.*** Considérant l'alignement en haut à gauche, l'arbre 1 est le plus parcimonieux. En effet les autres topologies impliquent un plus grand nombre de substitutions.

site, bien que variable pour la séquence 3 est *non informatif* puisqu'il implique le même nombre de substitutions quelle que soit la topologie considérée.

Une des caractéristiques principales de cette méthode est qu'elle peut aboutir à l'obtention de plusieurs arbres équiparcimonieux. Dans ces cas, il n'est pas possible de choisir l'un de ces arbres et le résultat est souvent représenté sous la forme de ce que l'on appelle un *arbre consensus* [234]. Celui-ci présente alors une polytomie au niveau des nœuds conduisant à des sous-arbres différents dans les arbres équiparcimonieux. Soient A, B et C trois séquences d'un arbre de gènes possédant trois topologies équiparcimonieuses : τ_1, τ_2 et τ_3. Si dans l'arbre τ_1, A est groupé avec B puis avec C mais qu'il est tout d'abord groupé avec C dans τ_2 et τ_3, alors l'arbre consensus dit *strict* sera formé d'une polytomie au niveau du nœud conduisant à ces trois feuilles. Il est néanmoins possible de construire des arbres consensus plus flexibles, appelés arbres *consensus majoritaires* [235], dans lesquels les nœuds ne sont pas regroupés si l'une des bipartitions est représentée majoritairement dans ces arbres équiparcimonieux. Ainsi, l'arbre consensus majoritaire dit à 60% de l'exemple précédant ne possédera pas de polytomie, la feuille A sera groupée avec C puis avec B car deux tiers des topologies équiparcimonieuses présentent cette topologie. Les arbres consensus stricts correspondent donc aux arbres majoritaires à 100%.

Par ailleurs, dans sa conception, le maximum de parcimonie considère chaque substitution comme équiprobable ; toutes ont donc le même poids lors du calcul de la longueur de l'arbre. Or, comme décrit dans la section 2.2, toutes les substitutions

ne sont pas observées avec la même fréquence et ne représentent donc pas un coût équivalent. La prise en compte d'une pondération dans le calcul de la longueur de l'arbre a été rendu possible grâce à l'algorithme développé en 1975 par Sankoff [236]. On parle alors de parcimonie pondérée.

Enfin, bien qu'il soit possible d'inférer des longueurs de branches à partir du nombre de substitutions impliquées [237, 238], les arbres obtenus par maximum de parcimonie sont le plus souvent représentés sous forme de cladogrammes.

La méthode par maximum de Parcimonie nécessite en théorie l'exploration de l'ensemble des topologies possibles. Leur nombre croissant de façon exponentielle avec la taille du jeu de données (équations 2.1 et 2.2), les logiciels d'inférence d'arbres ont généralement recours à des heuristiques itératives afin de rendre les calculs possibles. Les méthodes NNI (***N****earest* ***N****eighbor* ***I****nterchange*) et SPR (***S****ubtree* ***P****runing and* ***R****egrafting*) figurent parmi les plus utilisées. Le principe général est d'explorer successivement les topologies « proches » d'une topologie initiale, puis de sélectionner le meilleur arbre qui constituera la topologie initiale de l'itération suivante. Dans l'approche NNI, les topologies proches correspondent aux $(2n-6)$ arbres situés à une distance de Robinson et Foulds $d_T = 2$. De complexité en $O(n)$, ce processus permet donc une exploration rapide des topologies.

Plus sophistiquée, la méthode SPR considère quant à elle $4(n-3)(n-2)$ topologies proches. En effet, à chaque itération une des branches de l'arbre est aléatoirement « coupée » en deux, séparant ainsi l'arbre en une partie appelée *résiduelle* et une partie dite *élaguée*. Cette dernière est ensuite placée successivement sur chacune des branches de la partie résiduelle. Comme pour la méthode NNI, l'ensemble des topologies ainsi créées sont évaluées et la meilleure est sélectionnée. Légèrement moins rapide, cet algorithme est de complexité en $O(n^2)$. Si la topologie initiale considérée est proche de l'arbre vrai, ces processus permettent donc d'éviter l'exploration des topologies les moins probables. Néanmoins ils ne garantissent pas l'obtention de l'arbre le plus parcimonieux.

b) Méthodes de distances

Fondées sur l'utilisation de modèles évolutifs tel que ceux présentés à la section 2.3.2, les différentes méthodes des distances sont apparues pour la première fois à la fin des années 1960 [239, 205]. Leur objectif est d'obtenir un arbre dont l'ensemble des distances δ_{ij} (ou distances *patristiques*) séparant le nœud i du nœud j soient les plus proches possible des valeurs des distances d_{ij} calculées au moyen

d'un modèle donné. Parmi ces méthodes on distingue les algorithmes basés sur des regroupements (ou *clustering*) de ceux utilisant un critère d'optimisation. Le clustering par UPGMA [162] et le minimum d'évolution [240, 241, 242] constituent les principaux représentants respectifs de ces deux catégories.

UPGMA

UPGMA, est une méthode de classification ascendante hiérarchique au lien moyen, c'est-à-dire que cet algorithme regroupe successivement les feuilles de l'arbre jusqu'à aboutir à la racine de l'arbre. La conséquence de cela est qu'il s'agit d'une des seules méthodes construisant des arbres racinés automatiquement. Soit $\mathbf{D} = (d_{ij})$ la matrice des distances entre les séquences du jeu de données, l'algorithme UPGMA débute par l'identification des deux nœuds i et j les plus proches. La création d'un nœud ancestral à i et j nommé u implique l'attribution d'une longueur aux deux branches créées, qui est fixée à $d_{ij}/2$. L'algorithme consiste ensuite à calculer l'ensemble des distances séparant u de tous les autres nœuds de l'arbre et à remplacer les lignes et colonnes correspondant à i et j par les valeurs calculées pour u dans la matrice $\mathbf{D}$. Les trois étapes précédentes sont itérées tant qu'il reste plus d'un élément dans $\mathbf{D}$.

Pour n séquences, l'algorithme UPGMA répète ces opérations $(n-1)$ fois, ce qui le rend très rapide par rapport aux autres méthodes. Néanmoins, le résultat obtenu repose sur l'hypothèse de l'horloge moléculaire, considérant qu'à partir d'un nœud la même quantité d'évolution s'écoule dans les deux lignées filles. En effet, l'arbre généré par cette méthode est dit *ultramétrique*, c'est-à-dire que la distance de chaque feuille à la racine de l'arbre est égale.

Moindres carrés et minimum d'évolution

La méthode des moindres carrés consiste à minimiser la somme des carrés des écarts entre les distances patristiques et les distances calculées au moyen d'un modèle. Pour une topologie donnée, ceci revient à déterminer quelles sont les valeurs des longueurs de branches minimisant :

$$Q = \sum_{i<j} w_{ij}(d_{ij} - \delta_{ij})^2 \tag{2.11}$$

avec w_{ij} les valeurs de pondération associées à chaque paire (i, j). Il est ensuite nécessaire de répéter ce calcul pour chacune des topologies possibles.

Le problème est que la résolution l'équation précédente nécessite d'effectuer une

inversion de matrice pouvant être de grande dimension (proportionnelle au nombre de séquences), cette inversion pouvant s'accompagner de dérives numériques importantes. Pour pallier ce problème, Fitch et Margoliash [239] ont développé un algorithme permettant d'approximer la construction d'un arbre par la méthode des moindres carrés.

Soient trois séquences A, B et C et leurs distances respectives d_{AB}, d_{AC} et d_{BC} calculées à l'aide d'une matrice de substitutions (Figure 2.13). L'enjeu est donc de déterminer d'une part l'ordre de regroupement des séquences et d'autre part les longueurs de branches associées. De manière analogue à la méthode UPGMA, il est possible d'introduire un nœud ancestral u reliant les deux séquences les plus proches, A et B par exemple. Afin de s'affranchir de l'hypothèse d'horloge moléculaire, les distances de l'arbre seront calculées à l'aide du système de trois équations à trois inconnues suivant :

$$\begin{cases} d_{AB} &= \delta_{Au} + \delta_{Bu} \\ d_{AC} &= \delta_{Au} + \delta_{Cu} \\ d_{BC} &= \delta_{Bu} + \delta_{Cu} \end{cases} \Leftrightarrow \begin{cases} \delta_{Au} &= (d_{AB} + d_{AC} - d_{BC})/2 \\ \delta_{Bu} &= (d_{AB} + d_{BC} - d_{AC})/2 \\ \delta_{Cu} &= (d_{AC} + d_{BC} - d_{AB})/2 \end{cases}$$

L'arbre composé par ces trois séquences est donc entièrement déterminé. Dans le cas général d'un jeu de données composé de plus de trois séquences, Fitch et Margoliash ont montré que l'on peut se ramener à la situation précédente en considérant que C regroupe l'ensemble des séquences hormis A et B. Dans ce cas, les distances patristiques seront calculées avec d_{AC} (respectivement d_{BC}) égales à la moyenne des distances entre toutes les séquences de C et la séquence A (respectivement B).

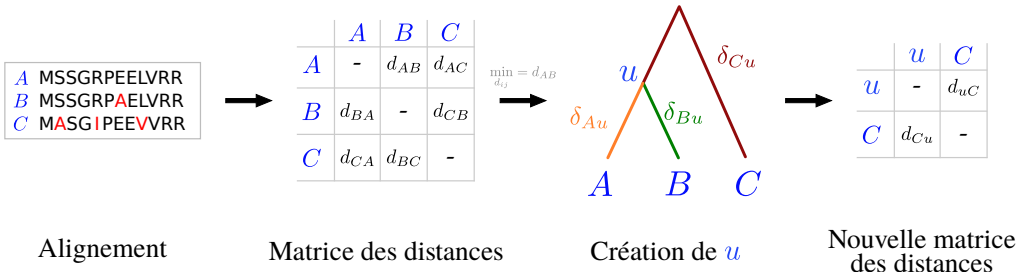

FIGURE 2.13 – ***Illustration de la méthode de Fitch et Margoliash [239] pour trois séquences protéiques.***

Néanmoins, Fitch et Margoliash soulignent que le choix des deux premières séquences à grouper impacte fortement la topologie obtenue. Ainsi, la seconde partie de leur algorithme consiste à utiliser différents groupements initiaux de séquences et comparer les topologies obtenues. Pour cela, chaque valeur δ_{ij} séparant deux

feuilles i et j de l'arbre sont calculées et l'arbre retenu est celui qui minimise le carré des écarts entre les distances d_{ij} et δ_{ij}. En théorie, cela implique d'appliquer l'algorithme des triplets évoqués à partir des $n(n-1)/2$ paires de séquences initiales possibles, avec n le nombre de séquences du jeu de données [121]. Du fait de sa complexité en $O(n^5)$ l'usage de cet algorithme est réservé aux jeux de données de petite taille.

Plus récemment, Rzhetsky et Nei [242] ont proposé une méthode exacte permettant d'effectuer le calcul des longueurs de branches aux moindres carrés sans passer par une étape de calcul matriciel. La complexité de cet algorithme est en $O(n^2)$ [243].

La méthode dite du *minimum d'évolution* est similaire à celle des moindres carrés excepté que le critère optimisé est la longueur totale de l'arbre. Ainsi, la topologie considérée la plus vraisemblable sera celle dont la somme des longueurs des branches est minimale [240, 241, 242]. Qui plus est, Rzhetsky et Nei ont également proposé de réduire l'espace des topologies explorées afin de réduire le temps de calcul [241, 242]. En effet, ils n'examinent que les topologies situées à une distance de Robinson et Foulds $d_T = 2$ et $d_T = 4$ de l'arbre obtenu par la méthode NJ.

Neighbour-Joining

Performante et rapide, la méthode NJ fut proposée en 1987 par Saitou et Nei [244]. Dans cet algorithme, n'inférant qu'un seul arbre, deux nœuds deviendront *voisins* (*i.e.* connectés par un nœud interne) que s'ils minimisent la longueur totale l'arbre résultant de leur regroupement. A partir d'une topologie en étoile (toutes les feuilles étant connectées par un unique nœud interne), la procédure de création de nouveaux voisins est répétée jusqu'à l'obtention des $n-3$ branches internes composant un arbre de n séquences. Une variante de cet algorithme plus rapide et donc plus utilisée par les logiciels de phylogénie moléculaire fut publiée un an plus tard par Studier et Keppler [245]. Néanmoins, c'est la variante avec pondération nommée BioNJ qui est la plus efficace dans le cas de séquences aux vitesses d'évolution très différentes [246].

c) Maximum de vraisemblance

Le *maximum de vraisemblance* est un concept statistique introduit par Fisher dans les années 1920 [247]. D'abord restreinte à certains domaines de la biologie tel

que la génétique des populations, son utilisation en phylogénie est introduite pour la première fois en 1971 par Neyman [248]. Néanmoins ce n'est que deux ans plus tard que Felsenstein a généralisé son emploi avec le développement d'un algorithme performant [249].

Soit S l'ensemble des sites sélectionnés et $\boldsymbol{\theta}$ le vecteur des paramètres caractérisant un arbre (sa topologie τ, ses longueurs de branches $\mathbf{b}$ et les paramètres du modèle d'évolution utilisé $\boldsymbol{\vartheta}$). La vraisemblance de $\boldsymbol{\theta}$ est notée $L(\boldsymbol{\theta})$ (pour ***L****ikelihood*) et est définie comme la probabilité d'observer S avec l'utilisation des paramètres $\boldsymbol{\theta}$. La reconstruction d'un arbre phylogénétique par maximum de vraisemblance implique donc de trouver les valeurs des paramètres $\boldsymbol{\theta}$ pour lesquelles la vraisemblance est maximale.

Sous l'hypothèse que les ℓ sites $S^{(i)}$ ($i \in \{1, 2, ..., \ell\}$) de l'alignement sont indépendants et identiquement distribués on a :

$$L(\boldsymbol{\theta}) = \mathbb{P}(S|\boldsymbol{\theta}) = \prod_{i=1}^{\ell} \mathbb{P}(S^{(i)}|\boldsymbol{\theta}) \tag{2.12}$$

Soit, par transformation logarithmique :

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{\ell} \ln \mathbb{P}(S^{(i)}|\boldsymbol{\theta}) \tag{2.13}$$

Ainsi, pour obtenir la vraisemblance d'un arbre il est nécessaire de calculer la vraisemblance de chacun des sites du jeu de données S. Soient τ une des topologies racinées possibles, $\mathbf{b}$ le vecteur des longueurs des six branches et $\boldsymbol{\vartheta}$ le vecteur des paramètres du modèles évolutif utilisé. Notons U_j et V_k, respectivement les feuilles et nœuds internes de l'arbre , avec $1 < j < 4$ et $1 < k < 3$. Enfin, désignons par u_j et $v_k \in \{\mathrm{A}, \mathrm{C}, \mathrm{D}, .., \mathrm{V}\}$, les états de caractère au site i des feuilles U_j et des nœuds V_k respectivement. La vraisemblance au site i est alors déterminée par :

$$L^{(i)} = \mathbb{P}(S^{i)}|\boldsymbol{\theta}) = \mathbb{P}(u_1, u_2, u_3, u_4, v_1, v_2, v_3|\boldsymbol{\theta}) \tag{2.14}$$

Soit (Figure 2.14) :

$$\begin{aligned} L^{(i)} &= \mathbb{P}(u_1|v_2, b_1)\mathbb{P}(u_2|v_2, b_2)\mathbb{P}(u_3|v_3, b_3) \\ &\quad \times \mathbb{P}(u_4|v_3, b_4)\mathbb{P}(v_2|v_1, b_5)\mathbb{P}(v_3|v_1, b_6)\mathbb{P}(v_1) \end{aligned} \tag{2.15}$$

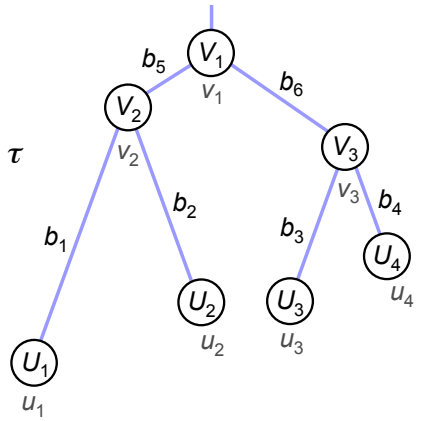

FIGURE *2.14* – ***Exemple de topologie pour un jeu de données de quatre séquences.*** Le jeu de données considéré est composé de quatre séquences protéiques nommées U_j. Les nœuds V_k correspondent à leurs derniers ancêtres communs hypothétiques tandis que b_n désignent les longueurs des branches.

Mais si les états de caractères u_j (séquences du jeu de données) sont connus, les états v_k qui concernent les séquences ancestrales V_k, ne le sont pas. Dans cet exemple à trois nœuds internes, il existe $20^3 = 8000$ scénarios possibles dont il est nécessaire d'évaluer les probabilités conditionnelles. La vraisemblance d'un site i de l'alignement sera donc la somme de tous les scénarios possibles soit :

$$\begin{aligned} L^{(i)} &= \sum_{v_1}\sum_{v_2}\sum_{v_3} \mathbb{P}(u_1|v_2,b_1)\mathbb{P}(u_2|v_2,b_2)\mathbb{P}(u_3|v_3,b_3) \\ &\qquad \times\mathbb{P}(u_4|v_3,b_4)\mathbb{P}(v_2|v_1,b_5)\mathbb{P}(v_3|v_1,b_6)\mathbb{P}(v_1) \end{aligned} \tag{2.16}$$

Les probabilités conditionnelles intervenant dans ce calcul sont déterminées à l'aide de modèles évolutifs tels que ceux décrits dans la section 2.3.2. La vraisemblance totale de $\boldsymbol{\theta}$ est ensuite calculée en sommant les vraisemblances de chaque site i de l'alignement. Pour une topologie donnée, ce calcul est répété en faisant varier les longueurs de branches jusqu'à trouver le vecteur $\mathbf{b}$ maximisant la vraisemblance. Enfin, les calculs doivent être théoriquement réitérés pour l'ensemble des topologies possible. Comme cela n'est généralement pas possible, une exploration d'un sous-ensemble des topologies est réalisé au moyen des algorithmes présentés dans la section sur le maximum de parcimonie (*i.e.* NNI et SPR).

Sous la forme donnée dans l'équation (2.16) la méthode du maximum de vraisemblance nécessite un nombre considérable de calculs. Si elle est aujourd'hui la méthode privilégiée dans les analyses phylogénétiques, c'est grâce à l'élaboration de plusieurs algorithmes permettant un gain de temps de calcul considérable, tout

en conservant sa capacité à s'approcher du maximum recherché. La principale amélioration a été proposée par Felsenstein [249] et est connue sous le terme d'*élagage* (ou *pruning* en anglais). L'idée est d'éviter de recalculer plusieurs fois les mêmes probabilités. En effet, dans l'équation (2.16) la somme sur v_3 est calculée pour chaque v_1 mais également pour chaque v_2 possible. La simplification proposée par Felsenstein s'écrit :

$$\begin{aligned} L^{(i)} &= \sum_{v_1} \mathbb{P}(v_1) \left[\sum_{v_2} \mathbb{P}(v_2|v_1, b_5)\mathbb{P}(u_1|v_2, b_1)\mathbb{P}(u_2|v_2, b_2) \right] \\ &\qquad \times \left[\sum_{v_3} \mathbb{P}(v_3|v_1, b_6)\mathbb{P}(u_3|v_3, b_3)\mathbb{P}(u_4|v_3, b_4) \right] \end{aligned} \tag{2.17}$$

d) Inférence Bayésienne

Bien que développées dès le XVIIIème siècle, les approches bayésiennes n'ont connu de véritable succès qu'avec l'apparition de moyen de calculs performants, c'est-à-dire à l'ère des ordinateurs du XXème siècle. En effet, fondées sur le théorème de Bayes, ces méthodes d'inférences statistiques font intervenir des calculs de probabilités conditionnelles qui nécessitent l'exploration de l'ensemble des événements possibles. En phylogénie cela se traduit par l'intégration d'un nombre très importants de paramètres tels que l'ensemble des topologies et des longueurs de branches (équation 2.19), c'est pourquoi l'utilisation de l'approche bayésienne en phylogénie moléculaire ne date que de la fin des années 1990 [250].

Pour des données continues, l'expression du Thèorème de Bayes est la suivante :

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})}{f(\mathbf{x})} = \frac{f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})}{\int f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \tag{2.18}$$

avec $f(\boldsymbol{\theta}|\mathbf{x})$ la *probabilité postérieure*, $f(\boldsymbol{\theta})$ la *distribution a priori* de $\boldsymbol{\theta}$ et $f(\mathbf{x}|\boldsymbol{\theta})$ la *vraisemblance* de $\boldsymbol{\theta}$. La constante de normalisation $f(\mathbf{x})$ est, quant à elle, obtenue en intégrant la vraisemblance sur la distribution *a priori*.

En phylogénie moléculaire, $\mathbf{x}$ correspond aux données disponibles S (*i.e.* aux sites gardés de l'alignement multiple) et $\boldsymbol{\theta} = \{\tau, \mathbf{b}, \boldsymbol{\vartheta}, \alpha\}$ à l'ensemble des paramètres caractérisant un arbre :

$$f(\tau, \mathbf{b}, \boldsymbol{\vartheta}, \alpha|S) = f(\boldsymbol{\theta}|S) = \frac{f(\boldsymbol{\theta})f(S|\boldsymbol{\theta})}{\int f(\boldsymbol{\theta})f(S|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \tag{2.19}$$

Dans ce cas, il n'existe habituellement pas de solution analytique du calcul de la constante de normalisation. Plusieurs techniques d'approximations sont possibles mais ce chapitre ne détaillera que celle du MCMC (***M**arkov **C**hain **M**onte-**C**arlo*), couramment utilisée en phylogénie moléculaire. Cette approche permet en effet de s'affranchir du calcul de l'intégrale au dénominateur de l'équation précédente.

La méthode MCMC est un processus d'exploration de l'espace des valeurs possibles des paramètres étudiés qui, dans le cas de la phylogénie moléculaire, est particulièrement considérable. Un chemin dans cet espace prend la forme d'une chaîne de Markov. Il peut être totalement aléatoire ou, comme dans l'algorithme de Metropolis-Hasting [251, 252], partiellement guidé. En effet, le but de cette approche étant de déterminer le jeu de paramètres $\boldsymbol{\theta}_{opt}$ maximisant $f(\boldsymbol{\theta}|S)$, il est nécessaire que les chaînes de Markov convergent vers les bonnes valeurs des dits paramètres. Partant d'un point quelconque de l'espace $\boldsymbol{\theta}_{t=0} = \{\tau_0, \mathbf{b}_0, \boldsymbol{\vartheta}_0, \alpha_0\}$, le déplacement vers un autre point $\boldsymbol{\theta}^*$ candidat à l'instant $t+1$ est déterminé par la *probabilité d'acceptation* r définie par :

$$r = \min\left[1, \frac{f(\boldsymbol{\theta}^*|S)}{f(\boldsymbol{\theta}_t|S)}\right] = \min\left[1, \frac{f(\boldsymbol{\theta}^*)f(S|\boldsymbol{\theta}^*)}{f(\boldsymbol{\theta}_t)f(S|\boldsymbol{\theta}_t)}\right] \tag{2.20}$$

Si r vaut 1, cela signifie que le jeu de paramètres $\boldsymbol{\theta}^*$ permet d'obtenir une meilleure probabilité postérieure, on fixe alors $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}^*$. Dans le cas contraire, le nouveau jeu de paramètres n'est pas forcément rejeté. En effet, pour éviter le piégeage de la chaîne dans un maximum local, il est nécessaire d'autoriser l'acceptation de probabilités postérieures moins élevées. Pour ce faire, un nombre u est aléatoirement tiré dans une distribution uniforme $\mathcal{U}(0,1)$. Si $u < r$ alors $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}^*$, sinon le nouvel état ne remplace pas l'ancien et la chaîne de Markov reste au même point de l'espace. Avec cette procédure, plus la diminution de probabilité postérieure induite par l'utilisation de $\boldsymbol{\theta}^*$ est importante, moins ce jeu de paramètres a de chances d'être gardé. A l'étape suivante, un nouveau jeu de paramètres $\boldsymbol{\theta}^*$ proche de $\boldsymbol{\theta}_t$ sera généré et, peut-être, accepté. Cette méthode nécessite une exploration judicieuse des vecteurs « proches » du vecteur de paramètres considéré au temps t ainsi qu'un nombre d'itérations suffisant pour atteindre la distribution stationnaire correspondant théoriquement à la probabilité postérieure maximale. Dans l'exemple de la Figure 2.15 il est ainsi nécessaire d'effectuer 10000 itérations MCMC pour explorer l'espace des paramètres représenté par deux pics dans un plan.

Par ailleurs, l'utilisation d'une seule chaîne MCMC peut conduire au piégeage

de celle-ci dans un maximum local des valeurs de probabilités postérieures. Pour s'affranchir de ce problème, il est courant de lancer plusieurs chaînes en parallèle, c'est ce que l'on appelle le MCMCMC ou MC^3 (***M**etropolis **C**oupling of **MCMC***). Dans ce cas, on considère que les bonnes valeurs des paramètres sont atteintes lorsque les différentes chaînes convergent vers les mêmes résultats. Afin de tester cette convergence, il est classique d'utiliser la méthode de Gelman et Rubin [253] qui compare les variances intra et inter chaînes. La convergence est considérée comme étant atteinte si le rapport de Gelman et Rubin est proche de 1.

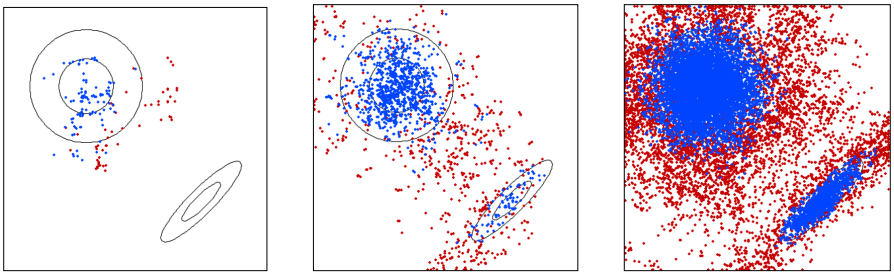

FIGURE *2.15* – ***Illustration de l'exploration de l'espace par deux chaînes de Markov en fonction du nombre d'itérations effectuées.*** Les maximums locaux sont représentés par deux ellipses concentriques. Les points de couleur bleue correspondent à une chaîne froide tandis que ceux de couleur rouge correspondent à une chaîne chaude (tiré de [121]).

Enfin, toujours dans l'objectif d'éviter le piégeage d'une chaîne dans un maximum local, il est également courant d'utiliser des chaînes MC^3 dont l'amplitude des pas (c'est-à-dire l'éventail des valeurs possibles pour $\boldsymbol{\theta}^*$) est différente. Dans ce cas on parle de chaînes *froides* lorsque cette amplitude est faible et de chaînes *chaudes* lorsque celle-ci est forte. L'idée étant qu'à intervalles réguliers la chaîne froide échange sa position avec la chaîne chaude. La Figure 2.15 montre un exemple graphique d'exploration avec une chaîne froide et une chaîne chaude d'un espace des paramètres formant un plan. Dans cet exemple, l'utilisation de la seule chaîne froide ne permet pas d'atteindre le second maximum alors que l'emploi des deux chaînes permet d'y parvenir.

La principale difficulté de l'approche bayésienne réside dans le choix des distributions *a priori* des paramètres considérés. Par exemple, Yang et Rannala [250] ont utilisé un modèle markovien de type naissance-mort pour établir certains de leurs *a priori*. Si aucune connaissance sur la distribution théorique de ces paramètres n'est disponible, il est d'usage d'utiliser un *a priori vague* ou *non informatif*. Ainsi, des distributions uniformes sont choisies dans le cas de la topologie τ de l'arbre ainsi

que pour le paramètre α de la loi Gamma. Les longueurs de branches sont, quant à elles, modélisées par une distribution uniforme ou une exponentielle décroissante qui présente l'avantage d'attribuer des probabilités faibles aux grandes longueurs de branches. Enfin, les paramètres $\boldsymbol{\vartheta}$ des modèles évolutifs, peuvent être représentés par des distributions de Dirichlet.

2.4 Analyses phylogénétiques et conclusions biologiques

2.4.1 Hypothèses sur les fonctions biologiques

Si la phylogénie moléculaire permet de déterminer les relations de parentés entre les espèces, elle peut également être utilisée en inférence fonctionnelle. En effet, comme théorisé en 1998 par Eisen [254], l'association de données biologiques fonctionnelles disponibles pour certains gènes et d'informations portées par les arbres phylogénétiques permet d'émettre des hypothèses quant à la fonction biologique de leurs homologues.

Soit un arbre de gènes composé des séquences de deux paralogues A et B, identifiés chez trois espèces différentes (Figure 2.16). Supposons que les protéines codées par les gènes A de l'Homme et du canard (respectivement gènes 1A et 3A sur la Figure 2.16) exercent une fonction biologique α. Supposons qu'au contraire, les protéines codées par les gènes B du chat et du canard (respectivement gènes 2B et 3B) possèdent quant à eux une fonction biologique β différente. Dans cet exemple théorique, le couplage de ces informations fonctionnelles et des information portées par l'arbre de gènes permet ainsi d'inférer que : i) le gène 2A du chat exerce la fonction α, et ii) le gène 1B de l'Homme possède la fonction β. Ces hypothèses fonctionnelles pouvant ensuite être validées expérimentalement.

Cette théorie a depuis été utilisée dans de nombreuses études [255, 256] et a conduit à l'élaboration de plusieurs outils d'inférence fonctionnelle [257, 258, 259].

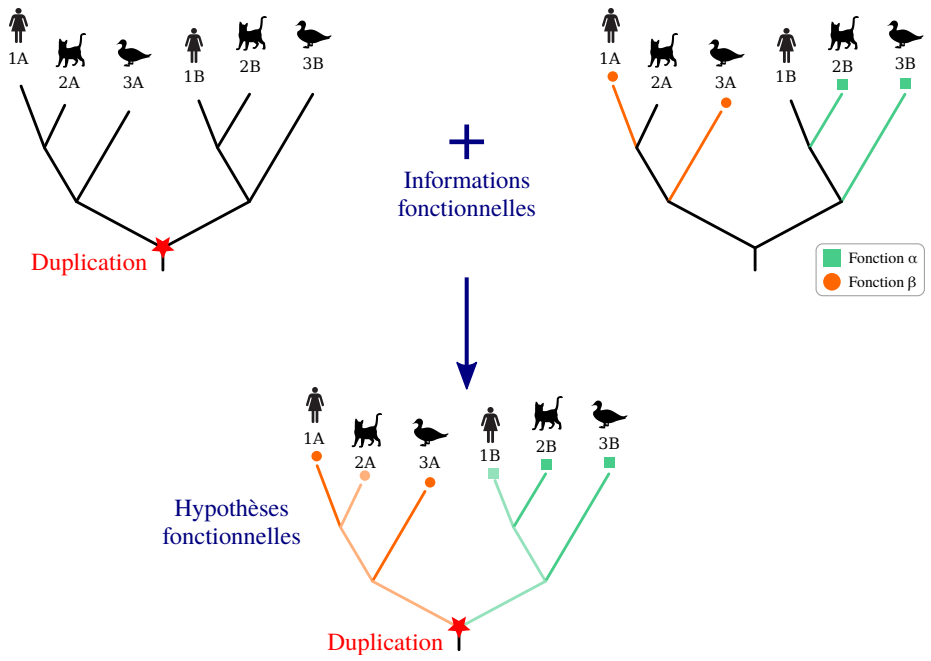

FIGURE *2.16* – ***Représentation théorique de l'apport du couplage de l'information phylogénétique et de l'information fonctionnelle.*** D'après [254].

2.4.2 SIDA et vaccins contre la poliomyélite

La phylogénie moléculaire constitue également un outil performant pour répondre à diverses questions biologiques. En virologie, elle peut en particulier permettre d'élucider les origines et/ou la transmission des virus au cours du temps entre différentes populations ou espèces [260, 261].

Dans ce domaine de recherche, une étude de Worobey *et al.* [262] s'est intéressée à l'origine de la transmission du virus du SIDA chez l'Homme dans la région de Kisangani, en République Démocratique du Congo. Cette analyse avait pour but de tester l'hypothèse de Hooper [263], selon laquelle la présence du VIH-1 (**V**irus de l'**I**mmunodéficience **H**umaine de type **1**) dans cette région était due à la campagne de vaccination massive contre la poliomyélite effectuée entre 1957 et 1960. En effet ces vaccins, développés dès les années 1950, étaient mis au point sur des cellules hépatiques de chimpanzés de l'Est du Congo. Or, comme l'illustre la Figure 2.17(a), deux sous-espèces de chimpanzés sont présentes en République Démocratique du Congo : *Pan troglodytes schweinfurthii* (ou *Pts*) qui est concentrée à l'Est, et *Pan troglodytes troglodytes* (ou *Ptt*) qui peuple l'Ouest de ce pays.

Figure omise pour des raisons de droits d'auteur.

(a)

Figure omise pour des raisons de droits d'auteur.

(b)

FIGURE *2.17* – ***(a) Répartition géographique des sous-espèces de chimpanzés en République Démocratique du Congo et (b) arbre phylogénétique des souches du SIDA.*** Issu de [262]. *HIV-1* : *Hhuman Immunodeficiency Virus type 1*, *SIVcpz* : *Ssimian Immunodeficiency Virus*, *Ptt* : *Pan troglodytes troglodytes*, *Pts* : *Pan troglodytes schweinfurthii.*

A partir du séquençage de virus simiens échantillonnés dans ces deux régions et des séquences du virus humain, Worobey *et al.* ont inféré un arbre phylogénétique par maximum de vraisemblance avec 1 000 réplicats de bootstrap. Comme le montre la topologie de cet arbre (Figure 2.17(b)), les souches de HIV-1 sont évolutivement plus proches des virus simiens de la sous-espèce *Ptt* que des virus des chimpanzés *Pts*. Il est donc plus probable que le virus du SIDA circulant chez les Hommes de la région de Kisangani soit issu d'une contamination depuis les chimpanzés *Ptt* plutôt qu'à cause du vaccin contre la poliomyélite, cultivé à partir des cellules hépatiques de *Pts*. A partir de ces résultats et des caractéristiques génétiques des différentes souches de SIDA, les auteurs ont ainsi pu invalider l'hypothèse de Hooper.

Chapitre 3

Sélection de transcrits alternatifs

3.1 Introduction

3.1.1 Transcrits alternatifs et phylogénie moléculaire

En phylogénie moléculaire, la qualité d'inférence des arbres dépend de la méthodologie employée, mais également de la qualité du jeu de données utilisé. Ainsi, après l'identification de l'ensemble des homologues du gène d'intérêt, une étape de sélection est primordiale. Parmi les séquences considérées « problématiques », on retrouve les séquences partielles mais également, dans le cadre d'études portant sur des protéines issues de gènes eucaryotes, les transcrits alternatifs. En effet, lors de la reconstruction phylogénétique d'un arbre de gène à partir de séquences protéiques, la présence de plusieurs protéines issues de la même séquence nucléique (ou isoformes) soulève plusieurs problèmes.

Tout d'abord, les transcrits alternatifs introduisent un premier biais lors de l'alignement multiple. En effet, si les transcrits alternatifs sont très similaires entre eux et ne diffèrent que par un exon (processus d'*exon skipping*), leur distance évolutive sera très faible. En conséquence, lors de l'alignement progressif ils seront groupés ensemble dès le début, introduisant un long gap au niveau de l'exon sauté. Or, dans leur conception, la plupart des logiciels d'alignement sont biaisés vers

la conservation des gaps introduits aux étapes précédentes (voir section 2.2). Par exemple, dans la Figure 3.1, le gap de cinq résidus introduit lors du groupement de la séquence T-a_1 avec la séquence T-a_2 dans la première étape de l'alignement, va être conservé lors de l'ajout progressif des autres séquences du jeux de données (région surlignée en bleu clair).

Ensuite, la présence de transcrits alternatifs dans le jeu de données induit un second biais au moment de la sélection des sites conservés. En effet, les logiciels de sélection tels que Gblocks ou BMGE, attribuent des poids aux sites selon leur degré de conservation entre les différentes séquences du jeu de données (voir section 2.2.3). Si à une position quelconque de l'alignement le même acide aminé est retrouvé dans plusieurs séquences, ces logiciels considéreront ce site comme très conservé et le sélectionneront. Néanmoins, s'il s'agit de transcrits alternatifs, la présence de cet acide aminé dans plusieurs séquences est artificielle puisqu'il s'agit en réalité du même résidu représenté plusieurs fois (voir position encadrée en jaune dans la Figure 3.1). Ce biais est d'autant plus marqué que les logiciels sélectionnent des régions conservées plus que des sites isolés et que les transcrits alternatifs sont généralement identiques sur plusieurs sites consécutifs (*i.e.* les régions correspondant aux exons).

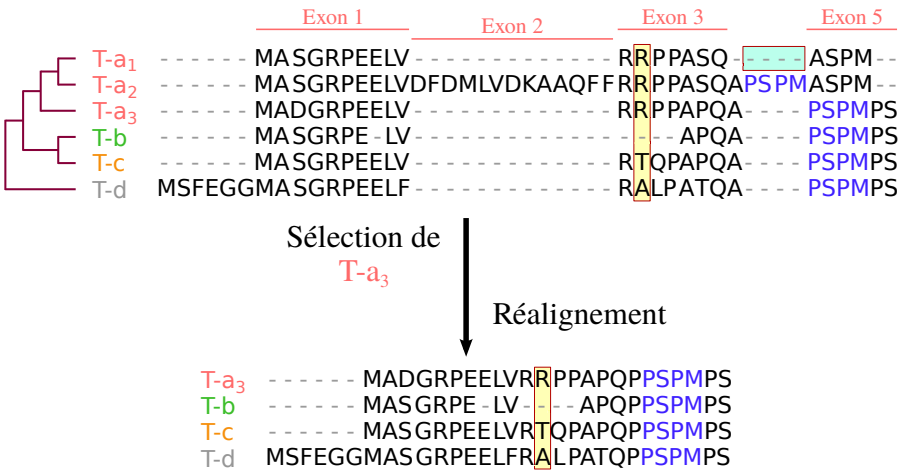

FIGURE *3.1* – ***Schématisation des biais introduits par la présence de transcrits alternatifs dans un jeu de données.*** Les T-x représentent des séquences issues d'un même gène x. En haut, l'alignement multiple avec les trois transcrits alternatifs du gène a. A gauche l'arbre représentant l'ordre d'insertion des séquences dans l'alignement. En bas, l'alignement obtenu en sélectionnant le transcrit T-a_3 pour le gène a.

Pour terminer, un arbre phylogénétique traduit les relations de parentés entre des séquences homologues, permettant ainsi d'inférer des événements de dupli-

cations et de spéciation. Conceptuellement, deux transcrits alternatifs ne correspondent pas à des séquences homologues, puisqu'elles ne sont pas issues d'un ancêtre commun et ne résultent d'aucun de ces deux événements.

La définition du meilleur transcrit alternatif va dépendre de l'objectif recherché. Dans le cadre d'une phylogénie moléculaire, on va chercher à sélectionner le transcrit minimisant les erreurs d'alignements, *i.e.* la séquence possédant le plus grand nombre de sites homologues aux sites des autres séquences du jeu de données. Dans la Figure 3.1, le transcrit T-a_1 n'est, par exemple, pas satisfaisant car contrairement au transcrit T-a_3 et aux autres séquences il ne possède pas la région `PSPM` (en bleu) en position C-terminale de l'alignement. Le transcrit T-a_2 comprend, quant à lui, un exon qui est apparemment spécifique du gène de l'espèce a et n'a donc pas de résidus homologues dans les autres séquences. Manuellement, un phylogénéticien choisira donc la séquence T-a_3. Dans cet exemple, la sélection du plus grand transcrit (T-a_2) produit l'insertion de nombreux gaps dans les autres séquences. Néanmoins, il est difficile de définir un critère de qualité permettant de déterminer automatiquement quel transcrit alternatif produira le moins d'erreur d'alignement une fois la sélection effectuée.

3.1.2 Méthodologies existantes

Peu de méthodes ont été développées pour la sélection automatique de transcrits alternatifs. La plupart des études phylogénétiques portant sur des gènes eucaryotes n'évoquent d'ailleurs pas cette sélection ni, par conséquent, la méthodologie employée. Le plus simple consiste à choisir un transcrit aléatoirement [264, 265] ou bien de sélectionner le plus long afin de maximiser l'information [266, 267]. Il est également possible de réaliser une sélection manuelle, comme j'ai pu le faire dans l'étude de la famille des PI3K.

La seule méthode actuelle dédiée à la sélection de transcrits alternatifs, baptisée PALO (***P****rotein* ***AL****ignment* ***O****ptimizer*), a été publiée en 2013 [268]. Un second programme, GUIDANCE, permet de filtrer les séquences et les positions problématiques d'un alignement (voir chapitre 2.2.3). Attribuant des scores aux séquences, il peut donc être utilisé pour la sélection du meilleur transcrit alternatif même s'il n'a pas été développé dans cet objectif.

PALO

Fondé sur la distribution des longueurs des séquences, PALO sélectionne l'iso-

forme ayant la taille la plus homogène aux autres séquences du jeu de données (Figure 3.2). Néanmoins, cet indicateur est inadapté lorsque les exons alternativement épissés sont de tailles similaires mais de séquences très différentes.

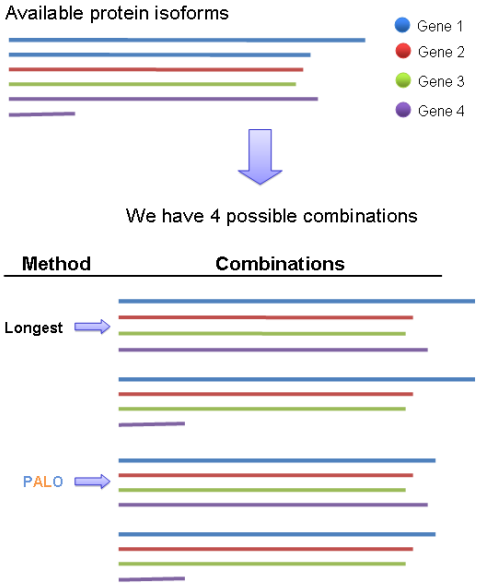

FIGURE 3.2 – ***Représentation schématique des sélections d'isoformes possibles pour un jeu de données composés de six séquences dont quatre transcrits alternatifs (tiré de [268]).***

Soit ℓ_{X_i} la longueur d'un des transcrits alternatifs du gène i, PALO choisit le transcrit minimisant la somme des carrés des écarts Q telle que :

$$Q = \sum_{m=1}^{m=k-1} \sum_{n=m+1}^{n=k} (\ell_{X_m} - \ell_{X_n})^2 \tag{3.1}$$

avec k le nombre de gènes considérés.

Pour tester l'efficacité de cette méthode, les auteurs l'ont comparée à : i) la sélection du plus long transcrit (méthode *longest*), ii) la sélection aléatoire d'un transcrit (méthode *random*) et iii) la sélection du transcrit maximisant la conservation des résidus de l'alignement (méthode *cons*). Cette dernière, nécessitant l'alignement de toutes les combinaisons possibles de transcrits alternatifs, est considérée par les auteurs comme produisant les meilleurs résultats. Cependant, cette approche présente l'important inconvénient d'une explosion du temps de calcul avec le nombre d'isoformes présentes dans le jeu de données. Concernant les données tests, les auteurs ont utilisé environ 21000 familles de gènes issues de la version 64 d'Ensembl.

Sur ces jeux de données, PALO sélectionne plus d'une fois sur deux une séquence différente du transcrit alternatif le plus long. Le transcrit retenu par PALO

est le même que celui désigné par la méthode *cons* dans 60-70% des cas, contre 16-21% pour les deux autres méthodes. Par ailleurs, les auteurs ont montré que PALO permet d'obtenir en moyenne des alignements multiples présentant moins d'insertions que les alignements obtenus avec les méthodes *longest* et *random*.

Du point de vue fonctionnel, PALO nécessite deux fichiers en entrée. Le premier contient les identifiants Ensembl des gènes homologues, chaque ligne correspondant à un jeu de données. Le second fichier est quant à lui composé de trois colonnes : la première contient les identifiants du gène, la seconde les identifiants des transcrits alternatifs et la dernière colonne la taille des séquences. PALO produit en sortie un fichier texte contenant les identifiants des séquences protéiques sélectionnées. PALO accédant directement aux annotations d'Ensembl au travers des identifiants, il n'est pas nécessaire de disposer de fichier contenant les séquences. Par contre, cette particularité rend ce programme inutilisable sur des données ne provenant pas de cette banque.

GUIDANCE

GUIDANCE est un algorithme développé pour évaluer la robustesse d'un alignement multiple. Le principe général de cet algorithme est d'engendrer des versions perturbées de l'alignement d'entrée et d'attribuer un score à chaque site selon s'il est affecté par la perturbation ou non.

A partir d'un ensemble de séquences non alignées, GUIDANCE génère l'alignement dit *de base* à l'aide de l'un des quatre programmes d'alignements multiples suivants : MAFFT, PRANK, CLUSTALW ou MUSCLE. Soit ℓ la taille de cet alignement, la seconde étape de l'algorithme consiste à inférer n arbres phylogénétiques par échantillonnage aléatoire avec remise de ℓ sites à partir de l'alignement de base. Cette étape de *bootstrap* permet d'inférer des arbres perturbés qui seront utilisés comme arbre guide pour le réalignement de l'alignement de base. Au total, n alignements dits *perturbés* sont ainsi inférés (Figure 3.3).

Une fois les alignements perturbés générés, GUIDANCE attribue un score de robustesse à chaque site de l'alignement de base. Trois distances sont communément utilisées dans le cadre de l'alignement de séquences :

- Le score par colonne CS (***C**olumn **S**core*) : si, à une position k de l'alignement perturbé, les résidus sont exactement les mêmes que dans l'alignement de base, un score de 1 est attribué à k (pas d'erreur). Un score de 0 est attribué à une colonne si au moins un résidu diffère (erreur).

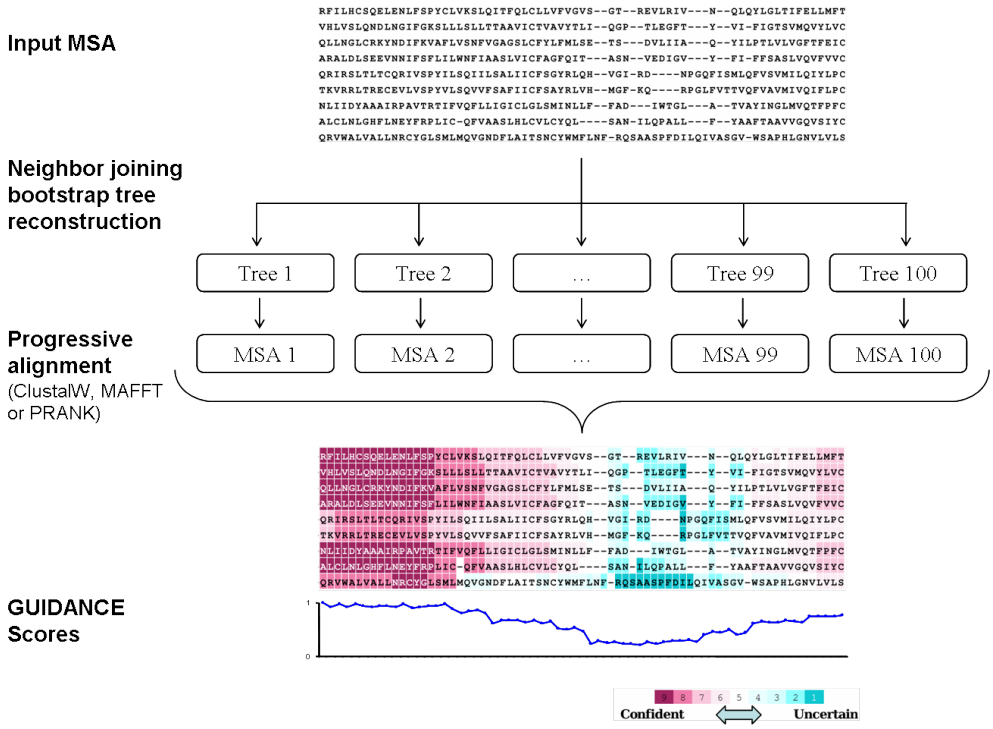

FIGURE *3.3* – ***Représentation schématique des étapes de l'algorithme de GUIDANCE [269].***

- Le score par paire de résidus SP (***S****um-of-****P****airs* score) : un score de 1 est attribué à chaque paire de résidu de l'alignement de base qui est retrouvée dans l'alignement perturbé (pas d'erreur). Si les résidus ne sont pas face à face, un score de 0 leur est assigné (erreur).
- Le score paire par colonne SPC (***S****um-of-****P****airs* ***C****olumn score*) : il s'agit de la moyenne des scores SP.

Le score CS ne distinguant pas les colonnes n'ayant qu'une erreur de celles contenant beaucoup d'erreurs, les auteurs de GUIDANCE ont décidé d'utiliser le score SPC moyenné sur l'ensemble des n alignements perturbés générés.

Le fichier d'entrée de GUIDANCE est un fichier au format Fasta contenant l'ensemble des séquences homologues étudiées. Étant développé pour évaluer la robustesse d'un alignement, GUIDANCE fournit de nombreuses informations en sortie telles que l'alignement de base, les scores par colonne, une sélection de sites conservés mais également les scores par séquences. Ces derniers reflètent la qualité de l'alignement de la séquence vis-à-vis du jeu de données et peuvent donc être utilisés afin d'identifier les séquences de mauvaise qualité. Dans la problématique

de sélection de transcrit alternatifs, il est donc possible de sélectionner le transcrit au score le plus élevé, *i.e.* le transcrit perturbant le moins l'alignement.

Contrairement à PALO, GUIDANCE utilise l'information portée par les séquences du jeu de données et est donc plus pertinent dans le cadre d'alignement de séquences et de reconstructions phylogénétiques. Néanmoins, si PALO présente l'avantage d'être rapide du fait d'un algorithme de faible complexité GUIDANCE est au contraire très lent [270], ceci même dans sa configuration la plus rapide (*i.e.* arbres inférés en NJ avec la matrice JTT et réalignements effectués avec MAFFT).

3.2 BATfinder

Dans ce contexte, j'ai décidé de développer BATfinder (***B**est **A**ligned **T**ranscript **finder***), un outil plus rapide et dédié à la sélection de transcrit alternatifs. Fondé sur le même principe que GUIDANCE, il permet, à partir d'un fichier d'entrée et d'un fichier contenant les identifiants des transcrits alternatifs, de fournir à l'utilisateur un jeu de données comportant une seule séquence par locus génomique. Implémenté en C/C++, BATfinder présente l'avantage d'être plus rapide que GUIDANCE notamment lorsque celui-ci est utilisé en mode multithread et que l'information sur l'identification des transcrits alternatifs lui est fournie.

BATfinder a fait l'objet d'un article soumis à *BMC bioinformatics* en Décembre 2015 (voir Annexe A).

3.2.1 Implémentation

Principe général

BATfinder est composé des trois mêmes grandes étapes que GUIDANCE à savoir : i) l'alignement dit de *référence* du jeu de données, ii) la génération d'alignements perturbés, et iii) le calcul des scores de SPC. Néanmoins, BATfinder propose un plus large choix de matrices de substitutions ainsi que l'utilisation de BioNJ au lieu de NJ pour l'inférence des arbres guides utilisés lors de la génération des alignements perturbés (Figure 3.4). De plus, BATfinder utilise un programme de calcul des matrices des distances très performant, nommé FastDist inclus dans la distribution du programme.

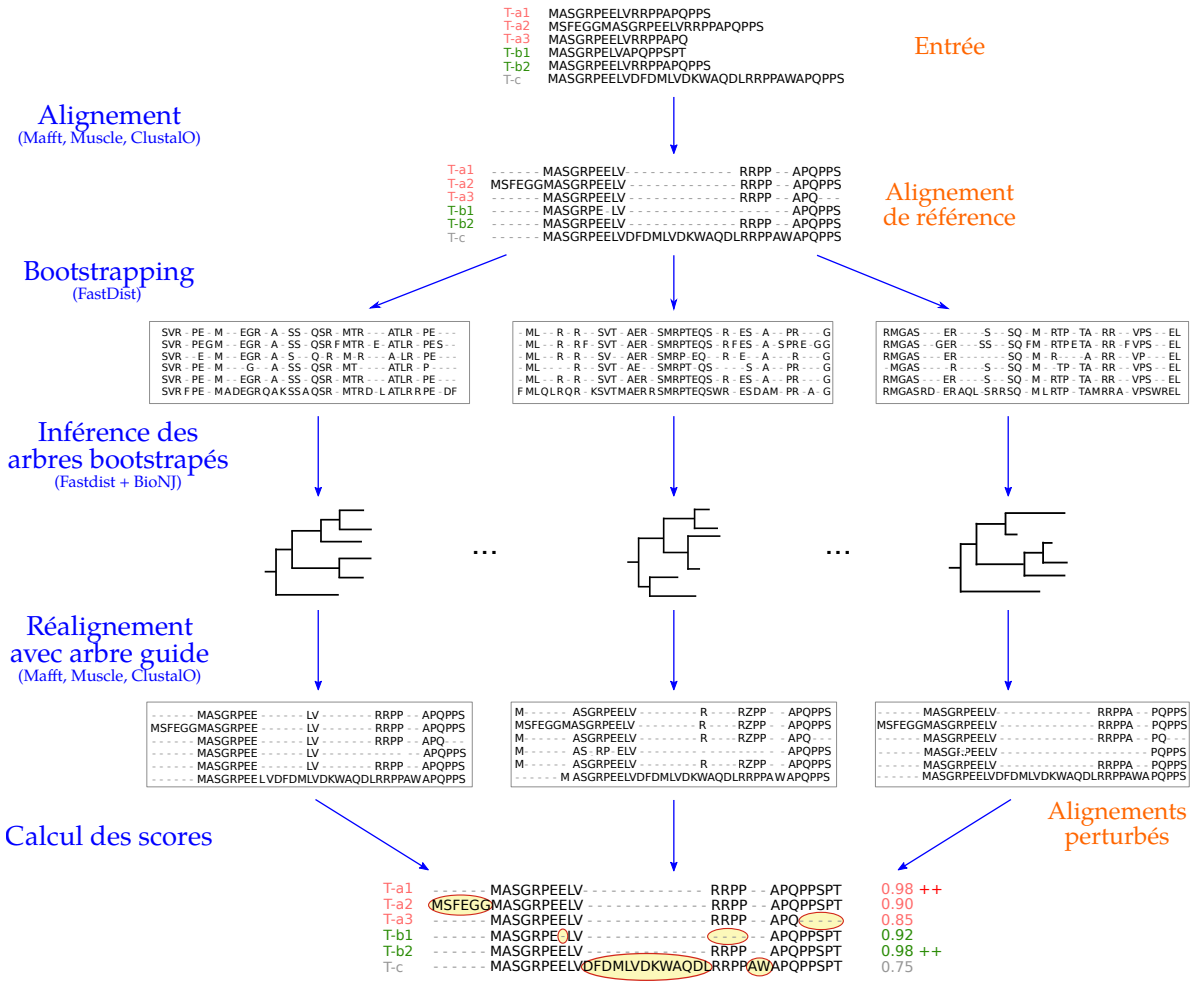

FIGURE *3.4* – ***Représentation schématique des étapes de BATfinder.***
Les programmes utilisés à chaque étape sont indiqués entre parenthèses.

Conceptuellement, BATfinder ne peut utiliser que des programmes d'alignement multiple permettant d'utiliser un arbre guide fourni en entrée (Figure 3.4). C'est pourquoi seuls CLUSTALO, MAFFT et MUSCLE ont été intégrés dans le programme. En effet, la comparaison des alignements perturbés et de l'alignement de base n'a de sens que si ces alignements ont été obtenus par le même algorithme. Cette contrainte explique également pourquoi BATfinder n'accepte pas de jeux de données préalignés en entrée.

FastDist

Une fois le fichier d'entrée aligné, l'échantillonnage aléatoire de sites de l'alignement de référence obtenu ainsi que le calcul des matrices des distances est effectué par le programme FastDist. Les modèles de substitutions pour les séquences pro-

téiques implémentés dans ce programme sont : Poisson, PAM (ou son approximation par Kimura), JTT, BLOSUM62, WAG et LG. Il est possible d'approximer les distances PAM et JTT en utilisant le modèle de Poisson couplé à une distribution Gamma. L'approximation de la distance PAM se fait en fixant $\alpha = 2.25$ tandis que l'approximation de la distance JTT se fait en fixant $\alpha = 2.4$ [271]. L'utilisation de ces approximations permet d'augmenter de façon notable la vitesse du programme. Enfin, l'utilisation des bibliothèques C/C++ Eigen [272] et OpenMP [273] permettent à FastDist d'effectuer des calculs parallèles très performants.

Calcul des scores

Considérons un jeu de données composé de m séquences et a l'alignement de référence composé de ℓ sites, obtenu à la première étape. Soit b ($1 \leq b \leq n$) l'un des n alignement perturbés générés à partir de a et l_i la taille de la séquence i ($1 \leq i \leq m$). Le score du $k^{\text{ème}}$ résidu ($1 \leq k \leq l_i$) de la séquence i est alors défini par :

$$R_{ik}^{(b)} = \frac{1}{m_k - 1} \sum_{j=1, j \neq i}^{m_k} p_{ijk} \tag{3.2}$$

où p_{ijk} vaut 1 ou 0 selon que le résidu à la position k de la séquence i est aligné avec le même résidu k' ($1 \leq k' \leq l_j$) de la séquence j dans l'alignement de référence a et dans l'alignement perturbé b. Le terme m_k correspond quant à lui au nombre de séquences possédant un résidu (*i.e.* donc pas de gap) à cette position dans a.

L'équation (3.2) permet de calculer les scores par résidu pour un seul alignement perturbé. Le score par résidu final est ensuite calculé en moyennant les scores obtenus pour chacun des n réplicats de *bootstraps* générés :

$$R_{ik} = \frac{1}{n} \sum_{b=1}^{n} R_{ik}^{(b)} \tag{3.3}$$

Enfin, le score de la séquence i est obtenu en moyennant les scores de chacun de ses résidus :

$$S_i = \frac{1}{l_i} \sum_{k=1}^{l_i} R_{ik} \tag{3.4}$$

Plus S_i est proche de 1, meilleur est l'alignement relatif de la séquence i au sein de a. Un exemple de calcul de ces différents scores est illustré dans la Figure 3.5.

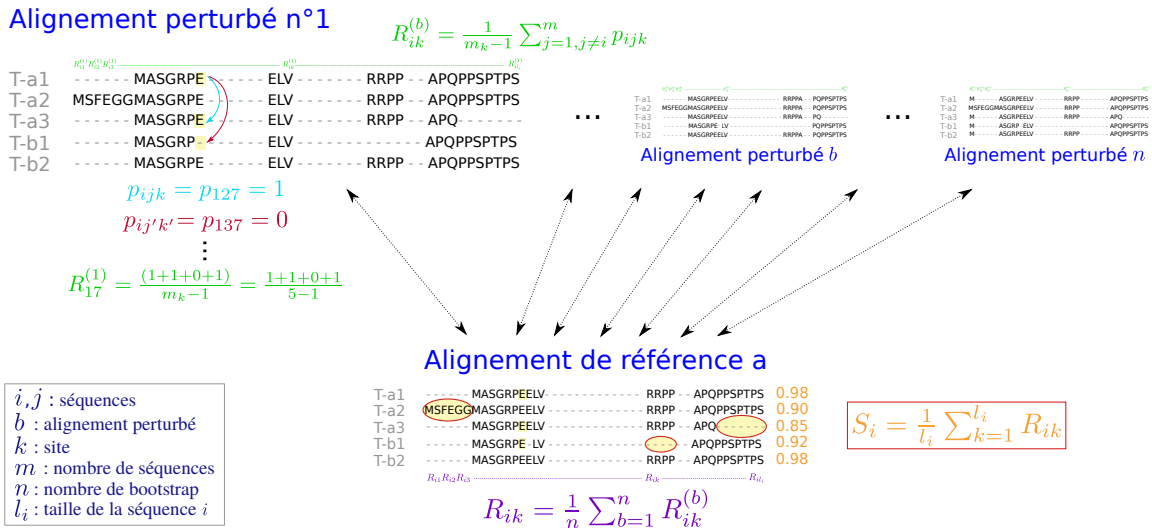

FIGURE *3.5* – ***Calcul des scores pour deux gènes possédant deux et trois transcrits alternatifs.***

Pénalisation des scores

L'un des points négatifs du score S_i précédemment défini est qu'il ne prend pas en compte la taille relative de la séquence i par rapport aux tailles des autres séquences. Ainsi, une séquence composée de résidus conservés situés dans une région conservée de l'alignement aura un très bon score S_i même si elle n'est composée de quelques résidus alors que la taille moyenne des séquences est de plusieurs centaines de résidus. Pour pallier ce problème, j'ai implémenté une option (`-short`) sous laquelle le score de la séquence n'est pas moyenné sur sa taille l_i mais sur la taille de l'alignement a :

$$S_i = \frac{1}{\ell} \sum_{r=1}^{l_i} R_{ik} \tag{3.5}$$

Les scores S_i obtenus seront de ce fait toujours compris entre 0 et 1 mais beaucoup plus faibles que lorsqu'ils sont calculés à l'aide de l'équation (3.4).Comme précédemment, le transcrit alternatif sélectionné sera celui possédant le score de plus élevé.

Un second problème de ce système de score est qu'il ne prend pas en compte les gaps situés dans les colonnes de l'alignement de référence. En effet, si à la position h ($1 \leq h \leq \ell$) de a, seules les séquences i et j possèdent un résidu tandis que les $m - 2$ séquences restantes présentent un gap à cette position, le poids des scores associés sera le même que si toutes les séquences possédaient un résidu homologue à cette position. Prenons l'exemple d'un jeu de donné dans lequel i correspond à un

transcrit alternatif d'un gène humain et j à un transcrit alternatif de son homologue chez le Chimpanzé. Si la position h se situe dans un exon cassette uniquement présent dans ces deux espèces, il est fort probable que dans les alignements perturbés b ces deux régions soient bien alignées. En effet, si l'exon n'est présent dans aucune autre séquence du jeu de données, l'alignement ressemblera à la situation de la Figure 3.1 où le transcrit T-a$_2$ est le seul à posséder l'exon 2, impliquant l'insertion de grands gaps dans les autres séquences. Ainsi, dans cette configuration, l'ensemble des résidus de cet exon cassette conservé entre l'Homme et le Chimpanzé auront de très bon scores $R_{ik}^{(b)}$, augmentant de fait les scores S_i de ces deux séquences. Afin de ne pas favoriser la sélection de ces transcrits, j'ai donc implémenté l'option `-gap` qui calcule ces scores selon la formule :

$$R_{ik}^{(b)} = \frac{1}{m_k - 1} \sum_{j=1, j \neq i}^{m} \frac{p_{ijk}}{m_g} \tag{3.6}$$

où $m_g = m - m_k$ est le nombre de gaps à cette position dans l'alignement de référence. Si aucune séquence n'a de gap à cette position, alors $m_g = 1$. De plus, p_{ijk} est calculé comme précédemment (équation 3.2) excepté qu'il est fixé à -1 si le résidu n'a aucun homologue dans le jeu de données (cas du transcrit T-a$_2$ de la Figure 3.1).

De même que pour l'option `-short`, les scores S_i obtenus avec l'option `-gap` sont par construction, plus faibles. Néanmoins, seule la comparaison des valeurs relatives des scores S_i est utilisée pour sélectionner le transcrit le mieux aligné.

Option `DS`

La phase de BATfinder qui nécessite le plus de temps est celle correspondant au *bootstrap*. Le temps d'exécution de BATfinder augmente également fortement avec le nombre de séquences du jeu de données et de leurs tailles. Afin de permettre une sélection de transcrit beaucoup plus rapide pour les gros jeux de données, j'ai implémenté une option baptisée `-DS` (***D****istance* ***S****core*). Sous cette option, la méthodologie de BATfinder est totalement différente. En effet, aucun alignement perturbé n'est nécessaire, aucun réplicat de *bootstrap* n'est effectué et le critère de sélection n'est pas le score SPC.

Dans cette utilisation de BATfinder, on considère la p-distance moyenne entre le transcrit alternatif et les autres séquences du jeu de données qui ne sont pas des transcrits alternatifs. L'utilisation combinée de l'option `-WOT` (***W****ith* ***O****ther* ***T****ranscripts*) permet de calculer cette p-distance moyenne en considérant l'en-

semble des séquences du jeu de données d'entrée. Néanmoins, pour prendre en compte les différences de tailles entres les transcrits, j'ai modifié le calcul classique de la p-distance. En effet, au lieu de ne considérer que les sites où deux séquences présentent un résidu, sont également pris en compte les sites où l'une des deux séquences présente un gap. Dans cette approche, un gap est donc considéré comme un état de caractère supplémentaire. Le score S_i de la séquence i est ensuite calculé en moyennant l'ensemble des distances observées entre i et les autres séquences, soit :

$$S_i = \frac{1}{m} \sum_{j=1}^{m} p_{ij} \tag{3.7}$$

avec m le nombre total de séquences si l'option `-WOT` est utilisée. Dans le cas contraire, m correspond au nombre de séquences qui ne sont pas des transcrits alternatifs. p_{ij} correspond à la p-distance « modifiée » entre les séquences i et j.

Le transcrit sélectionné correspond au transcrit le plus similaire aux autres séquences, c'est-à-dire le transcrit ayant la plus petite p-distance moyenne, donc le score S_i le plus faible.

3.2.2 Utilisation

Fichiers d'entrée et de sortie

Le seul fichier essentiel au fonctionnement de BATfinder est un fichier Fasta contenant des séquences homologues non alignées. Dans cette situation basique d'utilisation, un score sera calculé pour chacune des séquences d'entrée et sauvegardé dans un fichier de sortie. Par ailleurs, un fichier de log ainsi que celui de l'alignement de référence sont générés. Pour sélectionner un transcrit par gène homologue et fournir en plus un fichier de sortie contenant le jeu de données filtré, il est nécessaire d'indiquer à BATfinder quelles sont les séquences issues d'un même gène (Figure 3.6). L'option `-f` de BATfinder permet de spécifier le nom d'un fichier dans lequel le lien entre les noms des séquences et leur provenance génomique est indiqué.

Deux fichiers d'exemple sont inclus dans la distribution de BATfinder. Le fichier `example.fasta` contient 59 séquences eucaryotes homologues à la protéine humaine AKTS1 identifiées à l'aide d'une recherche BLASTp sur deux banques de données locales (voir chapitre 5.2.2). Le second fichier, `transcript_file.txt`, contient l'information sur la provenance des séquences protéiques. Chaque ligne de ce fichier correspond à un transcrit alternatif, la première colonne contenant les

noms des séquences présentes dans `example.fasta` et la seconde les identifiants Ensembl des gènes correspondants. Par exemple, les trois lignes ci-dessous :

```
>ENSP00000375710|Homo_sapiens ENSG00000204673
>ENSP00000375711|Homo_sapiens ENSG00000204673
>ENSP00000375706|Homo_sapiens ENSG00000204673
```

permettent de spécifier que les séquences `ENSP00000375710`, `ENSP00000375711` et `ENSP00000375706` correspondent à trois transcrits alternatifs codés par le gène `ENSG00000204673`.

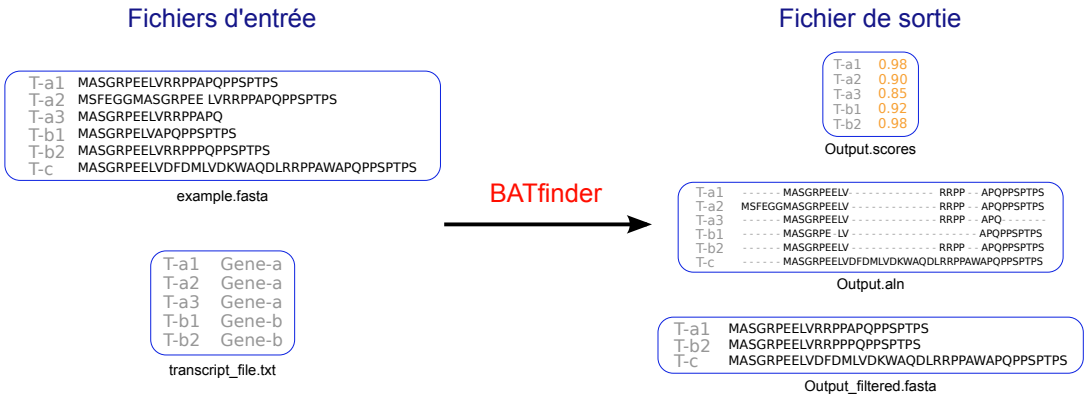

FIGURE *3.6* – ***Fichiers d'entrée et de sortie de BATfinder.***

Disponibilité

Toutes les informations concernant l'installation et l'utilisation de BATfinder ainsi que les fichiers associés sont téléchargeables à l'adresse `http://doua.prabi.fr/software/batfinder`. Les binaires de BATfinder et de FastDist distribués sur le site sont les versions mono- et multi-thread pour MacOSX et Linux. Dans le cas des versions multi-thread, celles-ci nécessitent que la bibliothèque OpenMP soit disponible sur la machine d'installation. Les sources C/C++ sont également disponibles pour une recompilation éventuelle.

Par ailleurs, BATfinder fait appel aux logiciels CLUSTALO v1.2.0, MAFFT v7.266, MUSCLE v3.6, SeaView v4.4.1 [274] et BioNJ. Les binaires des quatre premiers programmes figurent également dans les distributions de BATfinder ainsi que le source C de BioNJ. Tous ces programmes sont distribués avec l'autorisation de leurs auteurs respectifs.

3.2.3 Performances

Jeux de données tests

Pour tester BATfinder j'ai sélectionné 45 protéines humaines impliquées dans

la régulation de l'autophagie. J'ai identifié les homologues eucaryotes de chacune d'entre elles selon le protocole décrit dans la section 4.2.1. Pour les jeux de données correspondants aux protéines GSK3B, MTOR, PIK3C2A et PK3CA il n'a pas été possible de calculer les scores car les alignements avec MAFFT n'ont pas aboutis sur la machine virtuelle utilisée pour effectuer les tests. Les 41 jeux de données restants comportent de 42 à 1484 séquences parmi lesquelles les transcrits alternatifs représentent de 2.6% à 33.5% du total. Les alignements multiples obtenus avec MAFFT avant sélection des transcrits alternatifs contiennent de 280 à 11175 sites (Table 3.1).

Nom de la protéine	**Identifiant UniProtKB**	**Taille (aa)**	**Nombre d'homologues**	**Taille de l'alignement**	**Nombre de gènes**	**Pourcentage de transcrits alternatifs**
AKTS1	Q96B36	256	42	403	33	21.43
MCL1	Q07820	350	66	648	55	16.67
BIRC5	O15392	142	80	280	62	22.50
BAKOR	Q6ZNE5	492	81	1078	77	4.94
PIK3R5	Q8WYR1	880	90	1142	72	20.00
RBCC1	Q8TDY2	1594	100	2661	78	22.00
GOPC	Q9HD26	462	102	2081	79	22.55
UVRAG	Q9P2Y5	699	103	1943	92	10.68
SIN1	Q9BPZ7	522	111	846	77	30.63
TSC1	Q92574	1164	114	2130	88	22.81
ATG13	O75143	517	114	812	77	32.46
AMRA1	Q9C0C7	1298	115	3177	89	22.61
DPTOR	Q8TB45	409	125	1912	103	17.60
PATL1	Q86TB9	770	134	1895	115	14.18
SH3B4	Q9P0V3	963	137	1234	127	7.30
TSC2	P49815	1807	143	6126	109	23.78
PRR5	P85299	388	151	1171	123	18.54
RICTR	Q6R327	1708	158	6949	128	18.99
PIK3R4	Q99570	1358	168	6869	157	6.55
LST8	Q9BVC4	326	177	2165	163	7.91
BCL2	P10415	239	181	494	158	12.71
IRS1	P35568	1242	186	3660	159	14.52
MDM2	Q00987	491	187	1238	128	31.55
B2CL1	Q07817	233	197	621	162	17.77
RRAGA	Q7L523	313	206	1613	187	9.22
BECN1	Q14457	450	206	3359	186	9.71
RPTOR	Q8N122	1335	221	5977	203	8.14
SHLB1	Q9Y371	365	236	823	162	31.36
RRAGC	Q9HB90	399	247	3713	222	10.12
RHEB	Q15382	184	267	1067	260	2.62
FOXO3	O43524	673	290	2815	244	15.86
PIK3R1	P27986	724	299	2521	229	23.41
RUBIC	Q92622	972	313	4295	260	16.93
P53	P04637	393	322	1529	214	33.54
RS6	P62753	249	355	1256	340	4.23
TFEB	P19484	476	379	2360	262	30.87
PTEN	P60484	403	497	8188	420	15.49
HMGB1	P09429	215	523	1945	487	6.88
KAT5	Q92993	513	757	11175	627	17.17
1433G	P61981	247	1064	4348	994	6.58
CCR2	P41597	374	1484	1825	1378	7.14

TABLE *3.1* – ***Propriétés des 41 jeux de données utilisés pour tester BATfinder.***

Qualité des sélections

Je me suis intéressée à plusieurs critères phylogénétiques pour évaluer les sélections effectuées par les différentes options de BATfinder : i) la taille de l'alignement du fichier filtré, ii) le nombre de sites sélectionnés parmi les sites cet alignement, iii) la longueur totale de l'arbre phylogénétique inféré avec et sans sélection de sites.

Une fois la sélection de transcrits effectuée, j'ai aligné les jeux de données obtenus avec le logiciel MAFFT en mode automatique (option `-auto`) et avec 10 itérations maximum (`-maxiterate 10`). MAFFT produisant des alignements « éclatés », les alignements les plus longs ont donc beaucoup de sites isolés (*i.e.* entourés de gaps). En première approximation, on peut donc penser que plus les alignements obtenus seront longs, plus ils seront de mauvaise qualité pour une reconstruction phylogénétique. Comme l'illustre l'exemple de la Figure 3.7, l'alignement du jeu de données correspondant à la protéine FOXO3 est beaucoup plus éclaté lorsque le transcrit le plus long est sélectionné :

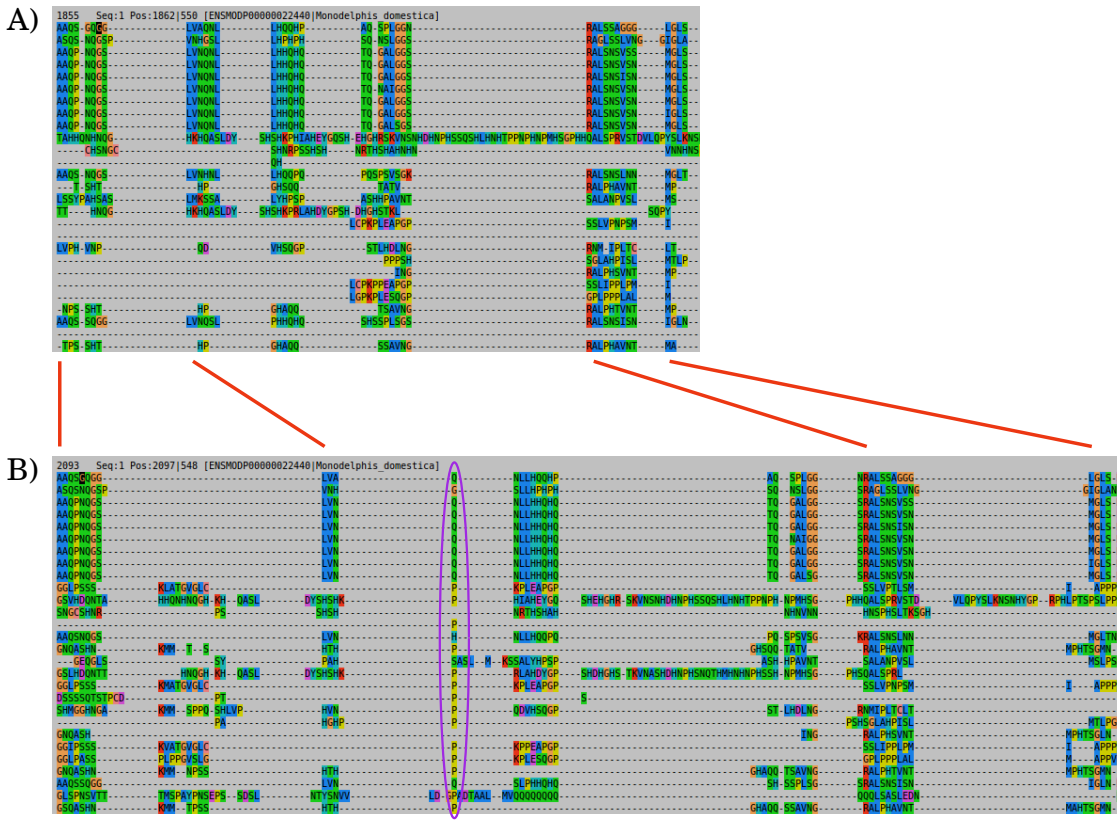

FIGURE *3.7* – ***Fragment de l'alignement des homologues de FOXO3 après : A) l'utilisation de BATfinder (paramètres par défaut) et B) la sélection du transcrit le plus long.*** Un exemple de site isolé est entouré en violet.

En moyenne, les alignements produits avec les différentes options de BATfinder sont plus courts que ceux obtenus en sélectionnant le transcrit alternatif le plus long, excepté avec les options par défaut. Un test de Wilcoxon apparié sur les longueurs d'alignements a été effectué pour chaque option par rapport à la sélection du plus grand transcrit. En utilisant le seuil $\alpha = 0.05$, seuls les résultats avec les options `-gap` et la combinaison `-gap -short` sont significatifs.

Néanmoins, cette mesure ne reste qu'une approximation de la qualité des alignements. Je me suis donc intéressée au nombre de sites conservés lorsque l'on utilise le programme de filtrage BMGE. L'hypothèse qui est faite est que la sélection du transcrit le plus semblable aux autres homologues du jeu de données doit conduire à un nombre plus important de sites conservés. Ainsi, moins de sites seront conservés si l'alignement fourni en entrée comprend de nombreux sites isolés (*e.g.* le site entouré en violet dans l'alignement B de la Figure 3.7). J'ai donc tout d'abord vérifié qu'il n'existait pas de corrélation entre la taille de l'alignement et le nombre de sites sélectionnés par BMGE (Figure A.1). Au seuil $\alpha = 0.05$, aucune corrélation significative n'est observée.

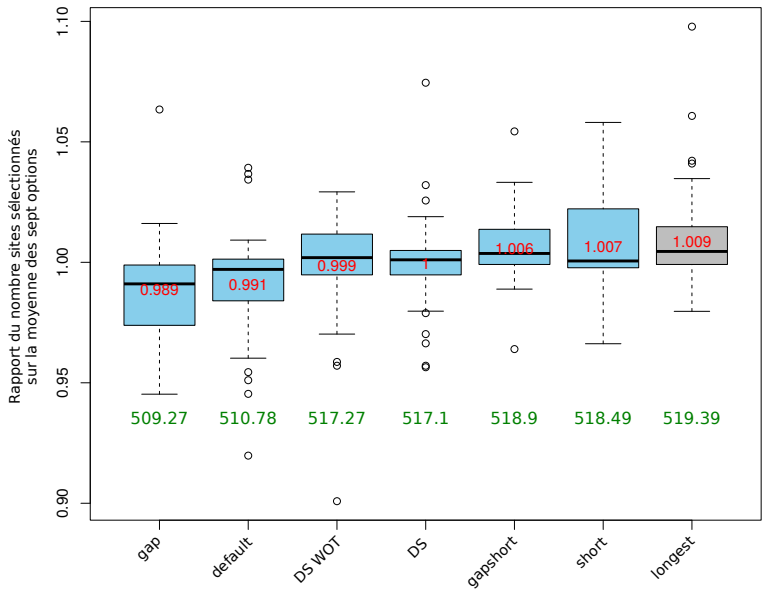

FIGURE *3.8* – ***Distribution du nombre de sites conservés par BMGE en fonction des différentes options de BATfinder.*** Les moyennes brutes (en aa) et normalisées sont respectivement indiquées en vert et rouge.

Les paramètres utilisés pour BMGE étaient : la matrice BLOSUM30, des blocs de longueur minimale de trois acides aminés et une proportion de gaps autorisée de 40%. Le choix de ces paramètres a été guidé par les résultats obtenus sur les PI3K

(voir chapitre 4). Les nombres de positions sélectionnées par BMGE sont présentés dans la Figure 3.8. Les alignements étant de tailles très variables (Table 3.1), le nombre de sites conservés pour chaque jeu de données est divisé par le nombre moyen de sites conservés, ceci pour les différentes options utilisées (transcrit le plus long et les six options de BATfinder). Pour une protéine et une option donnée, si ce rapport est inférieur à un, alors l'option en question a conduit à moins de sites sélectionnés que les autres options testées.

En moyenne, la sélection du transcrit le plus long conduit à un plus grand nombre de sites sélectionnés avec BMGE (519.39 sites sélectionnés en moyenne contre d'environ 509 à 518 avec BATfinder). Au contraire, l'option `-gap` de BATfinder produit des alignements filtrés plus courts (environ 509 sites en moyenne). Néanmoins, seules les différences observées avec les options par défaut et l'option `-gap` sont significatives ($P < 0.05$ selon un test de Wilcoxon apparié sur les données brutes). Ainsi, l'utilisation de BATfinder ne permet pas d'augmenter le nombre de sites conservés sélectionnés pour l'inférence des arbres phylogénétiques par rapport à la sélection du transcrit le plus long.

Dans un arbre phylogénétique, les séquences évoluant rapidement ou étant mal alignées sont souvent à l'origine de longues branches, entraînant de nombreux artefacts de reconstruction [275, 276]. Ainsi, dans le cadre de la sélection de transcrits alternatifs, on peut supposer que la longueur d'un arbre (*i.e.* la somme de ses longueurs de branches) est une approximation de la qualité du jeu de données. En effet, comme mathématiquement démontré par Rzhetsky et Ney en 1993 [242], l'arbre vrai est l'arbre le plus court parmi l'ensemble des arbres possibles. Ainsi, les transcrits alternatifs permettant d'obtenir un arbre plus court doivent être favorisés.

Pour chaque jeu de données j'ai donc inféré l'arbre correspondant, tout d'abord sans aucune sélection de sites conservés et ensuite avec une sélection effectuée par BMGE (mêmes options que précédemment). L'ensemble de ces arbres a été inféré en utilisant BioNJ et la p-distance. En effet, pour que ces longueurs d'arbres reflètent le plus précisément possible les similitudes entre séquences, je n'ai pas utilisé de modèle d'évolution. Par ailleurs, la longueur de l'arbre étant corrélée au nombre de branches (et donc de séquences), une normalisation par la moyenne des longueurs de branches est nécessaire pour une visualisation des résultats plus pertinente. Ainsi, de manière analogue à l'étude du nombre de sites sélectionnés par BMGE, j'ai normalisé la longueur d'un arbre par la moyenne des longueurs des arbres obtenus avec les autres options, ceci pour chacun des 41 jeux de données.

La Figure 3.9 montre les résultats obtenus avec et sans sélection des sites conservés. Dans les deux cas, l'utilisation de BATfinder conduit en moyenne à l'inférence d'arbres plus courts. Sans sélection, ces différences sont significatives pour toutes les options de BATfinder, excepté avec l'option `-short` ($P = 0.27$ avec un test de Wilcoxon apparié sur les données brutes). Les meilleurs résultats sont obtenus avec l'utilisation des options par défaut. Avec sélection de sites, seules les options par défaut et l'utilisation de l'option `-gap` donnent des arbres significativement plus courts que la sélection du transcrit le plus long.

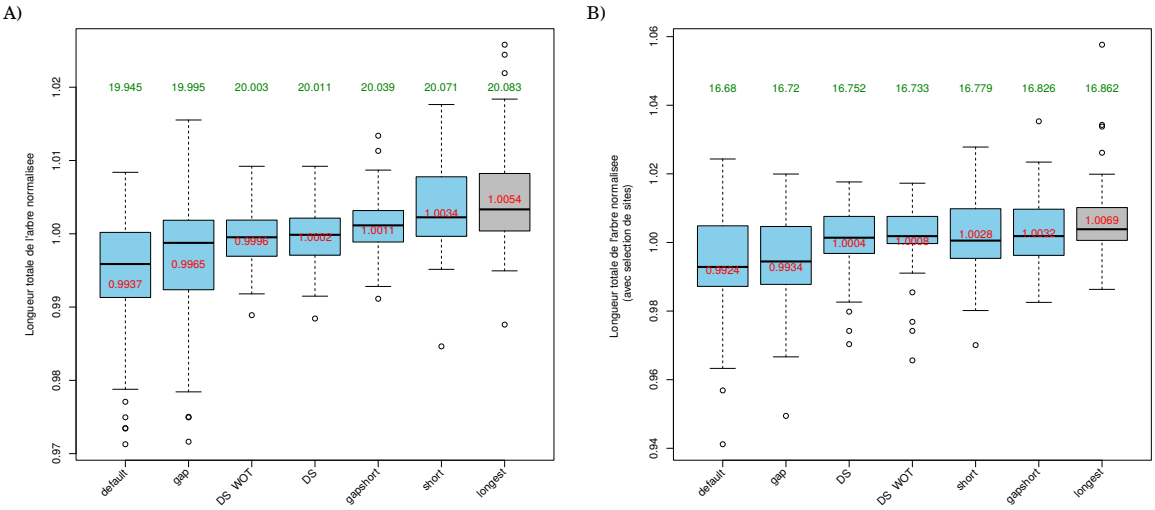

FIGURE *3.9* – ***Distribution des longueurs d'arbres en fonction des différentes options de BATfinder.*** (A) Aucune sélection de sites n'est effectuée. (B) Avec sélection de sites par BMGE. Les moyennes brutes et normalisées sont indiquées respectivement en vert et rouge.

Sans l'utilisation de BMGE, la sélection du plus long transcrit a donné l'arbre phylogénétique le plus court dans uniquement un cas sur 41. BATfinder permet donc d'obtenir le meilleur résultat dans les 40 cas restants. Cependant, les options permettant à BATfinder d'obtenir le meilleur résultat dépendent du jeu de données utilisé : options par défaut (meilleur dans 19 cas), `-gap` (9), `-gap -short` (4), `-DS -WOT` (4), `-DS` (4) et `-short` (1).

Avec l'utilisation de BMGE, la sélection du plus long transcrit a donné l'arbre phylogénétique le plus court dans trois cas sur 41. BATfinder permet donc d'obtenir le meilleur résultat dans les 38 cas restants. De la même façon que précédemment, les options permettant à BATfinder d'obtenir le meilleur résultat dépendent du jeu de données utilisé : options par défaut (meilleur dans 14 cas), `-gap` (10), `-gap -short` (4), `-DS -WOT` (3), `-short` (5) et `-DS` (2).

Vitesse d'exécution

Enfin, j'ai testé la rapidité de BATfinder par rapport à GUIDANCE. Les tests ont été effectués sur une machine virtuelle Linux CentOS de huit CPU, cadencée à 2.6 GHz et possédant 16 Go de mémoire vive. J'ai utilisé le programme d'alignement MAFFT et 20 réplicats de *bootstrap*. Comme le montre la Figure 3.10, BATfinder devient plus rapide que GUIDANCE lorsque le nombre de CPU utilisées augmente. Par exemple, le tri des transcrits alternatifs du jeu de données de la protéine ATG13 est effectué plus rapidement par GUIDANCE lorsqu'un seul ou deux CPU sont utilisées. En revanche, avec huit CPU, BATfinder devient deux fois plus rapide sur ce jeux de données. De même, le tri des transcrits alternatifs de SH3B4 est de 3.6 (une seule CPU) à 13.3 (huit CPU) fois plus rapide avec BATfinder qu'avec GUIDANCE.

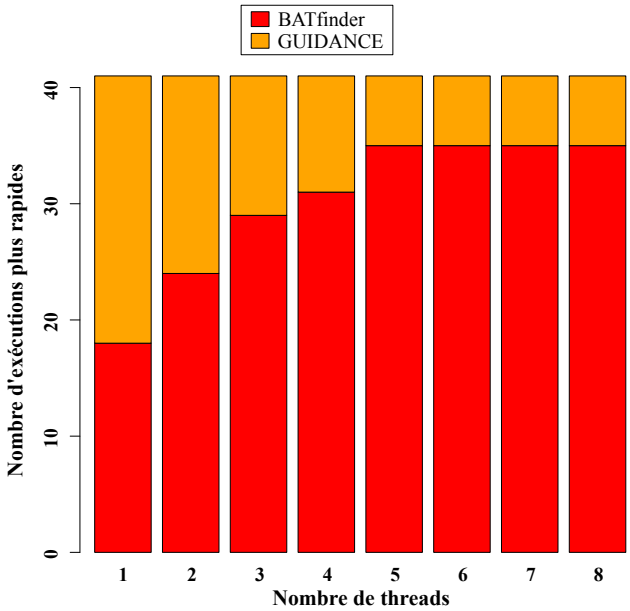

FIGURE *3.10* – ***Performance relative de BATfinder et GUIDANCE en fonction du nombre de threads utilisées.***

Ces différences en termes de performances s'expliquent en partie par le fait que GUIDANCE fournit de nombreuses informations supplémentaires par rapport à BATfinder et donc ce programme doit effectuer un certain nombre de calculs supplémentaires.

3.3 Conclusions

Fondé sur le même principe que GUIDANCE, BATfinder est un logiciel qui vise à sélectionner le transcrit alternatif qui soit le plus semblable aux autres séquences d'un jeux de données d'entrée. Plus rapide que GUIDANCE, notamment en mode parallèle, BATfinder produit directement un fichier contenant une séquence par locus génomique. Par ailleurs, BATfinder autorise l'utilisation d'un large panel de modèles évolutifs pour le calcul des distances évolutives entre les séquences tandis que GUIDANCE n'utilise que le modèle JTT.

En l'absence de critère de qualité intrinsèque pour les arbres phylogénétiques j'ai utilisé diverses mesures fondées sur certaines caractéristiques des alignements multiples. Tout d'abord, l'utilisation de BATfinder permet d'obtenir en moyenne des alignements plus courts que lorsque le transcrit alternatif le plus long est sélectionné, excepté avec les options par défaut. Les options `-gap` et `-gap -short` étant les seules à produire des résultats significatifs. Au niveau du nombre de sites sélectionnés par BMGE, le choix systématique du transcrit le plus long conduit à plus de sites gardés, bien que les alignements multiples obtenus soient plus longs et donc, en première approximation, de plus mauvaise qualité (comportant plus de sites isolés). Enfin, en ce qui concerne la longueur des arbres phylogénétiques inférés, qui semble être le meilleur critère de comparaison, l'ensemble des options de BATfinder permettent d'obtenir des arbres plus courts par rapport à la sélection automatique du plus long transcrit, justifiant son utilisation. Selon ce critère, ce sont les options par défaut et l'option `-gap` qui produisent les meilleurs résultats.

Pour terminer, l'option `-SD` permet d'obtenir de relativement bon résultats bien qu'elle implique une méthodologie très différente des autres options de BATfinder. En effet, aucune phase de *bootstrap* n'est effectuée avec cette option, ce qui permet de réduire drastiquement le temps d'exécution. Cette option est donc à privilégier pour les gros jeux de données.

Chapitre 4

Histoire évolutive des PI3K

4.1 Présentation des PI3K

Les Phosphatidylinositol-3-kinases (PI3K) sont des enzymes cytoplasmiques phosphorylant les inositols en position 3', générant ainsi différents types de phosphatidylinositols. Découvertes au début des années 1980, leur composition en modules fonctionnels ainsi que leurs structures tridimensionnelles sont connues depuis seulement une quinzaine d'années [277, 278]. A ce jour, on dénombre 14 protéines de la famille des PI3K chez l'Homme. Les PI3K sont impliquées dans de nombreuses fonctions biologiques telles que l'autophagie [279], l'angiogenèse [280], la motilité [281, 282, 283] ou encore la survie cellulaire [284, 285]. Situées principalement dans le cytoplasme de la cellule, elles sont les premières protéines activées par signaux extracellulaires lors de l'initiation de ces différents processus.

4.1.1 Une famille divisée

En fonction du type de phosphoinositides (PI) qu'elles phosphorylent, les PI3K ont été catégorisées en trois classes principales [286, 287] (Figures 4.1 et 4.2). Ainsi, la classe I des PI3K humaines transforment les phosphatidylinositols-4,5-bisphosphate (PI(4,5)P_2) en phosphatidylinositols-3,4,5-triphosphate (PI(3,4,5)P_3) tandis que la classe III permet la synthèse de phosphatidylinositols-3-phosphate

(PI(3)P) à partir de phosphoinositides. Enfin, on considère généralement que le substrat préférentiel de la classe II est le phosphatidylinositol-4-phosphate (PI(4,5)P_2), bien que les résultats varient selon les études *in vivo* et *in vitro* [288].

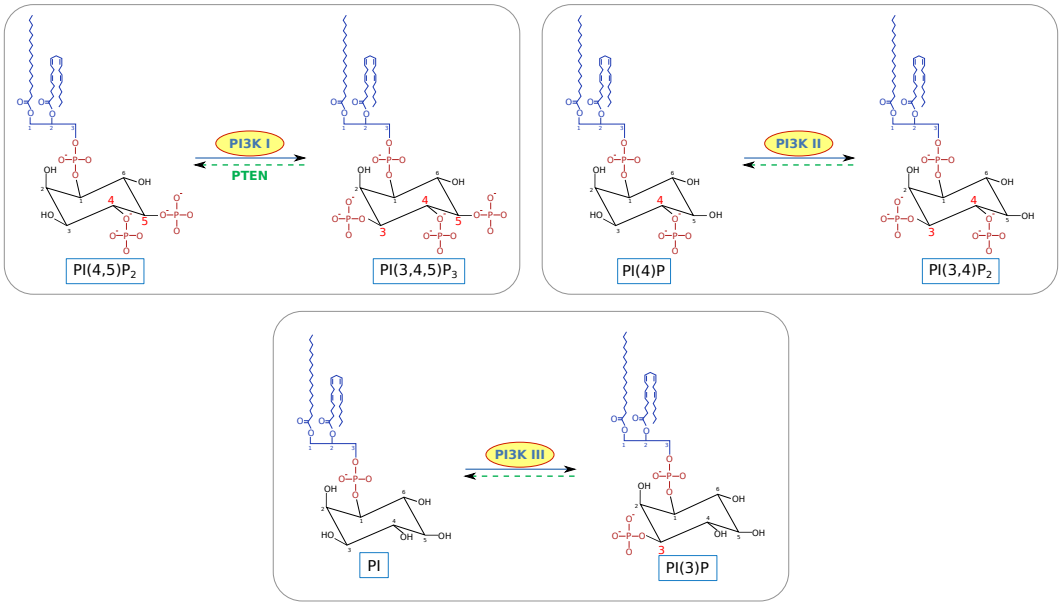

FIGURE *4.1* – ***Réactions enzymatiques catalysées par les différentes classes des PI3K humaines (adapté de [289]).***

Tandis que les protéines des classes I et III forment des hétérodimères composés d'une sous-unité catalytique et d'une sous-unité régulatrice, seules trois protéines catalytiques ont été recensées pour la classe II (Figure 4.2).

Classe I

Avec neuf gènes chez l'Homme, la classe I constitue la plus grande classe de PI3K. Selon leur capacité à lier ou non une sous-unité régulatrice de type p85, ces neuf protéines sont subdivisées en deux catégories :

- Le groupe IA, composé de p110α, p110β, p110δ, p85α (et ses formes alternatives p55α et p50α), p85β et p55γ.
- Le groupe IB, composé de p110γ, p87 et p101.

Chez l'Homme, l'ensemble des combinaisons entre les trois sous-unités catalytiques et les trois sous-unités régulatrices de la classe IA sont possibles [290, 291, 292]. De même, la seule protéine catalytique de la classe IB (p110γ) se lie alternativement aux deux sous-unités régulatrices de cette classe (p87 et p101) [293].

L'activation fonctionnelle de la classe IA est principalement réalisée par les RTK (***R**eceptor **T**yrosine **K**inases*) [294, 295] ainsi qu'à travers l'IGF1 (***I**nsulin-like **G**rowth **F**actor **1***) [296]. La classe IB est quant à elle essentiellement activée par les GPCR [294, 296].

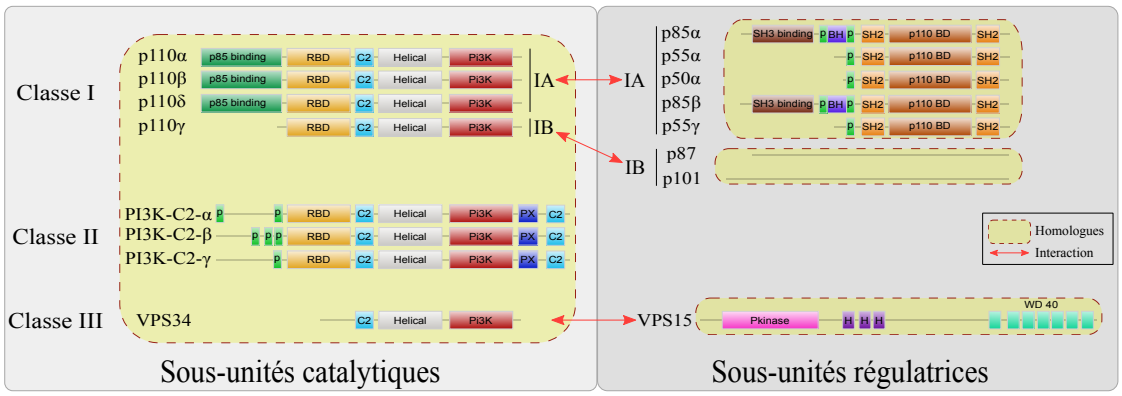

FIGURE *4.2* – ***Représentation des compositions en domaines des 14 protéines PI3K humaines réparties en trois grandes classes (adapté de [297]).*** Les cadres jaunes représentent les relations d'homologies entre les différentes protéines.

De par leur rôle précurseur dans de nombreux processus cellulaires, les variations d'activité ou d'expression des gènes de la classe I sont associées à de nombreuses pathologies. Les protéines p110α et p110β étant, entre autres, activées par les récepteurs à insuline, elles sont devenues des cibles de choix dans la lutte contre le diabète [298, 299]. La protéine p110δ, particulièrement exprimée dans les cellules hématopoïétiques [300] est, quant à elle, impliquée dans certaines pathologies hématologiques [301] telles que les leucémies lymphoïdes chroniques [302, 303] et les leucémies aiguës myéloblastiques [304, 305]. La protéine p110γ, de la classe IB est pour sa part particulièrement impliquée dans les maladies vasculaires comme l'athérosclerose [306, 307] et dans le cancer du sein [308]. De nombreuses études ont également démontré son rôle dans la motilité de cellules du système immunitaire [309, 310, 311, 312]. Enfin, la protéine p110α constitue une cible potentielle intéressante dans la lutte contre le cancer. En effet, le gène correspondant est muté dans environ 25% des cancers du sein [313, 314, 315], dans 15-20% des cancers colorectaux [316, 317, 318] et dans environ 10% des cancers de l'œsophage [319, 320]. Au total, l'expression de p110α est dérégulée dans plus de 30% de tumeurs solides [319].

Leurs impacts sur la santé humaine constatés, plusieurs inhibiteurs des PI3K de classe I ont rapidement été développés [321]. Ainsi, les deux principaux, nom-

més Wortmannin et LY294002, ont été respectivement découverts en 1993 [322] et 1994 [323]. Plus récemment, la molécule IC87114 a été identifiée comme inhibant spécifiquement la protéine p110δ [281, 324].

Classe II

Composée uniquement de trois protéines catalytiques (PI3K-C2α, PI3K-C2β et PI3K-C2γ), la classe II constitue la classe la moins bien caractérisée des trois. Si leur substrat de prédilection n'a toujours pas été identifié avec certitude [288], une étude récente a montré un lien entre la présence de PI(4,5)P_2, de la protéine PI3K-C2α et la production de PI(3,4,5)P_3 [325]. Concernant leur activation, elle est principalement réalisée par des chemokines de type MCP-1 [326], des cytokines (comme TNF-α et la leptin) [327] ou encore par les acides lysophosphatidiques (LPA) [328]. Au contraire, le Tamoxifène induit une baisse de leur expression chez la souris [329].

Les fonctions biologiques des PI3K de classe II n'ont été mises en évidence que très récemment par rapport aux deux autres classes. Plusieurs études ont tout d'abord montré une implication de ces protéines dans l'endocytose par clathrine [325], puis une étude de 2014 a démontré l'effet létal de l'absence de PI3K-C2α chez les embryons de souris [330]. Au niveau pathologique, la protéine PI3K-C2α est impliquée dans l'angiogenèse tumorale [329] et la protéine PI3K-C2β dans la tumorigenèse des neuroblastomes [331]. Enfin, en 2015, une des premières études consacrée à la protéine PI3K-C2γ, la moins étudiée des trois, a montré qu'elle était capable d'activer la protéine AKT2 suite à la présence d'insuline dans le milieu extracellulaire [332].

Classe III

Chez l'Homme, la classe III est uniquement composée d'une sous-unité catalytique (PIK3C3 ou VPS34 pour ***V****acuolar* ***P****rotein* ***S****orting* ***34***) et d'une sous-unité régulatrice (VPS15). Le principal rôle de cet hétérodimère est la régulation du trafic membranaire et l'induction de la formation des autophagosomes du processus d'autophagie [333, 334, 335].

Homologues répertoriés

Dès 1993, Schu *et al.* ont montré que le génome de la levure ne contenait que les gènes codant pour les deux protéines de la classe III [336]. Par la suite, des homologues de VPS34 et VPS15 ont été détectés chez d'autres eucaryotes unicellu-

laires (*S. pombe* [337], *Candida albicans* [338], *Dictyostelium discoideum* [339]), chez quelques plantes [340] et micro-algues [341], chez les vertébrés [342] et également dans les organismes modèles *C. elegans* [343] et *D. melanogaster* [344]. Concernant les classes I et II, des homologues ont été identifiés chez les vertébrés, le nématode, la drosophile [345] et les Amoebozoa [339]. L'absence d'homologues de ces deux classes chez la levure a par la suite été confirmée par plusieurs études [345, 346].

Si les PI3K ont été très étudiées chez l'Humain du fait de leur implication dans de nombreuses pathologies, très peu d'études fonctionnelles ont été réalisées chez d'autres organismes. Pour les Excavata et les SAR, les analyses se sont plus intéressées à la variation des niveaux d'expression des PI3K de l'hôte qu'aux fonctions des PI3K du pathogène [347, 348]. Une étude de 2014 a néanmoins démontré que chez l'Apicomplexa *Toxoplasma gondii*, la seule protéine PI3K détectée est impliquée dans la forme et la taille des apicoplastes [349]. De nombreuses études ont par contre été réalisées chez l'Amoebozoa modèle *D. discoideum*, démontrant le rôle clef des PI3K des classes I et II dans le déplacement par chimiotaxie de ces organismes [350, 282, 283]. Un parallèle avec la fonction motrice de la protéine p110γ chez les neutrophiles humains a d'ailleurs été établit par plusieurs groupes de chercheurs [309, 351, 352]. Enfin, les PI3K ont été décrites comme impliquées dans l'autophagie et plus généralement dans le trafic intracytoplasmique chez de nombreuses espèces eucaryotes [353, 354, 355, 356, 357].

4.1.2 État de l'art sur leur histoire évolutive

Malgré leurs rôles biologiques primordiaux, seules deux reconstructions phylogénétiques incomplètes des PI3K ont été publiées avant mon travail de thèse. La première, réalisée en 2003 par Kawashima *et al.* [358], concerne les protéines catalytiques (classes I, II et III) et régulatrices (exceptée la classe IB) des Opisthokonta. La seconde, publiée par Brown et Auger en 2011 [289], est plus complète au niveau taxonomique puisqu'elle considère l'ensemble des eucaryotes, mais ne porte que sur les protéines catalytiques.

Kawashima *et al.*

Les arbres phylogénétiques de Kawashima *et al.* ont été construits par la méthode NJ avec 1000 réplications de *bootstrap*. Le premier jeu de données, composé des homologues de la classe III catalytique, regroupe 19 séquences issues de cinq espèces d'Opisthokonta différentes : *C. elegans*, *Ciona intestinalis*, *D. melanogas-*

ter, *H. sapiens* et *S. pombe*. Pour les protéines régulatrices de la classe IA, l'arbre phylogénétique fut inféré à partir de dix séquences provenant de six espèces. Enfin, seuls cinq homologues issus de cinq espèces ont été utilisés pour la phylogénie de de la protéine régulatrice de la classe III (Figure 4.3).

Figure omise pour des raisons de droits d'auteur.

FIGURE *4.3* – ***Phylogénies des PI3K par Kawashima et al. [358].*** (A) Phylogénie des sous-unités catalytiques. (B) Phylogénie des protéines régulatrices de la classe IA. (C) Phylogénie de la protéine régulatrice classe III. CE : *C. elegans*, CI : *C. intestinalis*, DM : *D. melanogaster*, HS : *H. sapiens*, MM : *M. musculus*, SC : *S. cerevisiae*, SP : *S. pombe*, XL : *Xenopus laevis*.

Bien qu'ils n'aient pas inféré d'arbre phylogénétique pour les protéines régulatrices de la classe IB, les auteurs indiquent néanmoins avoir détecté chez *C. intestinalis* la première protéine non mammifère de cette classe.

Brown et Auger

Plus récente, la phylogénie construite par Brown et Auger, concerne uniquement les PI3K catalytiques. Une recherche par BLASTp sur la banque de séquences NR (***N**on **R**edundant*) [359] avec un seuil $E < 10^{-10}$ a permis aux auteurs d'identifier 157 séquences provenant d'Opisthokonta (21 espèces), de plantes vertes (*A. thaliana*, *Glycine max* et *Oryza sativa*), d'Excavata (trois espèces de *Leishmania*, deux espèces de *Trypanosoma* et deux espèces d'*Entamoeba*) et d'Apicomplexa (sept espèces de *Plasmodium*). L'alignement a été réalisé avec CLUSTALW et ses paramètres par défaut. Les blocs conservés ont ensuite été identifiés manuellement (236 sites gardés). Enfin, les auteurs ont inféré l'arbre phylogénétique correspondant par la méthode NJ mais également par une approche bayésienne. La robustesse de l'arbre NJ a été estimée au moyen d'une procédure de *bootstrap* (1000 réplications). Dans le cas de l'approche bayésienne, le programme utilisé était MrBayes [360]. Par ailleurs, le modèle de mélange protéique par défaut de MrBayes et 10^6 générations

dont 10^4 de *burn-in* ont été utilisés. L'arbre phylogénétique consensus correspondant (Figure 4.4) a été raciné par des séquences de PI4K.

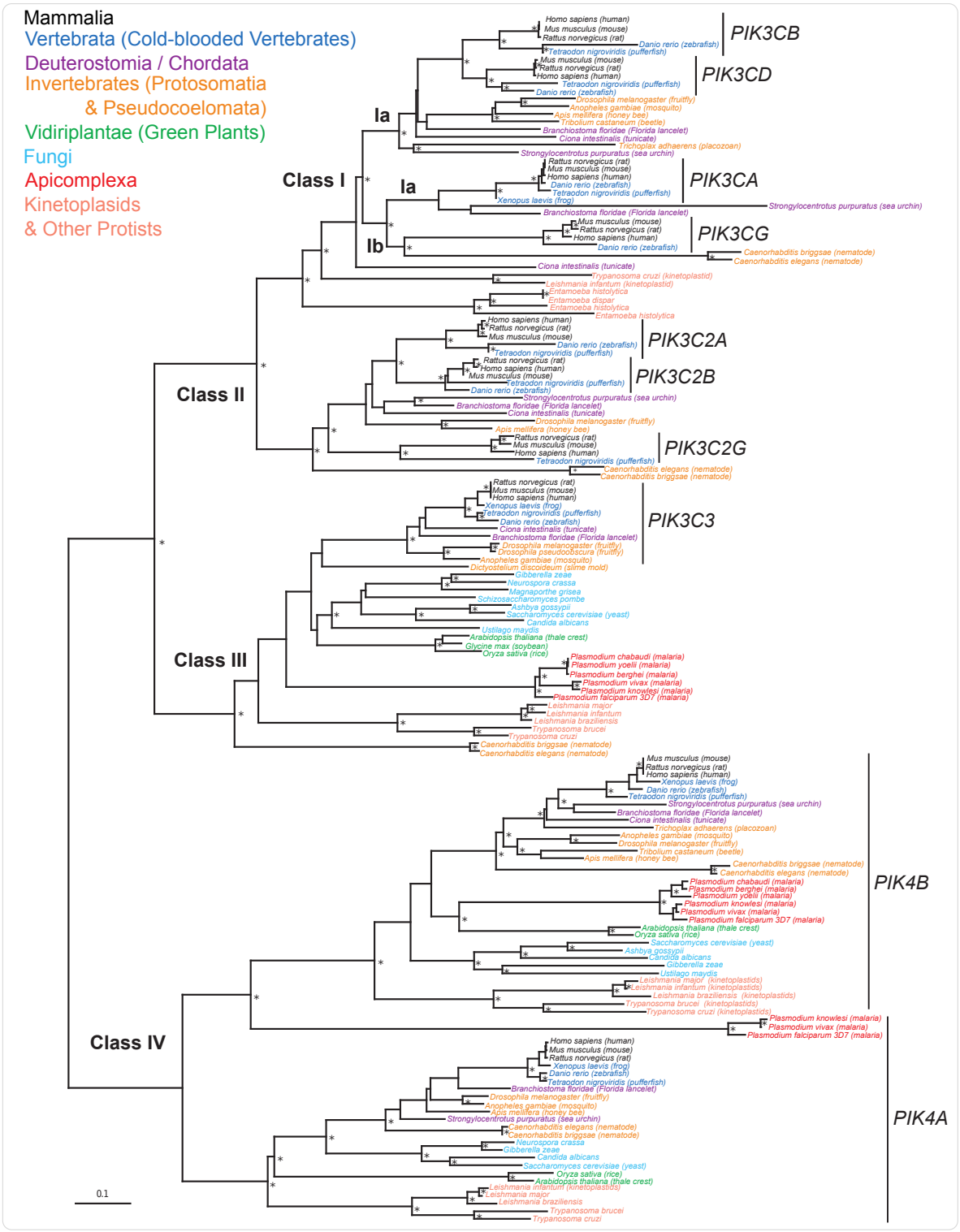

FIGURE 4.4 – ***Phylogénie des sous-unités catalytiques des PI3K, par Brown et Auger [289].*** Une astérisque (*) symbolise un nœud supporté par un *bootstrap* > 70% et une probabilité postérieure > 0.95.

Comparaison

Ces deux phylogénies soutiennent l'hypothèse d'une duplication ancienne conduisant à la séparation de la classe III et des classes I/II catalytiques, suivi d'une duplication plus récente à l'origine des classes I et II. Aucune duplication n'est détectée au sein des protéines catalytiques de la classe III. Les duplications intra-classe II sont identiques entre les deux analyses, avec les protéines PI3K-C2α et PI3K-C2β plus proches l'une de l'autre que de la protéine PI3K-C2γ. Concernant les duplications intra-classe I, la phylogénie de 2003 montre un placement (non soutenu) de la protéine p110γ à la base du groupe p110α-p110δ-p110β tandis que le phylogénie de 2011 regroupe les protéines p110γ et p110α ensemble (groupement soutenu).

L'émergence de la classe III catalytique semble très ancienne et datée d'avant LECA par Brown et Auger. Les résultats de Kawashima *et al.* montrent, quant à eux, que les protéines régulatrices semblent plus récentes. Néanmoins, comme expliqué ci-dessus, seules des espèces opisthocontes ont été considérées dans cette dernière analyse.

Objectifs de l'analyse

Malgré ces deux études, les connaissances de l'histoire évolutive des PI3K n'étaient que très partielles en 2014. En outre aucune phylogénie de la classe IB régulatrice n'était disponible de même qu'aucune information phylogénétique sur les protéines régulatrices des classes IA et III non Opisthokonta. Profitant de l'accumulation de données moléculaires et du nombre important de génomes complets séquencés ces dernières années, j'ai réalisé une analyse phylogénétique détaillée de la famille PI3K. Les résultats obtenus ont été publiés dans *BMC Evolutionary Biology* [361] (voir Annexe B) en Octobre 2015.

4.2 Histoire évolutive des PI3K

4.2.1 Matériel et méthodes

Pour cette analyse, j'ai utilisé les séquences protéiques des 14 protéines humaines issues d'UniProtKB [362] (Table 4.1). Après avoir déterminé les relations d'homologies entre ces protéines (Figure 4.2), j'ai reconstruit quatre phylogénies correspondant : i) aux sous-unités catalytiques, ii) aux protéines régulatrices de la classe

IA, iii) aux protéines régulatrices de la classe IB, et iv) à la protéine régulatrice de la classe III (VPS15).

	Nom commun	Nom du gène	Identifiants Ensembl	Identifiants UniProt	Taille
Sous-unités catalytiques					
	p110α	PIK3CA	ENSP00000263967	P42336	1068
IA	p110β	PIK3CB	ENSP00000418143	P42338	1070
	p110δ	PIK3CD	ENSP00000446444	O00329	1044
IB	p110γ	PIK3CG	ENSP00000392258	P48736	1102
	PI3K-C2α	PIK3C2A	ENSP00000265970	O00443	1086
II	PI3K-C2β	PIK3C2B	ENSP00000356155	O00750	1634
	PI3K-C2γ	PIK3C2G	ENSP00000266497	O75747	1445
III	VPS34	PIK3C3	ENSP00000262039	Q8NEB9	887
Sous-unités régulatrices					
	p85α	PIK3R1	ENSP00000274335	P27986	724
IA	p85β	PIK3R2	ENSP00000222254	O00459	728
	p55γ	PIK3R3	ENSP00000361075	Q92569	461
IB	p101	PIK3R5	ENSP00000392812	Q8WYR1	880
	p87	PIK3R6	ENSP00000475670	Q5UE93	754
III	VPS15	PIK3R4	ENSP00000349205	Q99570	B3

TABLE *4.1* – ***Nomenclature des 14 PI3K humaines.***

Identification des homologues

J'ai effectué la recherche de similarité des séquences à partir de chacune des protéine PI3K humaines à l'aide du logiciel BLASTp et les paramètres par défaut, excepté le seuil de similarité, fixé à $E < 10^{-30}$. Les homologues Metazoa ont été identifiés parmi les séquences de la banque Ensembl [363] tandis que les homologues eucaryotes non Metazoa sont issus des divers protéomes complets récupérés dans la banque BioProject du NCBI [359] (voir section 5.2.2). Suite à une première recherche BLASTp ayant conduit à l'identification de nombreuses séquences éloignées de l'Humain, certaines de ces séquences ont été utilisées comme graines pour une seconde itération avec les mêmes paramètres.

Face au nombre important d'homologues détectés, j'ai défini deux différents échantillons taxonomiques afin de réduire le bruit et la redondance phylogénétique. Les contraintes de cette sélection étant : i) d'obtenir une représentativité de chaque groupe eucaryote aussi équilibrée que possible, ii) de limiter le nombre de séquences évoluant rapidement, et iii) d'inclure les organismes modèles tels que *S. cerevisiae* et *C. elegans*. Ainsi, pour le jeu de données des protéines catalytiques, j'ai sélectionné 44 espèces représentatives parmi lesquelles seuls 10 des 42 mammifères de la banque Ensembl ont été conservés. Les jeux de données des protéines régulatrices ayant moins d'homologues, seule la sélection des espèces de mammifères a été réalisée pour ces phylogénies.

Les transcrits alternatifs des quatre jeux de données ont ensuite été identifiés à l'aide du système ACNUC [364] pour les espèces d'Ensembl et les E-utilities [365] pour les autres séquences. J'ai ensuite sélectionné manuellement le transcrit alternatif le plus conservé et/ou le mieux aligné pour chaque locus génomique.

Inférences des arbres

Pour les alignements multiples, j'ai comparé les résultats retournés par les logiciels PRANK et MAFFT grâce à l'indice norMD (***n**ew **o**bjective function for sco**r**ing using **M**ean **D**istance*) [366]. Ses scores étant largement supérieurs, j'ai décidé d'utiliser MAFFT. Suivant les recommandations des auteurs, j'ai utilisé l'algorithme `localpair` et fixé le nombre maximum d'itérations à 100.

A partir des alignements multiples obtenus, la sélection des blocs conservés a été réalisée à l'aide du programme BMGE, celui-ci présentant de meilleures performances que Gblocks. J'ai testé de nombreux jeux de paramètres (plusieurs matrices BLOSUM, plusieurs pourcentage de gap autorisés et différentes longueurs minimale des blocs) pour chacun des quatre jeux de données. Les valeurs choisies correspondent à celles offrant le meilleur compromis entre nombre et qualité des sites gardés (Table 4.2). Le nombre de gaps restant dans chaque alignement filtré est listé dans les tables B.1 à B.6 de l'annexe B.1. L'ensemble des jeux de données sont téléchargeables à l'adresse : `http://pbil.univ-lyon1.fr/datasets/Philippon2015/` .

Jeu de données	Nombre de paralogues humains	Nombre d'homologues	Paramètres BMGE utilisés			Nombre de sites sélectionnés	Modèle évolutif
			Matrice BLOSUM	Nombre maximal de gaps autorisé (%)	Taille minimale du bloc		
Sous-unités catalytiques (sélection d'espèces)	8	139	30	50	2	398	UL3+G
Protéine régulatrice classe III	1	117	30	40	3	839	UL3+G
Sous-unités régulatrices IA	3	126	30	40	4	539	JTT+G
Sous-unités régulatrices IB	2	67	30	50	4	599	JTT+G
Sous-unités catalytiques (total)	8	1055	30	70	2	468	JTT+G
Sous-unités catalytiques II (sélection MIC)	3	108	30	40	4	1113	JTT+G
Sous-unités catalytiques I (sélection MIC)	4	185	30	40	4	828	LG+G

TABLE *4.2* – ***Caractéristiques et paramètres utilisés pour les différents jeux de données.***

J'ai ensuite testé plusieurs modèles évolutifs. Les modèles standards ont été comparés à l'aide du logiciel ProtTest [367]. Selon les jeux de données, le modèle LG+Γ_4 ou le modèle JTT+Γ_4 ont été sélectionnés. Pour les modèles plus récents, tels que UL3 [229] et CAT20 [228], j'ai regardé la vraisemblance de chaque arbre inféré et calculé le BIC (***B**ayesian **I**nformation **C**riterion*) [368] pour chaque jeu de données. Les résultats sont présentés dans la Table 4.3 dans laquelle le score du meilleur modèle évolutif est indiqué en vert.

Jeu de données	Nombre de paralogues humains	Nombre de sites sélectionnés	LogLikelihood			
			LG+Γ4	JTT+Γ4	UL3+Γ4	CAT20+Γ4
Sous-unités catalytiques (sélection d'espèces)	139	398	-71987,680	-74278,953	**-71404,807**	
Protéine régulatrice classe III	117	839	-89690,029	-90488,929	**-89400,621**	-91287,484
Sous-unités régulatrices classe IA	126	539	-33798,331	**-33581,092**	-33783,764	-34544,087
Sous-unités régulatrices classe IB	67	599	-32063,930	**-31734,586**	-32151,918	-33200,347
Sous-unités catalytiques classe II (sélection MIC)	108	1113	-78459,294	**-78219,468**	-78659,953	-80768,245
Sous-unités catalytiques classe I (sélection MIC)	185	828	**-89058,436**	-89245,368	-89179,771	

Jeu de données	Nombre de paralogues humains	Nombre de sites sélectionnés	BIC			
			LG+Γ4 (2n-3+1 paramètres)	JTT+Γ4 (2n-3+1 paramètres)	UL3+Γ4 (2n-3+3 paramètres)	CAT20+Γ4 (2n-3+1 paramètres)
Sous-unités catalytiques (sélection d'espèces)	139	398	145627,62	150210,17	**144473,85**	
Protéine régulatrice classe III	117	839	180941,93	182539,73	**180376,58**	184136,84
Sous-unités régulatrices classe IA	126	539	69169,09	**68734,61**	69152,54	70660,60
Sous-unités régulatrices classe IB	67	599	64972,04	**64313,35**	65160,80	67244,87
Sous-unités catalytiques classe II (sélection MIC)	108	1113	158419,76	**157940,11**	158835,10	163037,66
Sous-unités catalytiques classe I (sélection MIC)	185	828	**180589,47**	180963,33	180845,58	

TABLE *4.3* – ***Valeurs de vraisemblance et de BIC pour chaque jeu de données.*** Les meilleures valeurs de vraisemblance et de BIC sont indiquées en vert.

L'inférence des arbres phylogénétiques a ensuite été réalisée à l'aide du logiciel PhyML [369] et des modèles évolutifs précédemment sélectionnés. Le paramètre de forme de la loi Gamma a été estimé au maximum de vraisemblance par PhyML et quatre catégories de taux de substitutions ont été utilisées. Pour estimer le support statistique des branches, j'ai utilisé le test de Shimodaira-Hasegawa (SH) ainsi qu'une approche de type *bootstrap* (BS) (100 réplications).

Pour le jeu de données composé des homologues catalytiques issus des espèces sélectionnées, j'ai également effectué une inférence bayésienne. Pour cela, j'ai utilisé le logiciel MrBayes avec les paramètres par défaut et un modèle de mélange. Sept millions d'itérations MCMC ont été nécessaires pour atteindre la convergence et la première moitié a été retirée en tant que zone de *burn-in*. Enfin, un arbre de

consensus majoritaire à 50% a été construit en échantillonnant 1000 arbres dans la distribution des probabilités postérieures. Cet échantillon a également été utilisé pour déterminer les probabilités postérieures (PP) des différents clades.

Composition en domaines

Dans le but d'éclaircir certaines relations de parenté et d'inférer les fonctions biologiques des différents homologues identifiés, je me suis intéressée aux compositions en domaines de ces séquences. Pour cela, j'ai utilisé le programme HMMScan du package HMMER [370, 371] avec les paramètres par défaut. Les deux bases de données PfamA et PfamB ont été interrogées. Pour les homologues des protéines régulatrices de classe IA, j'ai également utilisé le logiciel Batch CD-Search [372] pour confirmer la présence du domaine p110 binding chez les séquences provenant des espèces non Euteleostomi.

4.2.2 Résultats

Afin de reconstruire l'histoire évolutive de la famille des PI3K la plus complète et précise possible, j'ai procédé en deux étapes successives. Je me suis d'abord intéressée à la distribution taxonomique des PI3K chez l'ensemble des Eucaryotes et inféré les arbres phylogénétiques correspondants. Cette étape a servi à identifier les événements évolutifs majeurs subis par ces protéines. J'ai ensuite réalisé une étude plus précise chez les espèces appartenant au groupe MIC (***M****etaoza,* ***I****chthyosporea et* ***C****hoanoflagellida*) afin d'étudier les duplications plus récentes ayant conduit à la grande expansion des protéines PI3K dans ces lignées, y compris chez l'Homme.

a) Protéines catalytiques

Evénements majeurs

Pour les huit sous-unités catalytiques, la recherche de similarité a permis d'identifier 1055 homologues provenant de l'ensemble des grands groupes taxonomiques eucaryotes. De manière analogue à l'étude de Brown et Auger, j'ai raciné l'arbre correspondant par les séquences de PI4K (Figure annexe B.1).

Ce premier arbre montre quatre duplications majeures. Les deux premières sont antérieures à LECA et sont, respectivement, à l'origine de la divergence PI3K/PI4K et à la séparation de la classe III et d'une classe I/II ancestrale (séquences de SAR, d'Excavata et d'Amoebozoa). La troisième duplication majeure a lieu plus tard et a produit la séparation des classes I et II. Enfin, une dernière duplication, beaucoup

plus récente, est à l'origine de la séparation des classes IA et IB.

Après une sélection d'espèces représentatives (voir section 4.2.1), j'ai inféré un arbre réduit des sous-unités catalytiques dans lequel seules 139 séquences ont été gardées. Une sélection parmi les séquences de PI4K a également été effectué afin d'éviter tout artefact dû à un groupe externe trop étoffé. Les arbres reconstruits par maximum de vraisemblance et par inférence bayésienne sont respectivement présentés sur les Figures 4.5 et 4.6.

Ces trois arbres sont dans l'ensemble congruents et montrent une séparation claire entre la classe III et le regroupement des classes I et II. Ces deux groupements sont soutenus par des valeurs de BS égales à 97% et 86%, un SH de 0.94 et 0.97 et deux PP égales à 1.0.

Le premier groupe est composé des homologues de la protéine de la classe III (VPS34 ou PK3C3) provenant d'organismes de tous les grands groupes eucaryotes : SAR, Excavata, Archaeplastida, Amoebozoa et Opisthokonta (*i.e.* Fungi, Metazoa et organismes unicellulaires apparentés). Des séquences provenant de l'Haptophyta *Emiliana huxleyi*, du Cryptophyta *Guillardia theta* et de l'Apusozoa *Thecamonas trahens* sont également détectées. Néanmoins, aucune protéine PI3K catalytique de classe III ou d'autre classe n'a été décelée parmi les trois protéomes complets d'algues rouges de notre banque de séquences.

Le second groupe est composé des homologues des classes I et II provenant de l'ensemble des lignées eucaryotes excepté des champignons et des plantes. Une recherche de similarité spécifique dans ces deux groupes taxonomiques parmi les séquences de la banque NR du NCBI a confirmé cette absence. De manière surprenante, une séquence provenant de la plante *Selaginella moellendorffii* se groupe à la base de ce cluster. Cette présence sera discutée dans la section 4.3 de ce chapitre.

Ces résultats suggèrent donc deux duplications successives des gènes catalytiques. La première, très ancienne, a eu lieu avant LECA et a conduit à la séparation des classes III et I-II. La seconde duplication, plus récente, est quant à elle à l'origine de la divergence des classes I et II. Puisqu'une seule copie est présente chez les Bikonta (SAR, Excavata et Haptophyta) tandis que deux copies (correspondant aux classes I et II) sont détectées chez les Unikonta (Amoebozoa et Opisthokonta), on peut penser que cette duplication a eu lieu chez le dernier ancêtre commun des Unikonta. Néanmoins, le groupement des homologues Bikonta (SAR et Excavata) avec les protéines unicontes de la classe I (Figures 4.5 et 4.6) suggère plutôt une duplication antérieure à LECA. Cette deuxième hypothèse implique que les Bikonta auraient ensuite perdu le gène codant pour la protéine de la classe II.

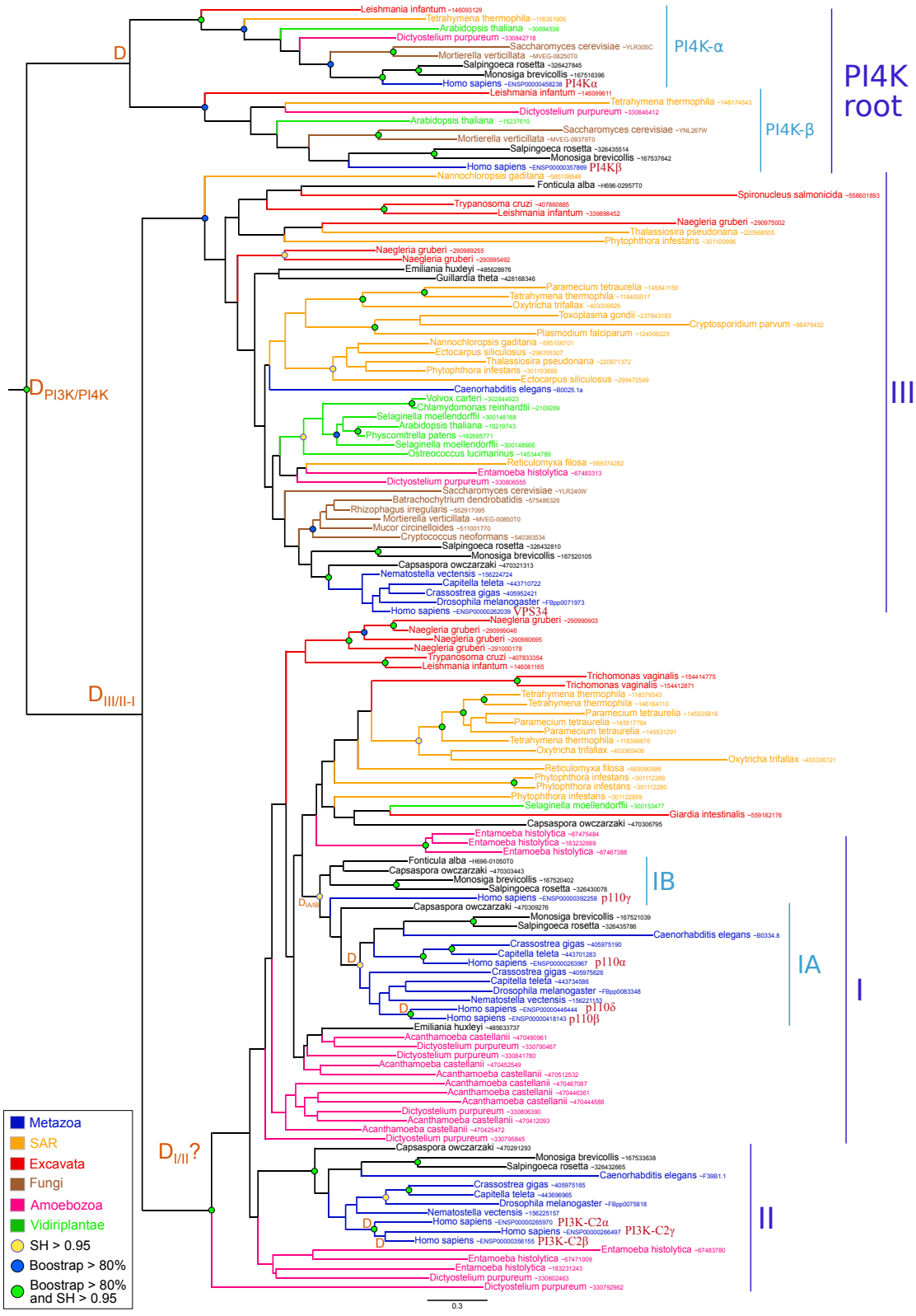

FIGURE 4.5 – ***Arbre phylogénétique des sous-unités catalytiques, raciné par les PI4K, après sélection d'espèces représentatives.*** L'arbre a été inféré par maximum de vraisemblance en utilisant le modèle évolutif UL3+Γ_4 (398 sites conservés, 139 séquences). Les cercles verts symbolisent des branches supportées par un SH > 0.95 et un BS $> 80\%$. Les cercles bleus et jaunes correspondent respectivement à un SH > 0.95 et à un BS $> 80\%$. Les événements de duplication sont matérialisés par un D orange. La barre d'échelle représente le nombre moyen de substitutions par site.

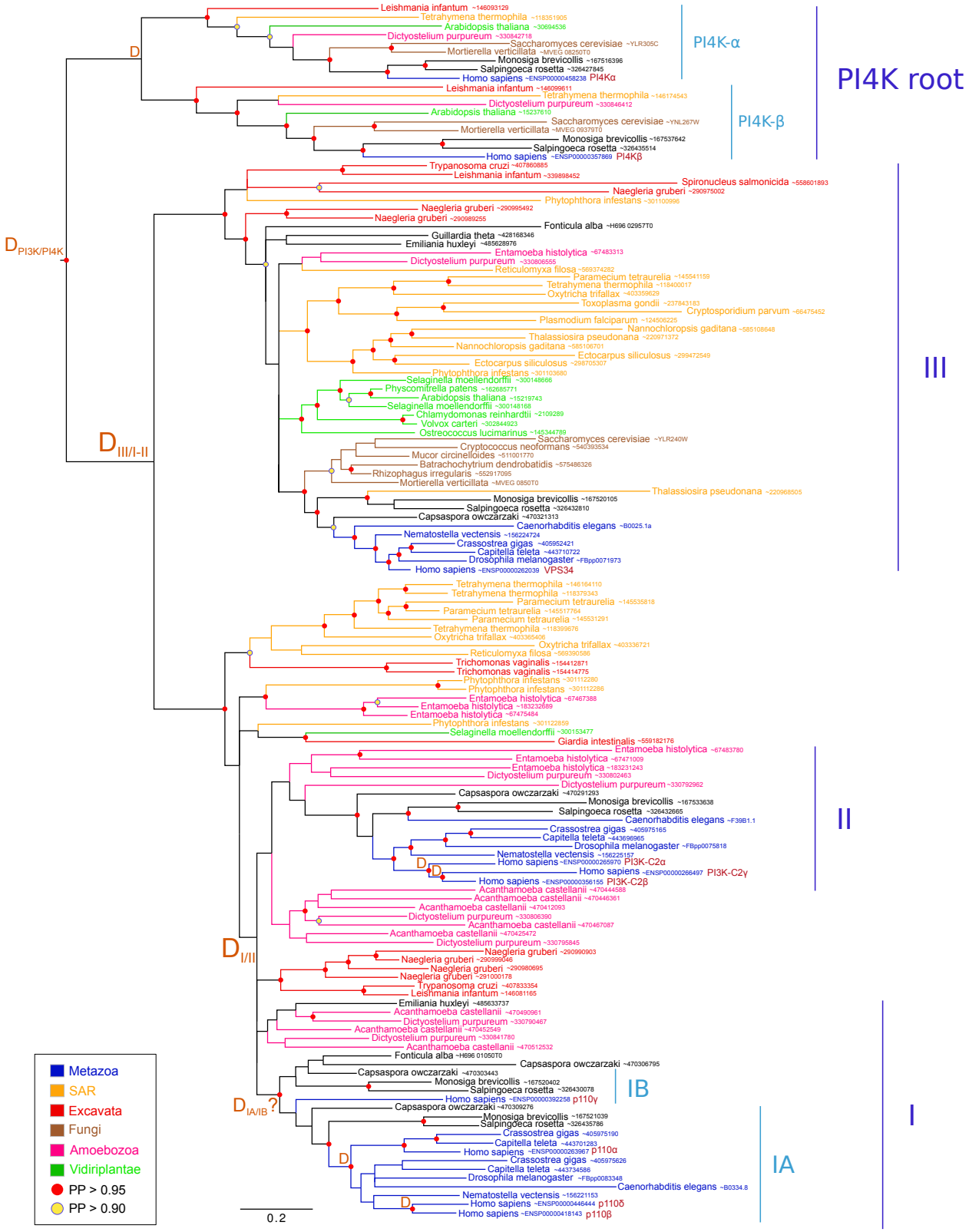

FIGURE *4.6* – ***Arbre phylogénétique des sous-unités catalytiques, raciné par les PI4K, après sélection d'espèces représentatives.*** L'arbre a été inféré par méthode bayésienne (398 sites conservés, 139 séquences). Les cercles jaunes et rouges correspondent respectivement à une PP > 0.90 et > 0.95. Les événements de duplication sont matérialisés par un D orange. La barre d'échelle représente le nombre moyen de substitutions par site.

Néanmoins, ce groupement n'est statistiquement supporté dans aucun de ces deux arbres (BS < 80%, SH < 0.95 et PP < 0.5). De plus, dans l'arbre inféré avec l'ensemble des 1055 homologues détectés (Figure B.1), ces séquences de Bikonta se placent à la base des deux copies d'Unikonta.

Au niveau du nombre de pertes, le second scénario implique quatre pertes indépendantes tandis que le premier n'en implique que trois : la classe I/II chez les plantes et les classes I et II chez les champignons. On peut donc penser que le premier scénario, dans lequel la duplication I/II a eu lieu chez le dernier ancêtre commun des Unikonta, est plus vraisemblable. Néanmoins, ces analyses ne permettent pas de l'affirmer avec certitude.

Je me suis donc ensuite intéressée à la composition en domaines des différents homologues catalytiques identifiés (Figure 4.7), composition précédemment déterminée pour les protéines humaines [373, 297].

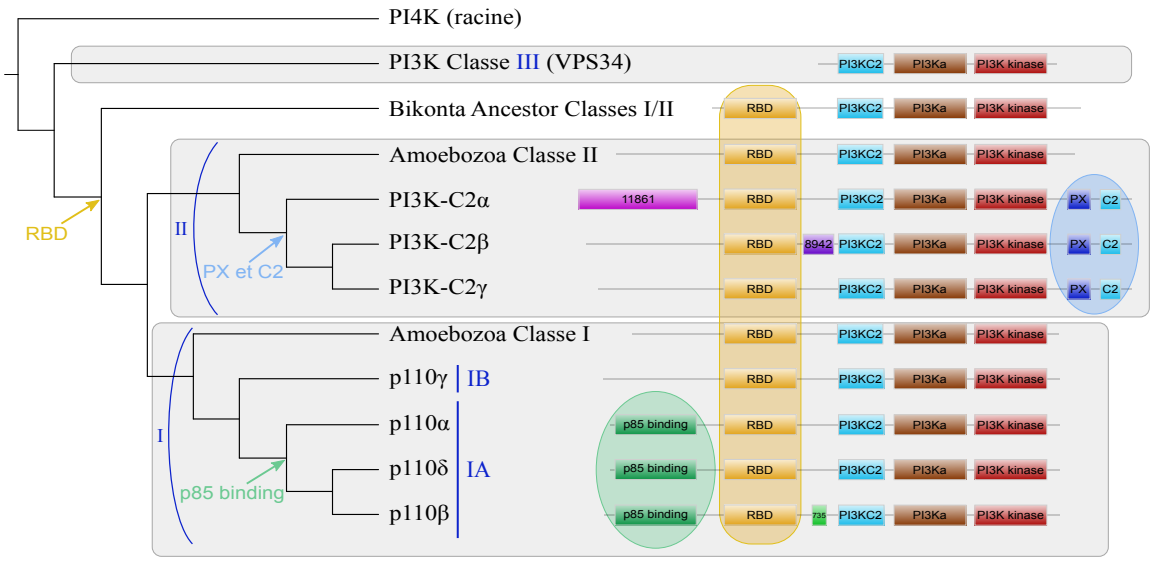

FIGURE 4.7 – ***Représentation schématique de la composition en domaines des protéines catalytiques eucaryotes.*** L'arbre à gauche correspond à la représentation schématique de l'arbre de la Figure 4.5 sous l'hypothèse d'une duplication spécifique des Unikonta. Les cercles colorés symbolisent les domaines spécifiques des sous-groupes des classes I et II. Les tailles des rectangles ne sont pas proportionnelles à la taille réelle des domaines.

J'ai tout d'abord confirmé la présence dans le même ordre de trois domaines protéiques chez l'ensemble des sous-unités catalytiques PI3K : PI3KC2 (numéro d'accession PF00792), PI3KA (PF00613) et PI3K kinase (PF00454). J'ai ensuite détecté la présence du domaine *Ras Binding Domain* (ou RBD, PF00794) en position N-terminale, chez l'ensemble des homologues du groupe formé par les classes

I et II. L'émergence de ce domaine coïncide avec la duplication à l'origine de la divergence entre les classes III et I-II (*i.e.* avant LECA). Essentiel pour l'activation des protéines PI3K par la protéine RAS (***RA**t **S**arcoma*) [374, 375], l'acquisition de ce domaine suggère un changement fonctionnel de cette copie suite à la duplication.

Malheureusement, la composition en domaines identiques des homologues classe I/II Bikonta et des homologues classes I et classe II provenant des Amoebozoa ne permet pas de départager les deux scénarios évolutifs considérés ci-dessus.

Duplications spécifiques des classes I et II.

L'analyse des premiers arbres obtenus pour les sous-unités catalytiques (Figures B.1, 4.5 et 4.6) montrent que les différentes protéines des classes I et II ont émergé après la divergence entre les Metazoa et les Choanoflagellida. Afin d'obtenir l'histoire évolutive de l'ensemble des sous-unités catalytiques plus précise, j'ai donc inféré les arbres phylogénétiques correspondants aux homologues métazoaires des classes I et II. J'ai également gardé les séquences d'Ichthyosporea et de Choanoflagellida afin de disposer d'un groupe externe pour l'enracinement des arbres. Pour plus de clarté, ces homologues sont regroupés sous l'acronyme MIC.

Sous-unités catalytiques MIC de la classe II

L'arbre phylogénétique obtenu à partir des séquences MIC de la classe II (Figure 4.8) montre que les trois paralogues de cette classe (PI3K-C2α, PI3K-C2β et PI3K-C2γ) sont détectés chez l'ensemble des Vertebrata (*i.e* les Petromyzontidae, les Chondrichtyes, les Actinoptérygii, les Lepidosauria, les Aves et les Mammalia). Au contraire, une seule copie de ce gène est présente chez les autres Metazoa tels que les Mollusca, les Cnidaria ou les Arthropoda. Cela suggère que les duplications à l'origine des ces trois protéines chez l'Homme se sont produites chez le dernier ancêtre commun des Vertebrata (SH > 0.95). Néanmoins, la topologie de l'arbre soulève quelques questions. En effet, les trois copies détectées chez *Petromyzon marinus* (un Petromyzontidae) sont groupées avec les protéines PI3K-C2α tandis que seules deux copies de ce gène sont détectées chez *Callorhincus milii* (un Chondrichtyes). Ainsi, cette topologie, couplée à l'hypothèse de deux duplications chez l'ancêtre commun des Vertebrata, implique une perte de la copie PI3K-C2γ chez les Chondrichtyes, deux pertes indépendantes des copies PI3K-C2γ et PI3K-C2β chez les Petromyzontidae ainsi que deux duplications successives dans cette lignée. Néanmoins, étant donné la taille des branches menant aux trois copies de *P. marinus*, la position surprenante des ces trois protéines peut être expliquée par le

phénomène d'attraction des longues branches ou être dû à une vitesse d'évolution rapide des gènes PIK3C2B et PIK3C2G.

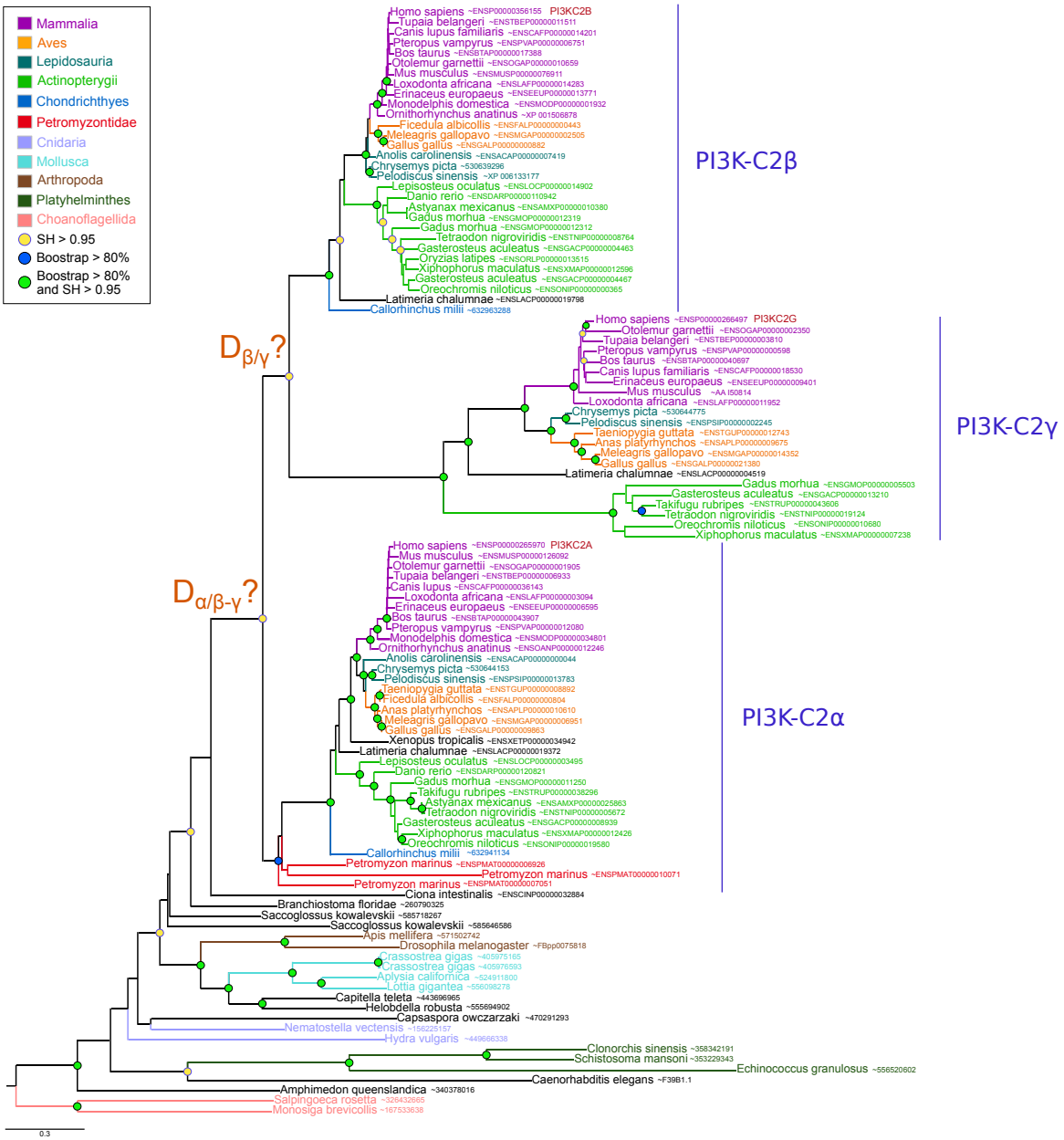

FIGURE *4.8* – ***Arbre phylogénétique des homologues MIC de la classe II catalytique.*** L'arbre a été inféré par maximum de vraisemblance sous le modèle évolutif JTT+Γ_4 (1113 sites conservés, 108 séquences). Les symboles utilisés sont identiques à ceux de la Figure 4.5.

Une hypothèse alternative serait une duplication à la base des Gnathostomata (*i.e.* le groupement des Chondrichtyes, Actinoptérygii, Lepidosauria, Aves et Mammalia) à l'origine des protéines PI3K-C2β et PI3K-C2γ, impliquant deux pertes in-

dépendantes du gène PIK3C2A chez *P. marinus*. Dans les deux cas, le gène codant la protéine PI3K-C2γ a été perdu dans la lignée des Chondrichtyes.

Au niveau de la composition en domaine des sous-unités catalytiques de classe II (Figure 4.7), j'ai confirmé la présence des quatre domaines : RBD (PF00794), PI3KC2 (PF00792), PI3KA (PF00613) et PI3K kinase (PF00454) chez tous les homologues. Tandis que les deux domaines PX (PF00787) et C2 (PF00168) en position C-terminale avaient reportés chez l'Homme [297], j'ai étendu cette observation à l'ensemble des Opisthokonta. Au contraire, les protéines issues des espèces d'Amoebozoa ne les possèdent pas. De plus, j'ai détecté deux autres domaines, PB011861 et PB008942, respectivement spécifiques des homologues de PI3K-C2α et PI3K-C2β. Ces différences de composition en domaine des trois protéines de la classe II peuvent suggérer des fonctions biologiques spécifiques différentes. Néanmoins, face à l'absence de données fonctionnelles concernant ces deux domaines et le manque de données de la littérature sur les protéines catalytiques de la classe II, il n'est pas possible d'émettre d'hypothèse quant aux fonctions biologiques des différentes copies.

Sous-unités catalytiques MIC de la classe I

Pour la classe I, des homologues des classes IA et IB sont retrouvés chez les Metazoa, Choanoflagellida, Ichthyosporea et une séquence de Nucleariidae (issue de *Fonticula alba*) est également détectée (Figures 4.5 et 4.9). Cela suggère que le dernier ancêtre commun des MIC possédait un gène de classe IA et un gène de classe IB (SH = 1 et BS > 80%) et que la duplication a eu lieu avant la divergence des MIC. Néanmoins, il n'est pas possible de dater plus précisément cet événement étant donné le faible nombre de protéomes complets disponibles pour les protistes apparentés au MIC (Nucleariidae et Apusozoa).

Au sein des Metazoa, la distribution taxonomique de la classe IB est plus réduite que celle de la classe IA. En effet, alors que des homologues aux protéines de la classe IB ne sont détectés que chez les chordés et chez *Amphimedon queenslandica*, des homologues de la classe IA sont également présents dans les protéomes de protostomiens (Annelida, Mollusca, Platyhelminthes et Arthropoda). Cette topologie soutenue indique des pertes secondaires du gène de la classe IB dans ces lignées. De plus, la présence de deux copies p110γ (classe IB) au sein des Actinopterygii et Chondrichthyes indiquent qu'une duplication a eu lieu chez l'ancêtre commun des Gnathostomata mais que l'un des deux paralogues a été perdu dans un second temps par les Sarcopterygii (BS > 80%).

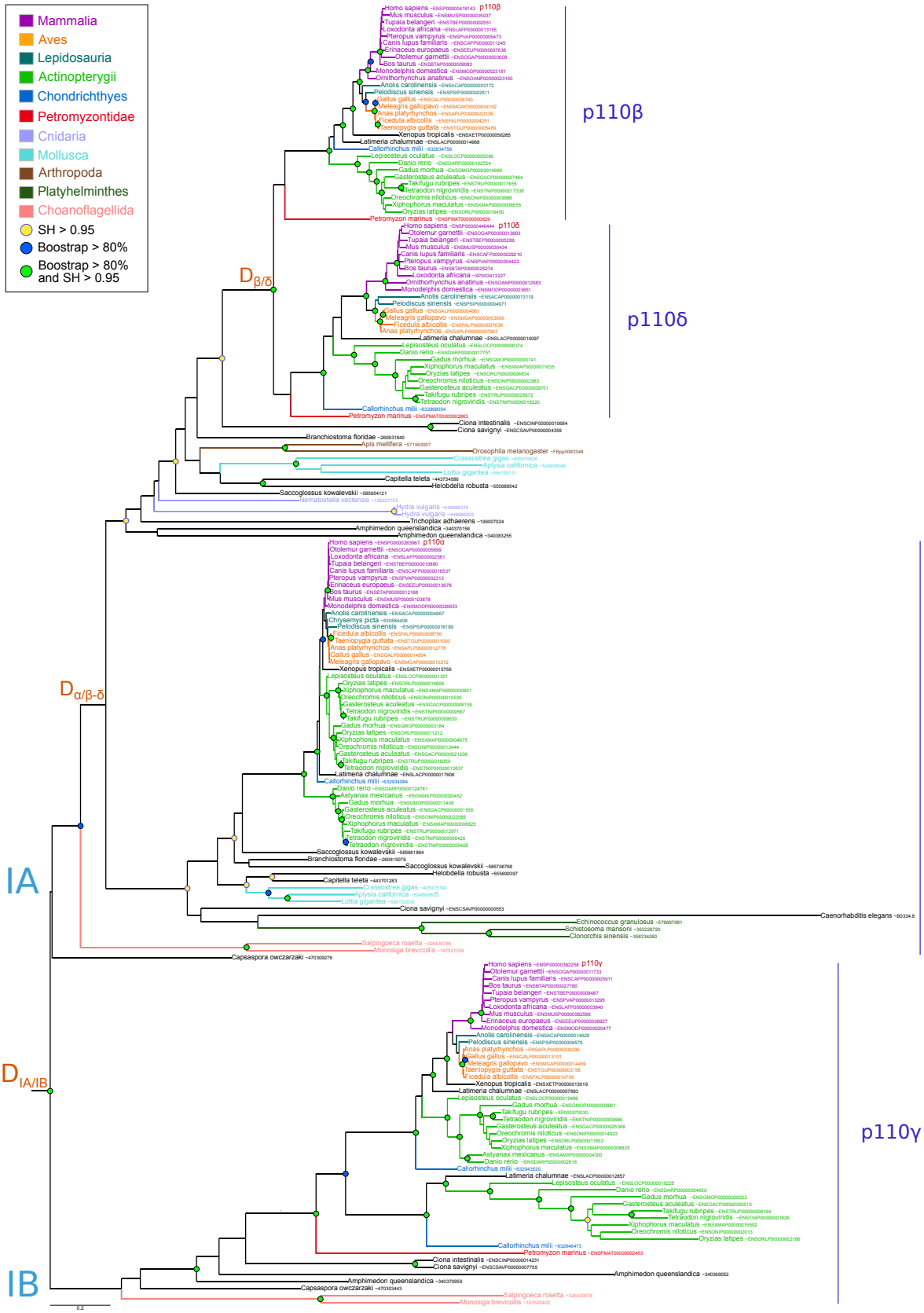

FIGURE 4.9 – ***Arbre phylogénétique des homologues MIC de la classe I catalytique.*** L'arbre a été inféré avec le modèle évolutif LG+Γ_4 (828 sites conservés, 185 séquences). Les symboles utilisés sont identiques à ceux de la Figure 4.5.

Au sein de la classe IA, les protéines p110δ et p110β sont évolutivement plus proches, tandis que la protéine p110α est plus divergente. L'arbre phylogénétique inféré (Figure 4.9) suggère qu'une première duplication a eu lieu chez le dernier ancêtre commun des Metazoa conduisant à la divergence du gène PIK3CA tandis qu'une seconde duplication, à l'origine des gènes PIK3CB et PIK3CD, s'est produite chez l'ancêtre commun des Vertebrata (SH = 1 et BS = 100%). Sous ce scénario évolutif, l'absence d'homologues à p110α chez les Arthropoda, Cnidaria et Placozoa (*Trichoplax adhaerens*) ainsi que l'absence de la forme ancestrale p110β/p110δ chez les Platyhelminthes sont à interpréter comme des pertes secondaires.

Au niveau de la composition en domaine des sous-unités catalytiques de la classe I (Figure 4.7), j'ai confirmé la présence du domaine p85 binding (PF02192) chez l'ensemble des homologues de la classe IA et son absence dans les protéines de la classe IB. De façon intéressante cette acquisition est antérieure à l'ancêtre commun des MIC et a eu lieu au moment de la divergence des classes IA et IB. Ces résultats sont cohérents avec la capacité spécifique des protéines catalytiques de la classe IA à se lier aux sous-unités régulatrices p85 de la classe IA [376, 377, 290, 378, 291]. Parmi ces protéines catalytiques, p110β possède un domaine spécifique (PB000735) situé entre les domaines RBD et PI3KC2 tandis que les protéines p110α et p110δ ont la même organisation en domaine.

b) Protéines Régulatrices

Classe III

La recherche de similarité de séquences pour la protéine régulatrice de classe III (VPS15) a permis d'identifier 117 homologues issus de l'ensemble des grands groupes eucaryotes, y compris des plantes et des champignons (Figure 4.10). Cette distribution taxonomique est cohérente avec celle des homologues catalytiques de classe III, signifiant que la présence d'une seule sous-unité catalytique et d'une seule sous-unité régulatrice a été conservée chez l'ensemble des Eucaryotes depuis LECA. A noter la position surprenante des séquences de *C. elegans* et *Fonticula alba* parmi le groupe des Bikonta, qui peut être due à un artefact d'attraction des longues branches.

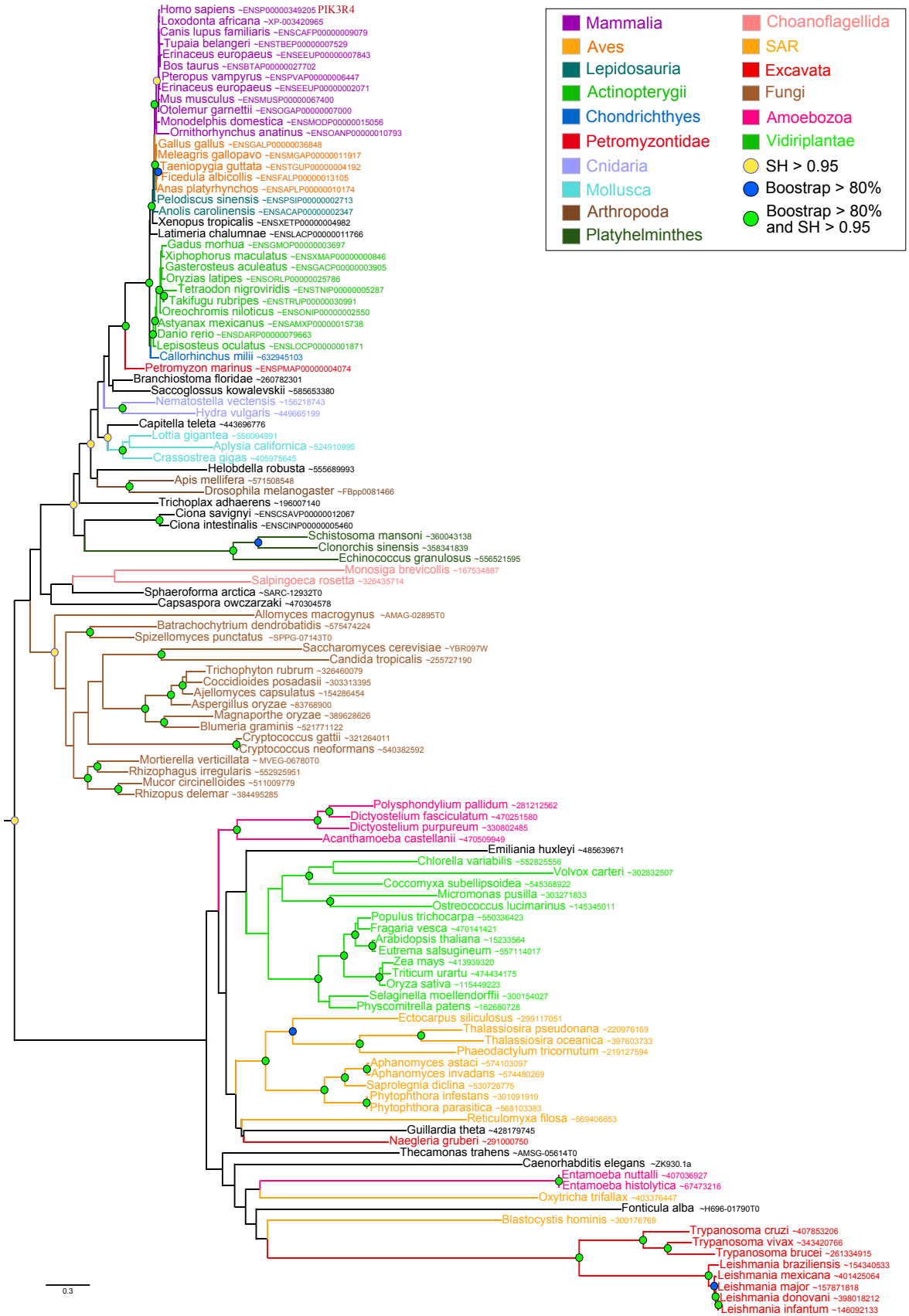

FIGURE *4.10* – ***Arbre phylogénétique de la protéine régulatrice classe III.*** L'arbre a été inféré par maximum de vraisemblance, avec le modèle évolutif UL3+Γ_4 (839 sites conservés, 117 séquences). Les symboles utilisés sont identiques à ceux de la Figure 4.5.

Si le gène codant la protéine régulatrice de classe III a été conservé chez l'ensemble des Eucaryotes actuels, la composition en domaine de cette dernière est assez différentes chez les Excavata (Figure 4.11). En effet, le domaine Pkinase (PF00069) situé en position N-terminale ainsi qu'au moins un domaine WD40 (PF00400) en position C-terminale sont présents chez toutes les séquences eucaryotes. En revanche, les séquences d'Excavata sont les seules à ne pas posséder les domaines PB007639, PB018740, PB001064, PB005262 et PB000285. De manière analogue, les Choanoflagellida ne possèdent ni le domaine PB010332 ni le domaine PB005367.

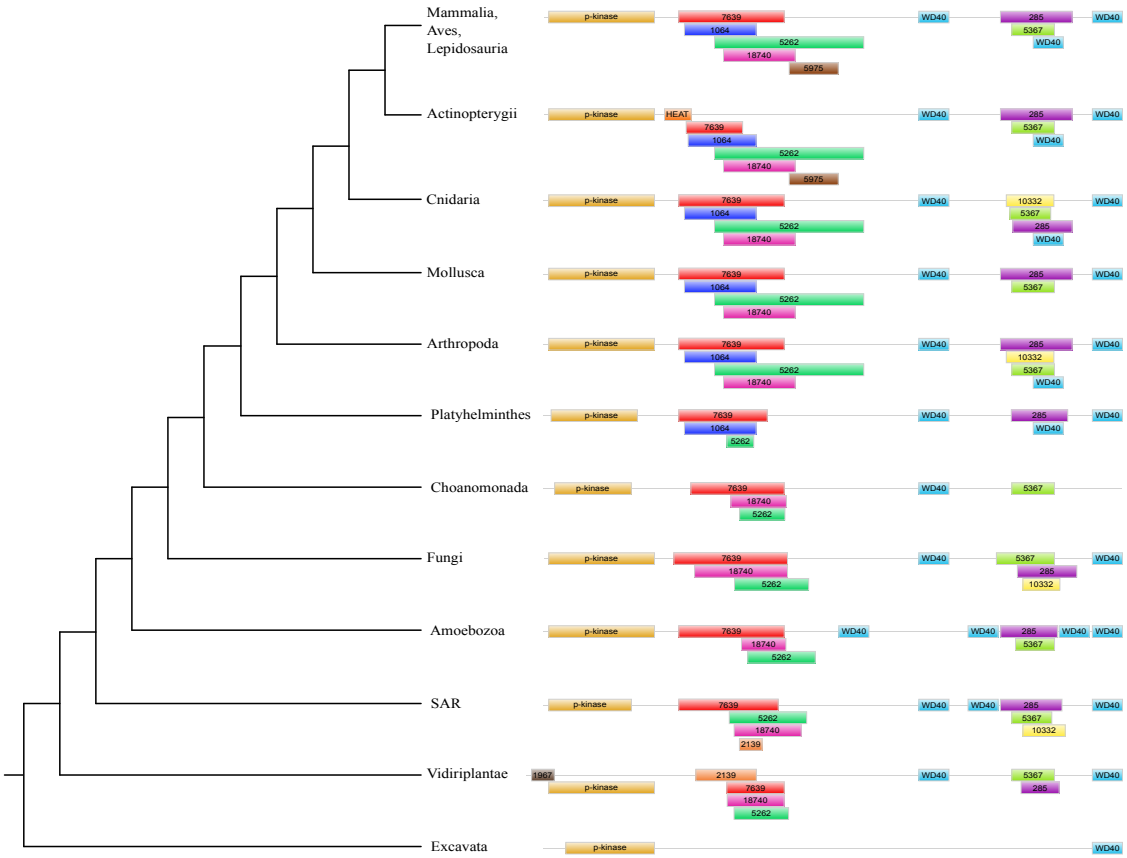

FIGURE *4.11* – ***Représentation de la composition en domaines des homologues à la protéine régulatrice de la classe III (VPS15).*** L'arbre à gauche correspond à la représentation schématique de l'arbre de la Figure 4.10. Les tailles des rectangles ne sont pas proportionnelles à la taille des domaines.

Classe IA

Contrairement à la protéine régulatrice de classe III, les protéines régulatrices de la classe IA ne possèdent d'homologues que chez les MIC (Figure 4.12).

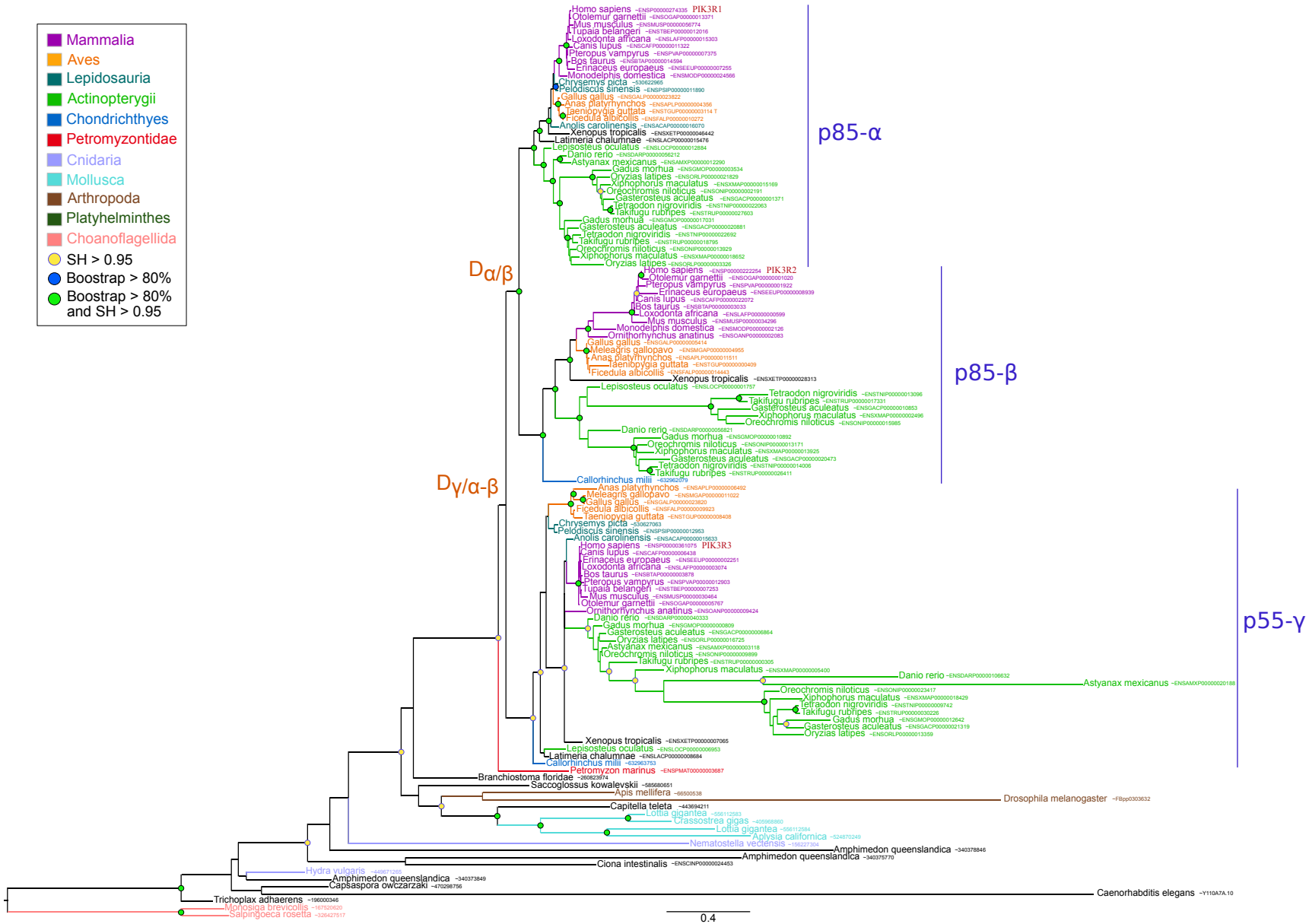

FIGURE *4.12* – ***Arbre phylogénétique des protéines régulatrices classe IA.*** L'arbre a été inféré par maximum de vraisemblance, avec le modèle évolutif JTT+Γ_4 (539 sites conservés, 126 séquences). Les symboles utilisés sont identiques à ceux de la Figure 4.5.

Tandis que trois protéines sont détectées chez les Euteleostomi, les protéines p85β et p85γ sont également détectées chez les Chondrichthyes. Au contraire, dans les autres lignées de Metazoa ainsi que chez les Ichtyosporea et les Choanoflagellida, une seule copie de ce gène est présente. Cet arbre phylogénétique suggère donc deux duplications successives Gnathostomata spécifiques suivi de la perte de la sous-unité p85α chez les Chondrichthyes. Néanmoins, seule la duplication à l'origine de la divergence entre les protéines p85α et p85β est statistiquement soutenue (BS = 91% et SH = 0.98). De plus, la présence d'un unique protéome de Chondrichthyes dans notre banque de données ne permet pas de conclure à une perte globale dans les lignées de ce groupe taxonomique. De façon surprenante, aucun orthologue à la protéine p85β n'est détectée chez les Lepidosauria.

Au niveau de la composition en domaines, toutes les séquences de la classe régulatrice IA présentent la même organisation en position C-terminale : un domaine ρ-gap (PF00620) suivi de deux domaines SH2 (PF00017) entrecoupés par un domaine p110-binding (PB011403) (Figure 4.13).

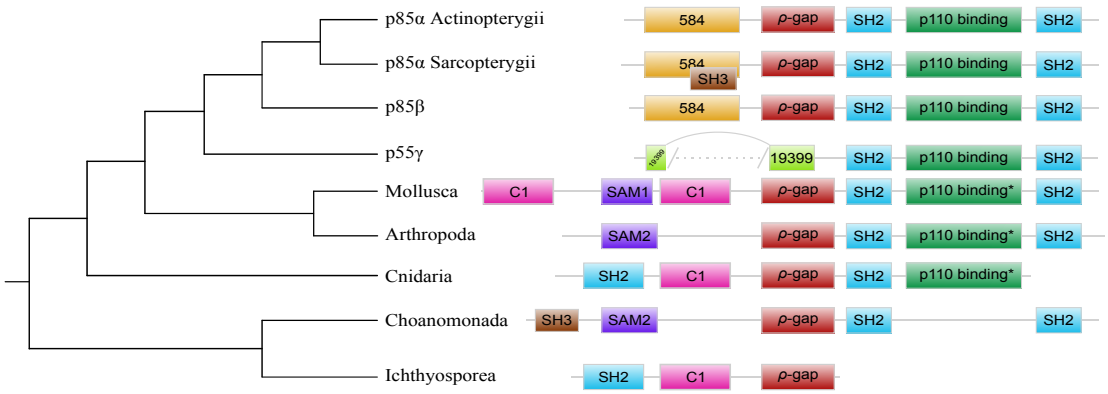

FIGURE *4.13* – ***Représentation de la composition en domaines des protéines régulatrices de la classe IA.*** L'arbre à gauche correspond à la représentation schématique de l'arbre de la Figure 4.12. Les tailles des rectangles ne sont pas proportionnelles à la taille réelle des domaines.

Trois exceptions sont les homologues Gnathostomata de la protéine p55γ qui ne possèdent pas le domaine ρ-gap et les protéines courtes issues des Ichthyosporea et des Choanoflagellida qui ne contiennent pas de domaine p110-binding. Ce résultat est étonnant étant donné que des protéines catalytiques de la classe IA ont été détectées chez les Ichthyosporea et chez les Choanoflagellida (protéine ancestrale aux protéines p110α-β-δ). Ainsi, dans ces espèces, soit le domaine p110-binding n'est pas requis pour la formation de l'hétérodimère de classe IA soit l'interaction entre la sous-unité catalytique et la sous-unité régulatrice n'existe pas. Néanmoins,

les Ichthyosporea n'étant représentés que par une seule séquence et les protéines issues des Choanoflagellida étant courtes (et sans doute partielles), il n'est pas possible de déterminer si cette absence est factuelle ou non.

En plus de ces trois domaines (ρ-gap, SH2 et PB011403), d'autres domaines sont présents en position N-terminale de certaines séquences. Ainsi, les protéines p85α et p85β chez les Gnathostomata possèdent le domaine PB000584 tandis que la copie présente chez les Mollusca, Cnidaria et Ichtyosporea possède un domaine C1 (PF00130) à cette position. Un domaine SAM_1 (PF00536) ou SAM_2 (PF07647) a également été détecté chez les Mollusca, Arthropoda et Choanoflagellida tandis que les Cnidaria et l'Ichtyosporea possèdent un domaine SH2. Enfin, un domaine SH3 (PF00018) est présent dans les deux séquences des Choanoflagellida et chez les homologues de p85α issus d'espèce d'Actinopterygii.

FIGURE *4.14* – ***Arbre phylogénétique des protéines régulatrices de la classe IB.*** L'arbre a été inféré par maximum de vraisemblance, avec le modèle évolutif JTT+Γ_4 (599 sites conservés, 67 séquences). Les symboles utilisés sont identiques à ceux de la Figure 4.5.

Classe IB

Pour finir, la recherche de similarité de séquences des deux protéines régulatrices de la classe IB (p87 et p101) a permis d'identifier 67 homologues issus uniquement de vertébrés (Figure 4.14). Alors que deux copies sont détectées chez les Chondrichthyes, les Sarcopterygii et les Actinopterygii, une seule copie est présente dans le protéome du seul Petromyzontidae disponible (*P. marinus*). L'émergence de la classe IB régulatrice date donc du dernier ancêtre commun des Vertébrés et a subit une duplication spécifique à la base des Gnathostomata.

La recherche des domaines protéiques parmi les 67 homologues de la classe IB régulatrice n'a révélé la présence que d'un seul domaine nommé PI3K_1B_p101 (PF10486) et n'apporte donc aucune information à l'analyse phylogénétique.

4.3 Conclusions

4.3.1 Comparaison avec les phylogénies précédentes

Comme détaillé en première partie de ce chapitre, seules deux phylogénies incomplètes des PI3K étaient jusqu'alors disponibles [289, 358]. En accord avec ces deux études, j'ai identifié deux duplications majeures ayant affecté l'histoire évolutive des sous-unités catalytiques. Comme Brown et Auger, j'ai déterminé que la première duplication, à l'origine de la séparation des classes III et I/II, a eu lieu avant LECA. Si l'étude de Kawashima *et al.* avait également permis d'identifier cette duplication, leur recherche restreinte aux Opisthokonta (et composée uniquement de cinq espèces) ne leur permettait pas de dater précisément cet événement évolutif.

La distribution taxonomique combinée aux phylogénies des classes I et II catalytiques suggèrent que ces deux classes sont originaires des lignées Unikonta tandis que les Bikonta possèdent une forme « ancestrale »de ces deux classes. Néanmoins, dans mon étude comme dans celle de Brown et Auger (Figures 4.5 et 4.4), les séquences de Bikonta se groupent avec les homologues de la classe I et non à la base de la duplication (Figure 4.15). Contrairement à leur phylogénie, cette position n'est pas statistiquement soutenue dans mes arbres (BS = 7% et SH = 0.63). Tous les homologues des classes I et II non Opisthokonta possédant la même organisation en domaines (Figure 4.7), je n'ai pas pu déterminer si l'hypothèse d'une duplication

chez les Unikonta est plus vraisemblable qu'une duplication antérieure à LECA, suivi d'une perte chez le dernier ancêtre commun de Bikonta. Si la première hypothèse s'avère la bonne, il est probable que la protéine catalytique de classe I/II présente chez LECA possédait les mêmes fonctions biologiques que la copie présente chez les Bikonta. Des études fonctionnelles sur des organismes modèles tels que des *Leishmania* ou des *Paramecium* seraient nécessaires pour apporter des éléments de réponse à ces questions.

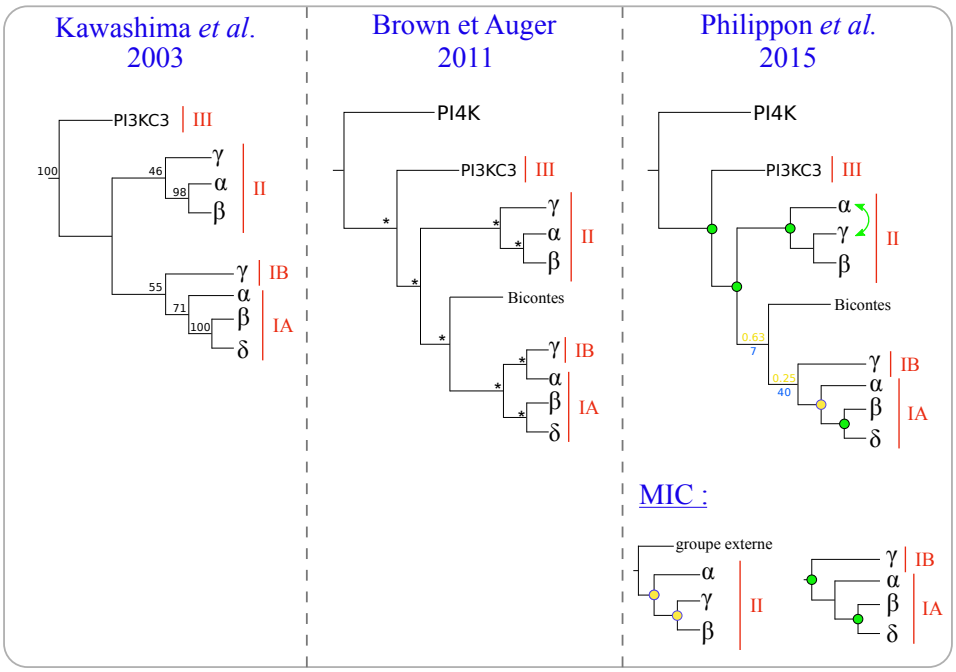

FIGURE *4.15* – ***Comparaison entre les deux phylogénies disponibles pour les protéines catalytiques et notre étude.*** Les valeurs de SH et de BS sont données respectivement en jaune et en bleu.

Au sein de la classe II, les deux phylogénies précédentes soutiennent une première duplication à l'origine de la divergence de PI3K-C2γ suivie d'une seconde duplication à l'origine des protéines PI3K-C2α et PI3K-C2β. Bien que statistiquement soutenu dans ces analyses, ce scénario évolutif est différent de celui suggéré par mes arbres phylogénétiques dans lesquels les protéines PI3K-C2β et PI3K-C2γ sont évolutivement plus proches. A noter que cette hypothèse est d'autant plus soutenue dans l'arbre MIC des protéines catalytiques de classe II (Figure 4.8).

Enfin, en ce qui concerne la classe I, si la phylogénie de Kawashima *et al.* rapporte une monophylie du sous-groupe IA non soutenue, la phylogénie de Brown et Auger regroupe les protéines p110γ et p110α d'un côté (statistiquement soutenu)

et les protéines p110β et p110δ de l'autre (également statistiquement soutenu). Dans mon analyse, l'arbre inféré à partir de l'ensemble des homologues PI3K (Figure B.1), celui du jeu de donnée réduit inféré par maximum de vraisemblance (Figure 4.5) ainsi que celui inféré par méthode bayésienne (Figure 4.6), la monophylie (très soutenue) du sous-groupe IA est retrouvée. De plus, contrairement aux deux phylogénies publiées, mes analyses ont été effectuées à l'aide de méthodes d'inférence plus récentes et plus précises. Par ailleurs, j'ai constitué des jeux de données beaucoup plus importants (homologues issus de 117 espèces Opisthokonta contre moins de 30 auparavant). Enfin, l'étude de la composition en domaine de ces protéines (Figure 4.7), soutient l'hypothèse d'une parenté proche des protéines IA puisqu'elles possèdent toutes le domaine p85-binding contrairement à la protéine catalytique de la classe IB (p110γ).

Au niveau des protéines régulatrices, seules les phylogénies des classes IA et III avaient été précédemment inférées [358]. En accord avec Kawashima *et al.*, je n'ai détecté aucun événement de duplication pour la protéine régulatrice de classe III (Figure 4.10). Pour la classe IA, leur phylogénie construite à partir de seulement dix séquences, place les homologues de la protéine p85β (PIK3R2) à la base du groupe formé par les homologues des protéines p85α (PIK3R1) et p55γ (PIK3R3). Ma phylogénie (Figure 4.12) soutient quant à elle une première duplication à l'origine de la divergence de p55γ, suivie par une duplication (non soutenue) à l'origine des sous-unités p85α et p85β.

4.3.2 Une histoire complexe

Les PI3K sont des enzymes au rôle clef dans de nombreuses voies de signalisation et processus cellulaires nécessaires au bon fonctionnement de la cellule. Comme montré précédemment, il apparaît que cette famille est très ancienne et est retrouvée chez l'ensemble des Eucaryotes actuels, excepté les algues rouges. Néanmoins, l'histoire évolutive de cette famille est très complexe et de nombreux événements de duplication et pertes de gènes ont eu lieu au cours de la diversification des Eucaryotes. De plus, l'analyse de la composition en domaine de toutes ces protéines homologues ainsi que les données fonctionnelles disponibles suggèrent que ces différents événements évolutifs se sont accompagnés de changements fonctionnels.

L'ensemble des résultats sont résumés par la Figure 4.16. L'arbre eucaryote de référence situé à droite a été réalisé en accord avec les études de Adl *et al.* [8], de Delsuc *et al.* [379] et du livre de Lecointre et Le Guyader [2]. Les couleurs

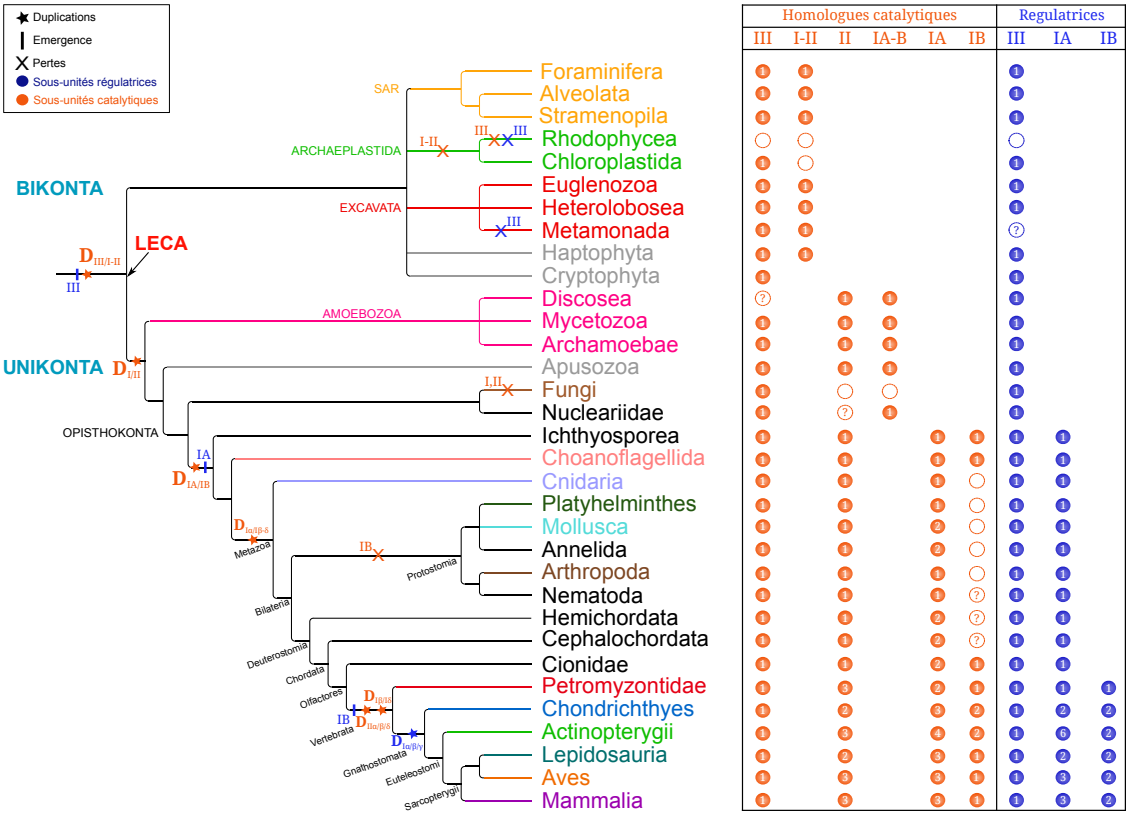

FIGURE *4.16* – ***Représentation de l'histoire évolutives des PI3K au cours de l'évolution des Eucaryotes.***

des groupes taxonomiques sont, quant à elles, identiques à celles utilisées dans la Figure 4.5. Pour chaque catégorie de PI3K, le nombre de copies présentes est indiqué dans les cercles colorés (tableau de droite). Les cercles en pointillés symbolisent pour leur part les pertes secondaires. Enfin, les points d'interrogation indiquent une absence supposée d'homologues au sein des groupes taxonomiques pour lesquels un seul protéome complet est disponible dans notre banque de données.

Les PI3K de LECA

L'analyse phylogénétique de la famille des PI3K a démontré que l'ancêtre commun des eucaryotes possédaient déjà les deux protéines (catalytique et régulatrice) de la classe III. Chez l'Homme et la levure, l'hétérodimère formé par ces deux protéines joue un rôle majeur dans l'induction de la formation des autophagosomes [333, 334, 335], première étape de la macro-autophagie. Ainsi, on peut supposer que ce mécanisme était déjà en place chez LECA même si des analyses plus poussées sur les autres composants de cette voie de signalisation sont nécessaires.

Par ailleurs, j'ai déterminé que la séparation entre les classe III et I/II est très ancienne et a eu lieu durant l'eucaryogenèse. Selon le scénario évolutif considéré pour la séparation des classes I et II, le génome de LECA codait donc également pour une ou deux protéines catalytiques homologues aux protéines des classes I et II humaines.

Des pertes majeures

De manière intéressante, aucune protéine PI3K (catalytique ou régulatrice) n'a été détectée chez les algues rouges, ce qui suggère trois pertes indépendantes dans cette lignée (Figure 4.16). Cependant, parmi les trois protéomes complets d'algues rouges de notre banque de séquences, le génome de *Chondrus crispus* est connu pour avoir une structure spécifique [380] tandis que les génomes de *Cyanidioschyzon merolae* et *Galdieria sulphuraria* sont vraiment très petits pour des génomes eucaryotes [381, 382]. Des analyses complémentaires sur des nouvelles données génomiques de cette lignée sont donc nécessaires pour confirmer ces conclusions.

Par ailleurs, en accord avec des études précédentes, j'ai confirmé que *S. cerevisiae* ne possède que les deux protéines (catalytique et régulatrice) de la classe III [346, 289]. Cette observation a cependant été étendue à l'ensemble des champignons et mon analyse a montré que deux pertes indépendantes des classes I et II avaient eu lieu dans cette lignée eucaryote.

De manière similaire, aucune protéine de classe I ou II n'a été détectée chez les Archaeplastida à l'exception d'une séquence de classe I catalytique chez la plante *S. moellendorffii*. Cette espèce n'étant pas présente dans le jeu de données de Brown et Auger, les auteurs ont conclut a une absence totale d'homologues de classes I et II chez des plantes. Trois hypothèses peuvent expliquer la présence de ce gène chez *S. moellendorffii* : i) des pertes de gènes multiples et indépendantes au sein des Archaeplastida sauf dans la lignée de *S. moellendorffii*, ii) une perte ancienne chez l'ancêtre commun Archaeplastida suivi d'une réacquisition par transfert horizontal chez cette plante, ou iii) la séquence détectée provient d'une contamination lors du séquençage. Afin de tester cette dernière hypothèse, j'ai effectué une recherche BLASTp pour identifier les homologues des six gènes situés à proximité du gène de *S. moellendorffii*. L'ensemble des meilleurs *hits* provenant de séquences de plantes, cette troisième hypothèse a été écartée. Si la première hypothèse est moins parcimonieuse que la seconde, il n'y a néanmoins pas assez d'éléments pour départager avec certitude ces deux hypothèses.

La diversification des classes I et II

L'histoire évolutive des classes I et II a été ponctuée de multiples événements de duplications et pertes de gènes. Au niveau de la classe I catalytique, cette analyse a permis de déterminer qu'une première duplication, à l'origine de la divergence entre les sous-groupes IA et IB, a eu lieu avant la divergence des MIC. Néanmoins, à cause d'un faible support statistique et d'une faible quantité de génomes complets disponibles pour les Nucleariidae et Apusozoa (un seul de chaque dans notre banque de donnée), je n'ai pas pu dater plus précisément cette duplication. On ne peut donc pas déterminer si le dernier ancêtre commun des Opisthokonta possédait une ou deux copies de la classe I catalytique. En terme de pertes de gènes, une duplication Opisthokonta specifique implique deux pertes indépendantes chez les champignons ainsi qu'un perte chez les Nucleariidae et Apusozoa. A contrario, une duplication MIC spécifique n'implique qu'un seule perte chez les champignons mais suppose un mauvais placement de la séquence de Nucleariidae.

Au sien de la classe IA catalytique, deux duplications successives ont eu lieu. La première chez le dernier ancêtre commun des Metazoa tandis que la seconde, à l'origine des protéines p110β et p110δ est spécifique des Vertebrata.

De façon très intéressante, l'émergence de la classe régulatrice IA coïncide parfaitement avec la divergence de la classe IA catalytique. En effet, ces deux groupes de protéines sont détectées chez l'ensemble des MIC (Figures 4.9 et 4.12). On peut donc supposer que, tout comme chez l'Homme actuellement, les deux sous-unités présentes chez le dernier ancêtre commun des MIC fonctionnaient également en hétérodimères. La duplication du gène codant pour la sous-unité catalytique ayant eu lieu avant la duplication du gène codant la protéine régulatrice (chez les Gnathostomata et Metazoa respectivement) ; on peut se demander si les deux sous-unités catalytiques présentes chez les espèces non Gnathostomata (*i.e.* Mollusca, Annelida) sont toutes les deux régulées par la seule protéine régulatrice codée par leur génome ou par une autre protéine non encore caractérisée.

Au contraire, les sous-unités régulatrices et catalytique de la classe IB n'ont pas émergé au même moment. En effet, les deux protéines régulatrices (p87 et p101) sont beaucoup plus récentes et uniquement présentes chez les espèces Gnathostomata, tandis que les MIC possèdent la sous-unité catalytique (p110γ). On peut donc se demander si et comment sont régulés les homologues catalytiques de p110γ dans les autres animaux, les Choanoflagellida et les Ichthyosporea.

Enfin, en ce qui concerne la classe II, j'ai pu déterminer que son expansion est relativement récente. En effet, les deux duplications à l'origine des trois protéines

humaines, sont postérieures à la divergence Vertebrata-Olfactores. Malheureusement les phylogénies de cette classe (Figures B.1 et 4.8) ainsi que la caractérisation de la composition en domaines de ces homologues (Figures 4.7) n'a pas permis d'apporter d'information supplémentaire quant à leur fonction biologique. On peut seulement supposer que la présence des domaines PB011861 et PB008942 implique des fonctions différentes pour les protéines PI3K-C2α et PI3K-C2β.

Hypothèse sur la fonction ancestrale de la classe I

Si de nombreuses études ont démontré l'implication de la classe I dans la chimiotaxie des organismes unicellulaires du genre *Dyctiostelium* [350, 282, 283] ; il est connu que chez l'Homme, la classe IB permet la motilité de différentes cellules telles que les leucocytes [309, 383, 384, 352] ou les cellules T CD4$^+$ [306]. Au contraire, les protéines humaines de la classe IA régulent plutôt la mitose [385], la croissance et la prolifération cellulaires [346]. En accord avec ces observations et l'analyse phylogénétique de la classe I (Figure 4.5), on peut supposer que la copie ancestrale à la duplication à l'origine des classes IA et IB était impliquée dans la chimiotaxie des cellules et que la copie IB a conservé cette fonction. Si tel est le cas, on peut également se demander si cette duplication, qui a eu lieu chez les Opisthokonta, est directement lié au passage à la multicellularité.

4.3.3 Conclusion générale

Les PI3K forment une famille ancienne, diverse et complexe. Grâce à un grand nombre de données disponibles combinées à des méthodologies récentes, cette analyse a permis de reconstruire une histoire évolutive plus complète que celles précédemment inférées. Néanmoins, certaines parties des arbres phylogénétiques obtenus sont peu résolues et certains événements de duplication n'ont pas pu être datés avec précision. L'ajout de génomes complets issus de certains groupes eucaryotes tels que les algues rouges, les Ichthyosporea ou encore les Nucleariidae permettront peut-être de lever les dernières zones d'ombres de l'histoire de cette famille.

Chapitre 5

Histoire évolutive de la voie de signalisation AKT/mTOR

L'étude de l'histoire évolutive des PI3K (chapitre 4) m'a permis d'établir une méthodologie pouvant être appliquée aux voies de signalisation. Avec cette première analyse détaillée, j'ai pu identifier certains problèmes soulevés par les reconstructions phylogénétiques et expérimenter les différents logiciels existants. Ce chapitre concerne l'étude de la mise en place, durant l'évolution des Eucaryotes, de la voie AKT/mTOR qui est essentiellement impliquée dans l'induction de l'autophagie. Cette analyse fut l'occasion de démarrer le développement d'EPINe (***E**asy **P**hylogenetics for **I**nteraction **Ne**tworks*), un pipeline destiné à l'analyse de l'histoire évolutive des voies de signalisation eucaryotes.

5.1 L'autophagie

5.1.1 Introduction

Le terme d'autophagie, dont la traduction grecque littérale signifie « manger soi-même », a été introduit par Duve en 1963 [386, 387]. A l'origine considérée comme un simple mécanisme de dégradation des composés cellulaires, de nombreuses études récentes ont au contraire démontré qu'il s'agit d'un processus clef situé au carrefour entre la mort programmée, la survie et la prolifération cellulaire [386, 388, 389].

Depuis les années 2000, l'autophagie a été mise en évidence comme dérégulée dans de nombreuses pathologies humaines telles que les maladies du foie [390, 391], les maladies hématologiques (myélomes [392], leucémies [393], etc.) ou encore dans divers cancers (cancer colorectal [394], carcinome épidermoïde [395], cancer du sein [396], etc.). En quelques années, l'autophagie est ainsi devenue l'un des processus les plus ciblés par le développement de nouveaux médicaments [397, 398] et de nombreuses données sont ainsi disponibles.

5.1.2 Mécanisme

Si la cellule synthétise divers composés (protéines, lipides, etc.), il lui est indispensable de maintenir son homéostasie en éliminant les éléments qui ne lui sont plus utiles ou qui se sont détériorés au cours du temps. Avec le système protéasomes-ubiquitine, l'autophagie constitue le second mécanisme majeur de dégradation des composés intracellulaires des cellules eucaryotes [399].

Chez les mammifères, le processus d'autophagie se décompose en trois grandes étapes. Après l'identification des composés à détruire, la seconde étape consiste à transporter ces derniers jusqu'aux lysosomes, qui, en dernier lieu, se chargeront de les dégrader et de les recycler [400]. Par ailleurs, on distingue trois grandes catégories d'autophagie : i) l'autophagie par protéines chaperones, ii) la microautophagie et iii) la macroautophagie (Figure 5.1). Dans ce chapitre, seule la macroautophagie (ci-après nommée simplement autophagie) sera considérée.

Figure omise pour des raisons de droits d'auteur.

FIGURE *5.1* – ***Les trois autophagies des cellules de mammifères [400].***

La macroautophagie se distingue principalement des autres catégories d'autophagie par l'utilisation de vésicules de transport nommées *autophagosomes*. Composées d'une double membrane phospholipidique, elles permettent de transporter les composés intracellulaires (aussi nommés *cargo*) jusqu'au lysosome. La fusion de ces deux entités génère les *autolysosomes*, au sein desquels le cargo ainsi que la double membrane des autophagosomes sont dégradés. Les produits issus de cette dégradation, tels que les acides aminés ou les acides gras, sont enfin relâchés dans le cytoplasme cellulaire grâce à des perméases lysosomales [386]. Ainsi, l'autophagie est un processus clef pour le bon fonctionnement de la cellule, notamment en cas de manque de nutriments où elle devient cruciale pour la survie de la cellule.

Figure omise pour des raisons de droits d'auteur.

FIGURE *5.2* – ***Rôle des gènes ATG dans l'induction, la formation et l'élongation des autophagosomes [393].***

L'identification des protéines impliquées dans l'autophagie commença dès 1993 avec l'étude de Tsukada et Oshumi [401] qui a permis de caractériser quinze gènes chez la levure. De nombreux autres gènes ont ensuite été découverts dans différentes espèces. Devant l'hétérogénéité des noms qui leur ont été donnés, Klionsky *et al.* [402] proposèrent en 2003 une nomenclature unique dans laquelle ces gènes sont désignés par l'acronyme ATG (***A****u****T****opha****G****y related genes*) suivi d'un nombre distinctif. De leur côté, les protéines codées par ces gènes sont désignées par l'acronyme Atg. On dénombre ainsi plus de 37 gènes ATG chez la levure [403]. Bien qu'essentiellement impliquées dans la phase d'élongation des autophagosomes (Figure 5.2), certaines protéines telles que Atg14 et Atg6 (aussi connue sous le nom Beclin1) sont à l'origine de leur initiation.

5.1.3 L'autophagie chez les Eucaryotes

Si chez les Opisthokonta les processus d'autophagie observés jusqu'à maintenant sont très semblables, plusieurs différences ont été mises en évidence chez les plantes [404, 405]. La principale concerne les vésicules impliquées dans ce processus. Chez la levure, l'Homme ou encore le nématode, les hydrolases responsables de la dégradation du cargo se trouvent à l'intérieur des lysosomes. Comme expliqué ci-dessus, chez ces espèces, les autophagosomes doivent donc transporter le cargo jusqu'à un lysosome afin d'induire sa dégradation. Au contraire, une étude récente [404] a démontré que cette phase de transport n'existe pas chez les plantes. En effet, la dégradation se fait directement au sein de la première vésicule qui possède des hydrolases. Nommées à tort autophagosomes, les auteurs suggèrent donc de les appeler autolysosomes ou *autophagosome-like*.

Par ailleurs, bien que la levure, le nématode et l'Homme présentent de grandes similitudes fonctionnelles pour ce processus, certains gènes ATG de la levure, tels que ATG29, ATG31 et ATG17, ne possèdent pas d'homologues chez l'Homme [406]. De manière surprenante, l'interaction de RBCC1 avec Atg1 (ou ULK1) chez l'Homme est indispensable à son fonctionnement, tout comme l'est celle entre Atg17 et Atg1 chez la levure [407]. Pourtant, RBCC1 et Atg17 ne présentent pas de similarité de séquence. Enfin, chez le nématode, quelques gènes spécifiques tels que epg-2 et epg-7 ont été identifiés. Ces gènes ne semblent posséder d'homologues ni chez l'Homme ni chez la levure [408].

Plutôt que de s'intéresser aux gènes ATG, déjà très étudiés et pour lesquels la similarité de séquence n'est parfois pas un critère adapté, j'ai décidé de me focaliser sur l'activation de l'autophagie par la voie de signalisation AKT/mTOR (aussi nommée PI3K/AKT/mTOR). Ce processus ayant été détecté chez les plantes, les champignons et les mammifères, cette étude a été effectuée en considérant l'ensemble des Eucaryotes. De plus, la mise en place d'une voie de signalisation concernant l'émergence de ses composés mais également la conservation de leurs interactions (voir chapitre 1.2.2), cette analyse a inclus une comparaison entre l'interactome de l'Homme et celui de la levure.

5.1.4 La voie AKT/mTOR chez l'Homme

La première étape de cette analyse a consisté à inventorier l'ensemble des protéines et des interactions de la voie de signalisation AKT/mTOR. Les données chez l'Homme étant nombreuses et de bonne qualité, j'ai donc choisi cet organisme comme point de départ.

En collaboration avec Evelyne Goillot [1], cette recherche bibliographique m'a permis d'élaborer la Figure 5.3 où les numéros associés aux interactions correspondent aux articles listés dans la Table C.1 (en annexe). Au total, 62 protéines humaines ont été identifiées comme impliquées dans cette voie de signalisation humaine (Table C.2). Parmi ces dernières, onze correspondent aux PI3K des classes I et III.

[1] Laboratoire BMC, UMR CNRS 5239, ENS Lyon

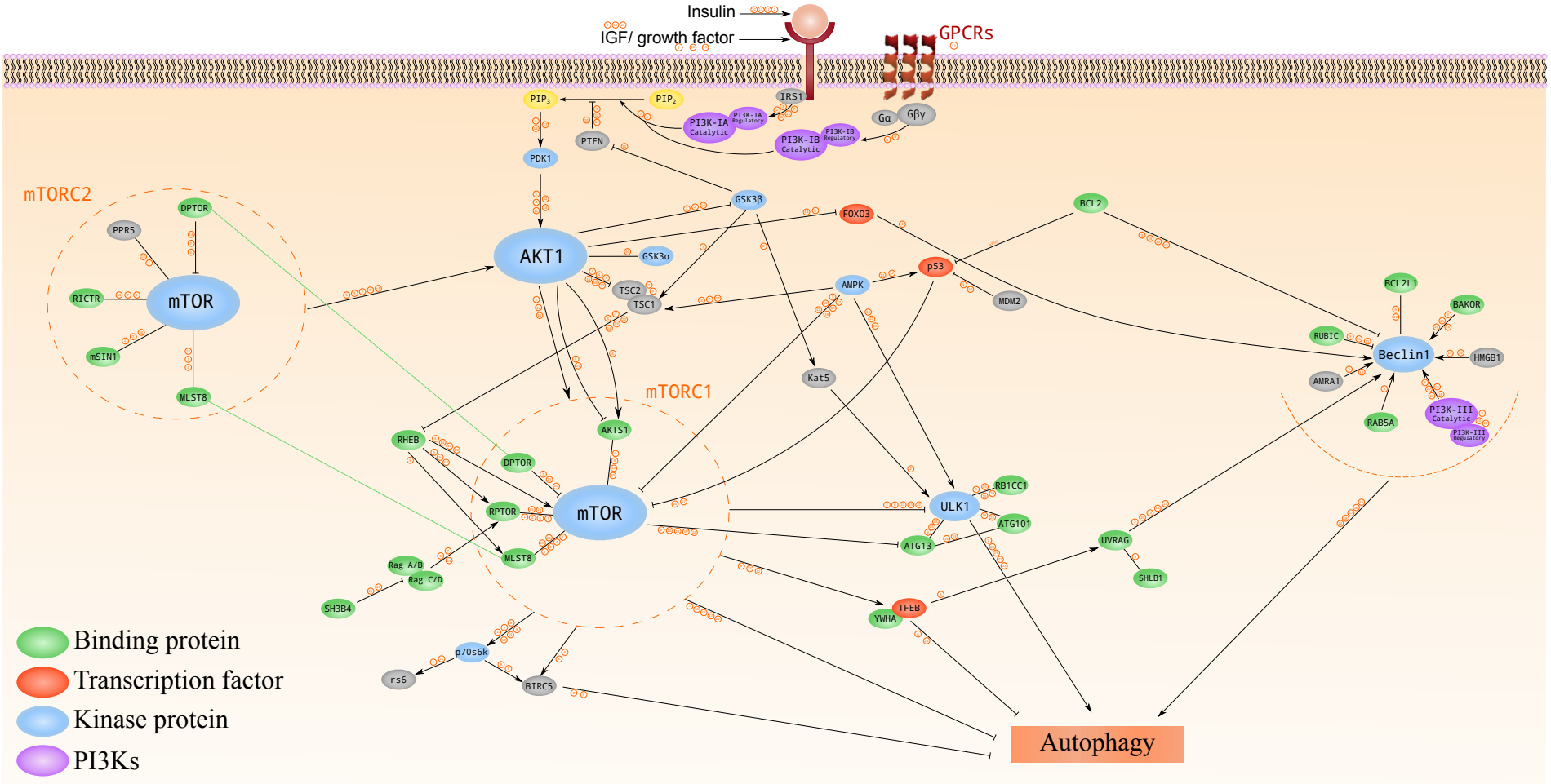

FIGURE *5.3* – ***Schéma de la voie AKT/mTOR chez l'Homme.*** Chaque numéro entouré correspond à un article de la Table C.1.

5.2 Reconstruction automatique d'un ensemble d'arbres de gènes

5.2.1 Principe général

L'étude de l'histoire évolutive d'un réseau d'interactions ou d'une voie métabolique implique la reconstruction des arbres de gènes de chacune des protéines le composant. Dans ce cadre, j'ai mis en place le pipeline EPINe qui permet de reconstruire automatiquement les arbres de gènes à partir d'une liste d'identifiants UniProtKB. A terme, ce pipeline devrait également permettre de dater automatiquement l'émergence de ces protéines ainsi que d'étudier la conservation de leurs interactions (voir Figure 5.4 et perspectives développées dans le chapitre 6).

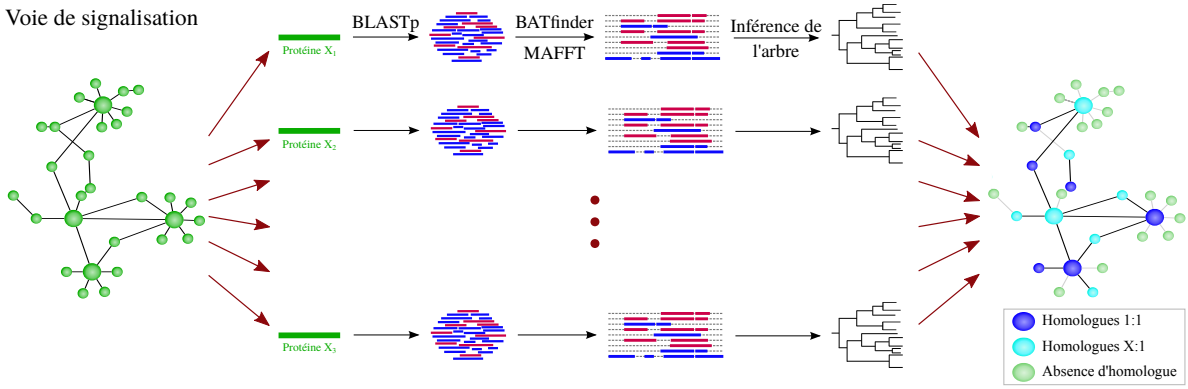

FIGURE 5.4 – ***Première étape d'EPINe : inférence des arbres phylogénétiques.***

La première étape de l'analyse a donc été l'inférence des arbres phylogénétiques des 62 protéines de la voie de AKT/mTOR. La même méthodologie que celle appliquée aux PI3K et l'utilisation de BATfinder ont été intégrés dans EPINe.

5.2.2 Banques de séquences utilisées

Comme indiqué dans le chapitre 2 de ce manuscrit, une reconstruction phylogénétique commence par l'identification des séquences homologues à une séquence d'intérêt. Dans ce but, j'ai constitué deux différentes banques de données. La première est composée d'une sélection des séquences protéiques d'Ensembl, et la seconde, de génomes complets de divers eucaryotes issus de GenBank.

Séquences d'*Ensembl*

Comme pour l'étude des PI3K, seule une partie des 87 espèces présentes dans la version 80 d'Ensembl ont été sélectionnées afin d'éviter, entre autres, une surreprésentation d'espèces mammifères dans les arbres phylogénétiques inférés. En effet, 31 espèces représentatives de la diversité des Bilateria (Figure 5.5) ainsi que le génome de l'espèce modèle *S. cerevisiae* ont été inclus dans ma banque de données.

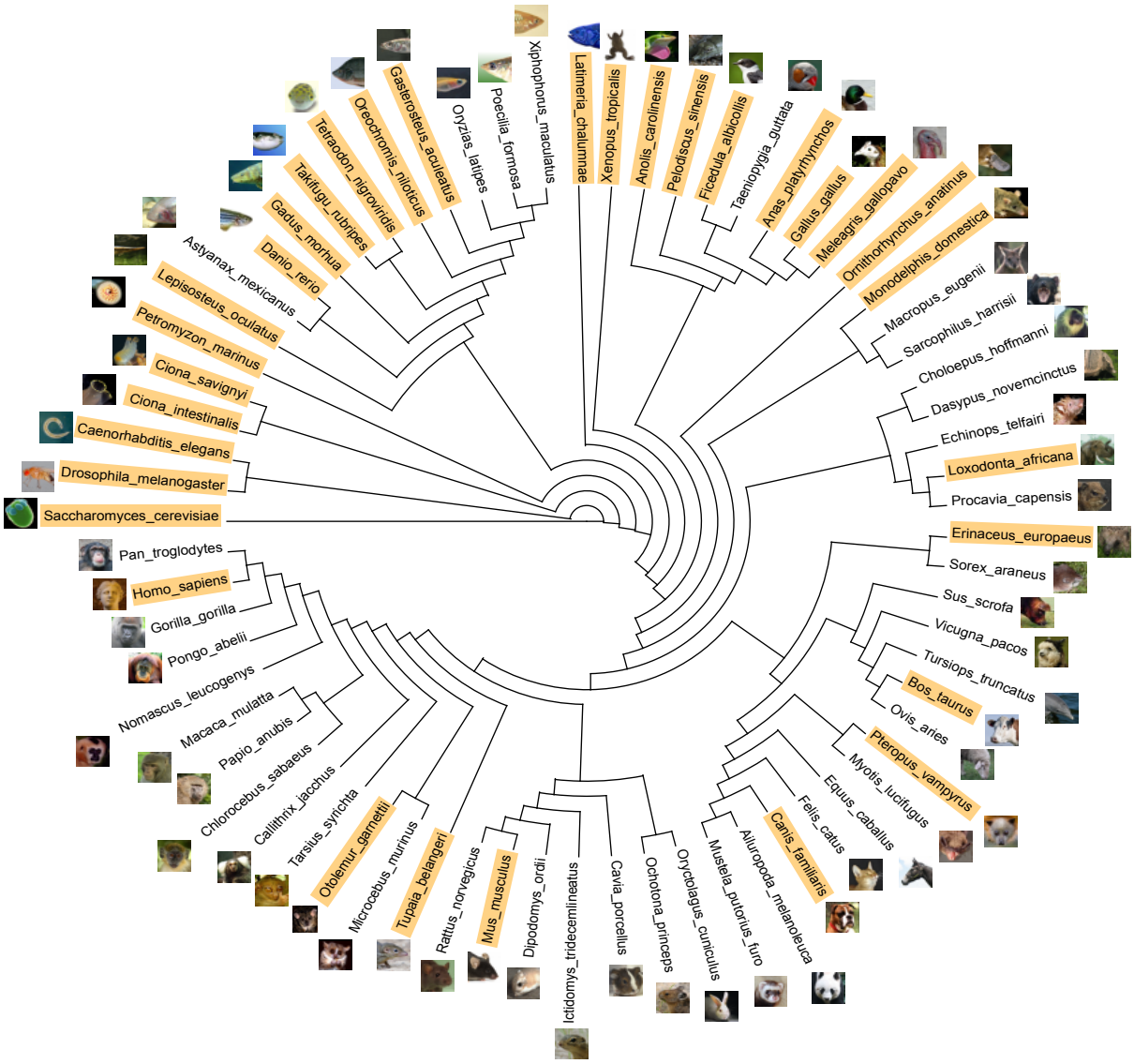

FIGURE 5.5 – ***Arbre des espèces de référence d'Ensembl [198].*** Les 32 espèces sélectionnées pour EPINe sont surlignées en jaune.

Au total, ce sont plus de 824 000 séquences protéiques qui sont codées par l'ensemble de ces 32 génomes eucaryotes (Table 5.1).

Espèce	# seq	Espèce	# seq	Espèce	# seq	Espèce	# seq
A. platyrhynchos	16353	*D. melanogaster*	30362	*L. oculatus*	22483	*P. sinensis*	20669
A. carolinensis	19177	*E. europaeus*	14605	*L. africana*	25635	*P. marinus*	11442
B. taurus	22118	*F. albicollis*	15983	*M. gallopavo*	16494	*P.vampyrus*	17053
C. elegans	30939	*G. morhua*	22100	*M. domestica*	22310	*S. cerevisiae*	6692
C.familiaris	25160	*G. gallus*	16354	*M. musculus*	54883	*T.rubripes*	47841
C. intestinalis	17302	*G. aculeatus*	27576	*O. niloticus*	26763	*T. nigroviridis*	23118
C. savignyi	20155	*H.sapiens*	100778	*O. anatinus*	23584	*T. belangeri*	15475
D. rerio	44487	*L. chalumnae*	23601	*O. garnettii*	19986	*X. tropicalis*	22718

Total : 824196

TABLE *5.1* – ***Nombre de séquences protéiques par espèce d'Ensembl sélectionnée.*** « # seq » : nombre de séquences.

Parmi ces séquences se trouvent des transcrits alternatifs, rendant indispensable l'étape de sélection par BATfinder. Afin de générer automatiquement le fichier contenant l'information liant la séquence protéique à son gène, j'ai renommé (automatiquement) chacune de ces séquences. Leurs noms contiennent ainsi l'identifiant du transcrit suivi du nom de l'espèce et de l'identifiant du gène, comme le montre l'exemple ci-dessous :

```
>ENSAPLT00000000572 Anas_platyrhynchos ENSAPLG00000000568
MSKTSQQNTATDANGVSVIHTQAHTSGLQQVPQLVPVSPGGGGKAVPPSKQGKKNSFVDR
NSDEYRQRRERNNMAVKKSRLKSKQKAQDTLQRVTQLKEENERLEAKIKLLTKELSVLKD
LFLEHAHNLADNVQPVGTETTTTNPENSGQ
```

où `ENSAPLT00000000572` est l'identifiant du transcrit et `ENSAPLG00000000568` celui du gène.

Ainsi, pour l'exécution de BATfinder, il suffit de récupérer le nom des séquences identifiées comme homologues pour générer le fichier liant la séquence à son gène. Cette étape m'a également permis d'homogénéiser les noms des séquences et d'introduire celui de l'espèce.

Génomes complets eucaryotes

Parmi les génomes eucaryotes disponibles dans GenBank en Mars 2015, Céline Brochier-Armanet, Lily Bickerstaffe et moi avons sélectionné 84 génomes complets dont nous avons récupéré les séquences protéiques correspondantes afin d'obtenir une banque de données composée de plus de 1,5 millions de séquences (Table 5.2).

Espèce	# seq	Espèce	# seq	Espèce	# seq	Espèce	# seq
A. castellanii	15014	*C. gigas*	27900	*L. gigantea*	23877	*R. filosa*	39963
A. macrogynus	19446	*C. neoforman*	7826	*M. oryzae*	12836	*R. irregularis*	30312
A. queenslandica	9914	*C. merolae*	4803	*M. pusilla*	10269	*R. delemar*	17574
A. deanei	16888	*D. purpureum*	12399	*M. brevicollis*	9203	*S. kowalevskii*	22093
A. astaci	26259	*E. granulosus*	23963	*M. verticillata*	12660	*S. rosetta*	11731
A. mellifera	21772	*E. siliculosus*	16417	*M. circinelloides*	12368	*S. diclina*	18229
A. californica	26005	*E. aedis*	4211	*N. gruberi*	15711	*S. japonicum*	25276
A. thaliana	35378	*E. tenella*	8599	*N. gaditana*	19601	*S. moellendorffii*	69887
A. oryzae	12074	*E.huxleyi*	38555	*N. parisii*	5387	*S. arctica*	18730
A. anophagefferens	11520	*E. cuniculi*	6676	*N. vectensis*	50384	*S. salmonicida*	8333
B. dendrobatidis	17425	*E. nuttalli*	6187	*N. bombycis*	4764	*S. culicis*	12083
B. graminis	6525	*E. bieneusi*	7297	*O. sativa*	28553	*T. thermophila*	24770
B. floridae	28623	*E. salsugineum*	27741	*O. trifallax*	24578	*T. oceanica*	34642
C. milii	28224	*F. alba*	6309	*P. tetraurelia*	39578	*T. trahens*	10659
C. tropicalis	6254	*G. sulphuraria*	7174	*P. tricornutum*	10408	*T. parva*	4079
C. teleta	32070	*G. intestinalis*	6098	*P. patens*	73923	*T. gondii*	7987
C. owczarzaki	8381	*G. theta*	24822	*P. infestans*	17797	*T. adhaerens*	11520
C. reinhardtii	31871	*H. robusta*	23468	*P. parasitica*	27942	*T. congolense*	5984
C. variabilis	19733	*H. vulgaris*	17795	*P. falciparum*	5337	*T. cruzi*	10847
C. crispus	9807	*L. major*	8316	*P.pallidum*	12367	*V. carteri*	14436
C. picta	18838	*L. mexicana*	8147	*P. yezoensis*	1210	*Z. mays*	62721

Total : 1577333

TABLE *5.2* – ***Nombre de séquences protéiques par espèce de GenBank sélectionnée.*** « # seq » : nombre de séquences.

J'ai ensuite renommé ces séquences à l'aide des E-utilities du NCBI. Comme pour la banque de données réduite d'Ensembl, leurs noms comportent ainsi l'information de leur provenance génomique, le nom de l'espèce mais également la taille de la séquence et sa fonction biologique. Dans l'exemple ci-dessous :

```
>gi110749931 Apis_mellifera 724196 retinal_homeobox_protein_Rx 282aa
MDSQQLVDTASQNSQDIVLPKPASSTPRHSIDAILGLANNKRSHQEMEDNGRDAQENAGENSCN
STGGGSDEELGAGCGDDLNGNSGKKKHRRNRTTFTTYQLHELERAFEKSHYPDVYSREELAMKV
NLPEVRVQVWFQNRRAKWRRQEKMEAARLGLSEYHHPGNMRNVAGPALGLPGDPWLTPPGLLSA
LPGFLAAPHTGYPSYLTSPRRLPSPPNVGAVGSAVPGGLSAGMTSIGSGGHVPAAPPSPPGHDP
RTTSIQALRMRAKEHVESITKGLQMV
```

`gi110749931` correspond à l'identifiant de la séquence, `724196` à celui du gène et `retinal_homeobox_protein_Rx` à la fonction biologique de la protéine codée par ce gène.

Enfin, à partir des deux fichiers des séquences renommées, j'ai créé les deux banques au format BLAST correspondantes.

PSI-BLAST

Pour chaque identifiant UniPprotKB fourni en entrée, EPINe commence par récupérer la séquence protéique associée grâce aux fonctions `esearch` et `efetch`

des E-utilities. L'ensemble des homologues à cette séquence se trouvant dans les deux banques de données décrites ci-dessus sont ensuite identifiés par une recherche PSI-BLAST ($E < 10^{-15}$ et cinq itérations par défaut).

5.2.3 Tri de séquences et reconstruction phylogénétique

Afin de réduire le temps de calcul de l'étape de sélection de transcrits alternatifs par BATfinder, un premier tri est effectué par EPINe. Pour les transcrits alternatifs parfaitement identiques, une seule séquence est gardée. Ensuite, BATfinder est exécuté avec 30 réplicats de *bootstrap* et le logiciel d'alignement MAFFT. Ce dernier étant utilisé en mode automatique, EPINe adapte les options de BATfinder en fonction des caractéristiques du jeu de données de façon à contre balancer le temps de calcul de MAFFT (Table 5.3).

		Taille maximale	
		< 2000 aa	≥ 2000 aa
Nombre d'homologues	< 50	LG	LG
	< 100	LG	JTTfast
	< 150	LG	DS
	< 200	LG	DS
	< 300	JTT	DS
	< 400	JTTfast	DS
	≥ 450	DS	DS

TABLE *5.3* – ***Options de BATfinder utilisées en fonction des caractéristiques du jeu de données.*** L'utilisation de l'option `-DS` réduit considérablement le temps d'exécution (voir chapitre 3).

En effet, cette option de MAFFT (`-auto`) permet de sélectionner l'algorithme d'alignement présentant le meilleur rapport temps de calcul / qualité de l'alignement en fonction des caractéristiques du jeu de données d'entrée. Néanmoins, pour les jeux de données aux limites de ces seuils, le temps d'alignement peut être très important. Par exemple, les protéines SHLB1 et PTEN possèdent respectivement 459 et 511 homologues dont les plus longs font respectivement 2172 et 2208 acides aminés (Table C.2). En mode automatique, MAFFT aligne le premier jeu de données en environ quatorze secondes tandis qu'il met un peu plus de neuf minutes pour le second qui est pourtant composé de moins de séquences. Cela est dû au fait que MAFFT utilise un algorithme plus rapide lorsque le jeu de données comporte plus de 500 séquences. De même, MAFFT adapte l'algorithme en fonction

de la taille de la séquence la plus longue avec un seuil à 1 000 et un seuil à 3 000 acides aminés. Il n'y a néanmoins pas de jeux de données aux alentours de ces seuils possédant un nombre comparable de séquences parmi les 62 protéines de la voie AKT/mTOR pour illustrer concrètement ce cas.

Le fichier filtré obtenu grâce à BATfinder est ensuite aligné à l'aide du logiciel MAFFT, toujours en mode automatique. A partir de l'alignement obtenu, le logiciel BMGE est utilisé pour la sélection des sites conservés. Par défaut, les paramètres utilisés sont la matrice BLOSUM30, une proportion autorisée de gaps de 40% et des régions conservées d'au moins trois acides aminés consécutifs. Ces paramètres correspondent à ceux choisis pour l'étude des PI3K (voir chapitre 4.2.1), suite aux tests de plusieurs jeux de paramètres.

Le modèle évolutif le plus adéquat est ensuite sélectionné à l'aide du logiciel ProtTest. Enfin, l'arbre phylogénétique est inféré par le logiciel PhyML et le modèle évolutif précédemment sélectionné.

Dans le but de faciliter la lecture de l'arbre phylogénétique obtenu pour chaque protéine d'entrée, une simplification des noms ainsi qu'une coloration automatique des séquences selon leur appartenance taxonomique sont effectuées. Dans un premier temps l'arbre est enraciné au poids moyen par SeaView puis transformé en format `.xtg` grâce au logiciel TreeGraph [409]. Les noms de séquences sont ensuite modifiés et coloriés à l'aide d'un script Python. Pour finir, une version bitmap de l'arbre colorié (format `.png`) est créée grâce à TreeGraph.

Exemple de la protéine MDM2

Par exemple, le fichier image produit pour l'arbre de la protéine MDM2 est présenté dans la Figure 5.6(a). Néanmoins, comme on le voit, la barre d'échelle n'est pas présente et l'enracinement au poids moyen ne correspond pas à l'enracinement par groupe externe. Dans ce cas, l'utilisateur peut raciner l'arbre par la séquence de *T. adhaerens* en utilisant le logiciel TreeGraph et le fichier `MDM2_colored.xtg` également fournit par EPINe. L'utilisation de ce fichier permet également à l'utilisateur de modifier le rendu visuel de l'arbre en ne conservant par exemple que les valeurs de *bootstrap* supérieurs à 95% (Figure 5.6(b)).

Dans cet exemple, on voit également l'intérêt de l'insertion du nom du gène ou de sa fonction dans le nom de la séquence. D'un seul coup d'œil, l'utilisateur peut déterminer qu'une duplication chez l'ancêtre commun des Gnathostomata est à l'origine des protéines MDM2 et MDM4.

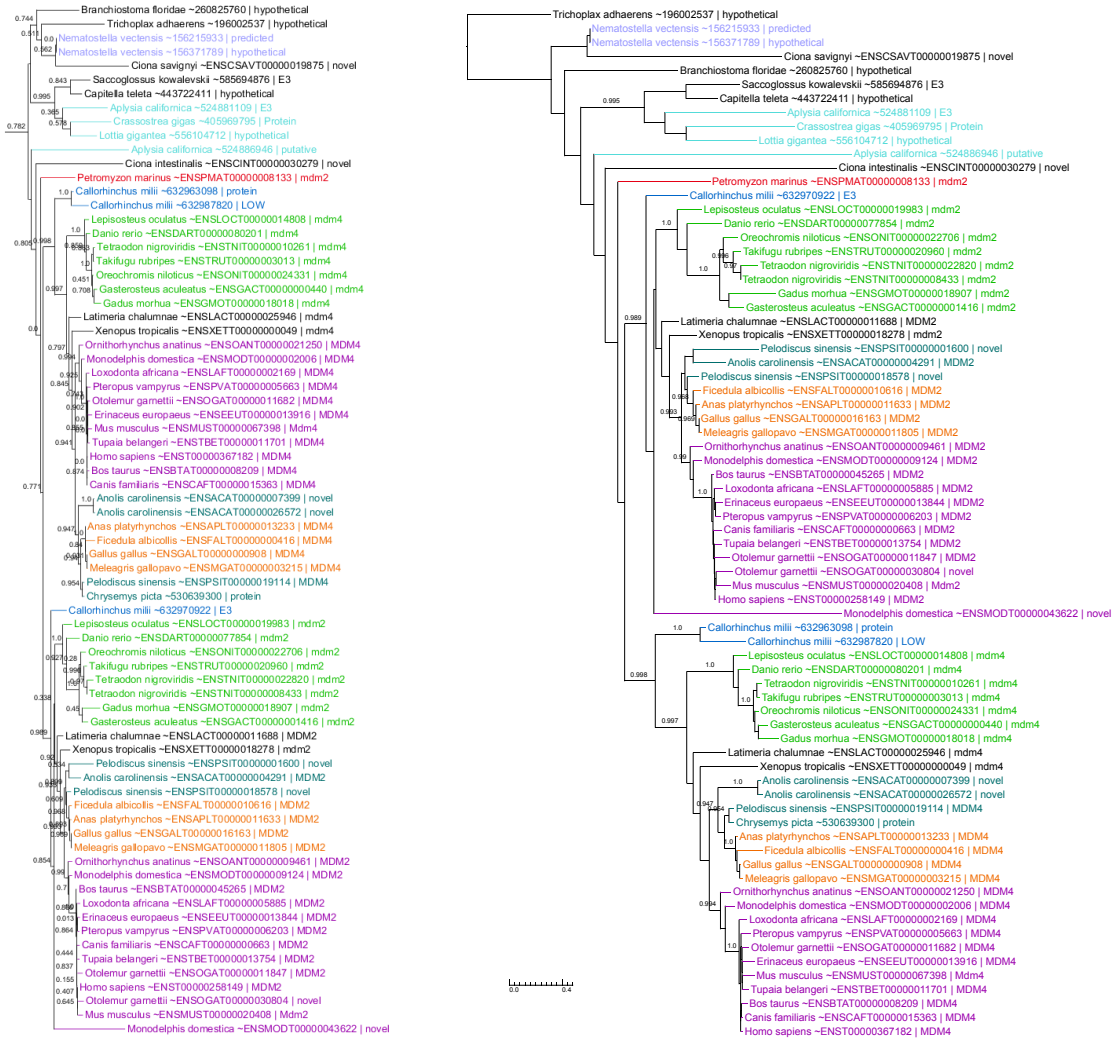

FIGURE 5.6 – ***Arbres phylogénétiques de la protéine MDM2.*** (a) Produit automatiquement sous forme d'image par EPINe, et (b) mis en forme à l'aide de TreeGraph. Les couleurs sont les même que pour l'analyse des PI3K.

5.2.4 Workflow d'EPINe

L'ensemble des étapes et logiciels utilisés par EPINe sont résumés dans la Figure 5.7. Afin d'intégrer les données interactomiques des espèces modèles, j'ai décidé d'utiliser Cytoscape [410] qui est un logiciel dédié à l'analyse de graphes. Il permet notamment de colorier les nœuds et les arêtes en fonction de différents attributs. Dans le cadre d'EPINe, il s'agit par exemple d'indiquer sur la voie de signalisation de départ (celle de l'Homme dans le cas de la voie AKT/mTOR), la présence ou

l'absence d'homologue chez un organisme ou une lignée Eucaryote. A partir des interactomes des organismes modèles, l'utilisation de Cytoscape permettra également de visualiser la conservation des interactions entre les homologues identifiés.

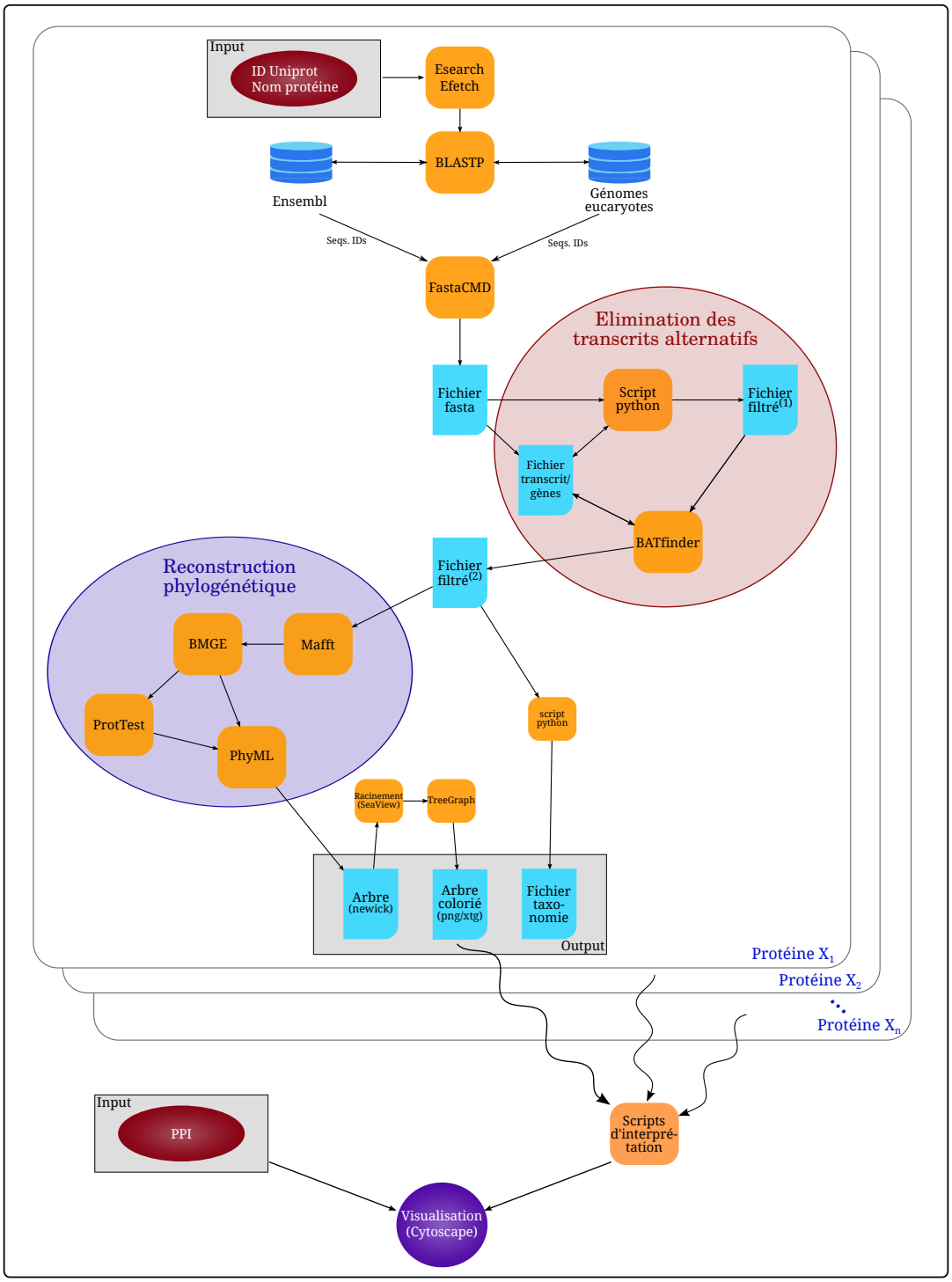

FIGURE *5.7* – ***Enchaînement des programmes utilisés par EPINe.*** (1) Fichier sans transcrits alternatifs identiques et (2) fichier contenant une séquence par locus génomique.

5.3 Principaux résultats pour la voie AKT/mTOR

5.3.1 Analyse phylogénétique

A partir de la liste des 62 identifiants UniProtKB des protéines de la voie AKT/mTOR, j'ai lancé EPINe avec les paramètres présentés dans la table de l'annexe C.2. Le plus petit jeu de données, correspondant aux homologues de la protéine AKTS1, est composé d'uniquement 26 séquences. Au contraire, les protéines AKT1, PDPK1, AAPK1, AAPK2, GSK3A, GSK3B, RAB5A, IGF1R, ULK1 et RHEB possèdent plus de 1000 homologues, ce qui pose des problèmes tant au niveau du temps d'exécution des logiciels qu'au niveau de l'interprétation des arbres phylogénétiques correspondants. Parmi ces séquences, en moyenne 22.37% correspondent à des transcrits alternatifs.

Au niveau de l'alignement, trois différentes options de MAFFT ont été choisies automatiquement : `FFT-NS-2` (pour les 25 plus gros jeux de données), `FFT-NS-i` (pour 25 jeux de données) et `L-INS-i` (pour les 12 jeux de données composés de moins de cent séquences). En ce qui concerne la sélection de sites par BMGE, en moyenne seuls 11.23% des sites ont été conservés. Pour certaines protéines telles que RPTOR ou DEPTOR, moins de 1% des sites ont été sélectionnés. Ceci représentant respectivement 58 et 39 positions pour respectivement 432 et 310 séquences homologues, les arbres obtenus doivent donc être interprétés avec précaution. La meilleure solution restant d'effectuer la reconstruction phylogénétique manuellement et d'adapter les paramètres des logiciels utilisés pour ces jeux de données.

Les modèles évolutifs sélectionnés par ProtTest sont principalement LG+Γ_4 et JTT+Γ_4. L'exécution de ProtTest n'ayant pas abouti pour les protéines possédant plus de 900 homologues, j'ai choisi d'utiliser le modèle LG+Γ_4 pour ces jeux de données.

5.3.2 La voie AKT/mTOR chez différents Eucaryotes.

A partir des arbres phylogénétiques obtenus, j'ai manuellement daté l'émergence de chacune des protéines de la voie AKT/mTOR et déterminé les événements évolutifs qu'elles ont subis de façon analogue à l'analyse des PI3K. Ainsi, il est possible d'inférer la présence ou l'absence d'un gène pour un groupe eucaryote

donné. Les paragraphes suivants concernent la voie telle qu'elle devait être chez LECA, chez le dernier ancêtre commun des Archaeplastida, des Amoebozoa, et enfin des Fungi.

LECA

Parmi les 62 protéines de la voie AKT/mTOR (dont onze sont les PI3K humaines de classes I et III), LECA possédait déjà 18 gènes correspondant à 25 protéines humaines (Figure 5.8). Quatre protéines supplémentaires (SIN1, RBCC1, RHEB et TSC2), étaient peut être codées par le génome de LECA, néanmoins l'absence de séquences de plantes et d'Excavata couplée à l'incertitude du placement de LECA dans l'arbre eucaryote ne nous permet pas de conclure pour ces protéines.

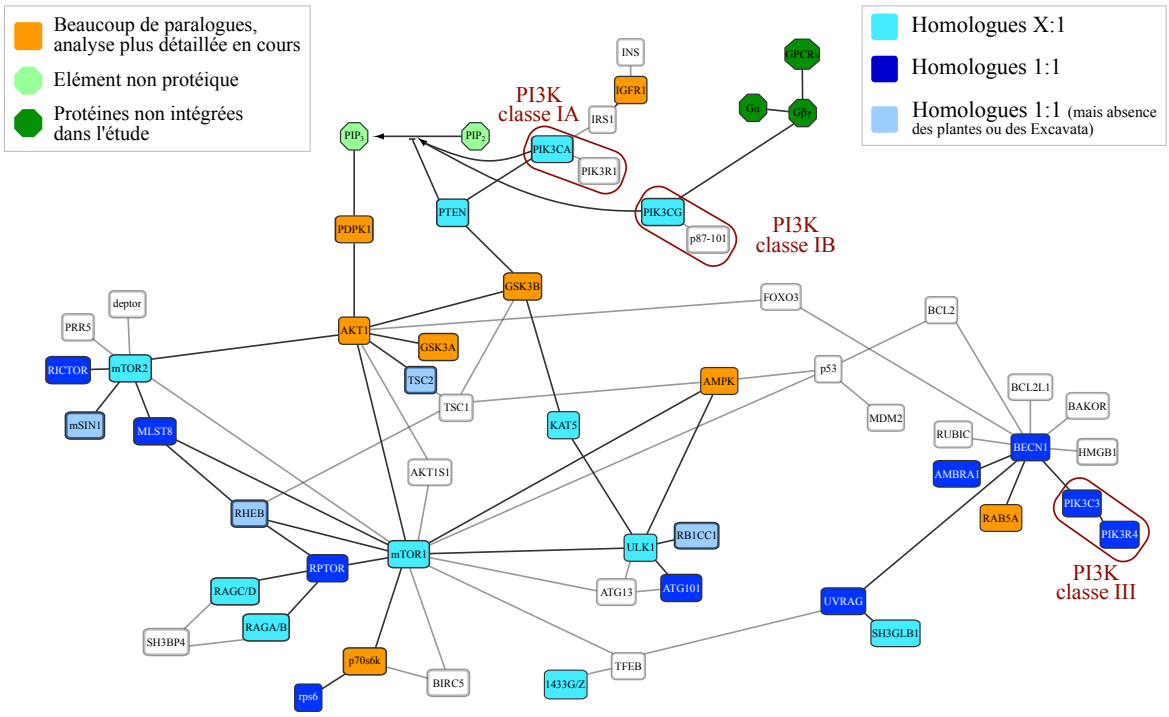

FIGURE 5.8 – ***Voie AKT/mTOR de LECA*** Les protéines présentes et absentes du génome de LECA sont représentées respectivement en bleu et en transparent. Le bleu clair indique que les protéines ont subi des duplications chez certains groupes d'Eucaryotes. Graphique généré à l'aide de Cytoscape.

De façon intéressante, les protéines clefs comme Beclin et mTOR étaient déjà présentes en une seule copie chez LECA. La troisième protéine centrale, AKT1 nécessite une analyse phylogénétique plus approfondie. En effet, de nombreuses duplications ont eu lieu et le jeu de données composé de 1326 protéines est difficilement interprétable en l'état. Néanmoins, un résultat intéressant est que la duplication

à l'origine des protéines AKT1, AKT2 et AKT3 humaines semble avoir eu lieu avant la divergence des Euteleostomi, suggérant la présence d'une seule copie chez la majorité des Eucaryotes.

Les Archaeplastida

Peu d'événements évolutifs se sont produits depuis LECA jusqu'au dernier ancêtre commun des plantes. Seule le gène codant pour la protéine RICTOR a été perdu dans l'ensemble de cette lignée eucaryote. A noter que malgré la présence d'homologues de SIN1 et de RBCC1 chez les SAR et les Unikonta, aucune séquence de plante n'a été détectée pour ces deux protéines. Enfin, parmi les Archaeplastida, seules les algues rouges possèdent un homologue des protéines RHEB, RRAGA, RRAGB, RRAGC et RRAGD, suggérant des pertes indépendantes chez les plantes vertes.

Les Amoebozoa

Seules quatre espèces d'Amoebozoa sont présentes dans la banque de données de génomes complets eucaryotes utilisée par EPINe (voir section 5.2.2). Ainsi il est difficile d'émettre des hypothèses solides pour ce groupe eucaryote. Néanmoins, les 62 arbres phylogénétiques montrent qu'une seule protéine supplémentaire, à savoir RUBIC, a émergé depuis LECA jusq'au dernier ancêtre commun des Amoebozoa. A noter que pour la protéine AMRA1, seuls deux homologues codés par le génome de *A. castellanii* ont été détectés, bien que cette protéine soit retrouvée chez l'ensemble des autres groupes eucaryotes, donc probablement présente chez LECA.

Les Fungi

Si les champignons ont subi les pertes (majeures) des PI3K des classes I et II, seules les pertes des protéines AMRA1 et RUBIC dans cette lignée ont été observées dans les 62 arbres phylogénétiques (Figure 5.9). Par ailleurs, bien que la présence de deux protéines homologues à mTOR aient été précédemment détectée chez *S. cerevisiae* [411], cette duplication ne s'est pas produite chez le dernier ancêtre commun des champignons. En effet, seules quelques lignées de champignons possèdent deux copies de ce gène, suggérant des duplications secondaires indépendantes.

Le croisement des données des arbres phylogénétiques obtenus par EPINe et des données interactomiques disponibles pour la levure et l'Homme montrent qu'il y a autant d'interactions conservées que perdues entre les homologues de ces deux

organismes (Figure 5.9). De façon surprenante, sept protéines (indiquées par une étoiles sur la Figure 5.9) sont présentes chez la majorité des champignons mais sont absentes chez *S. cerevisiae*. Il n'est donc pas possible de conclure quant à la conservation des interactions de ces protéines. Enfin, bien que la levure possède un homologue à la protéine RHEB (nommé RHBP1), aucune de ses interactions ne semblent être conservées. On ne peut néanmoins pas savoir si ces interactions ont été testées et non détectées, ou si aucune expérience n'a été réalisée. En effet, les bases de données d'interactions répertorient la présence des interactions mais n'indiquent généralement pas les résultats négatifs des tests d'interactions.

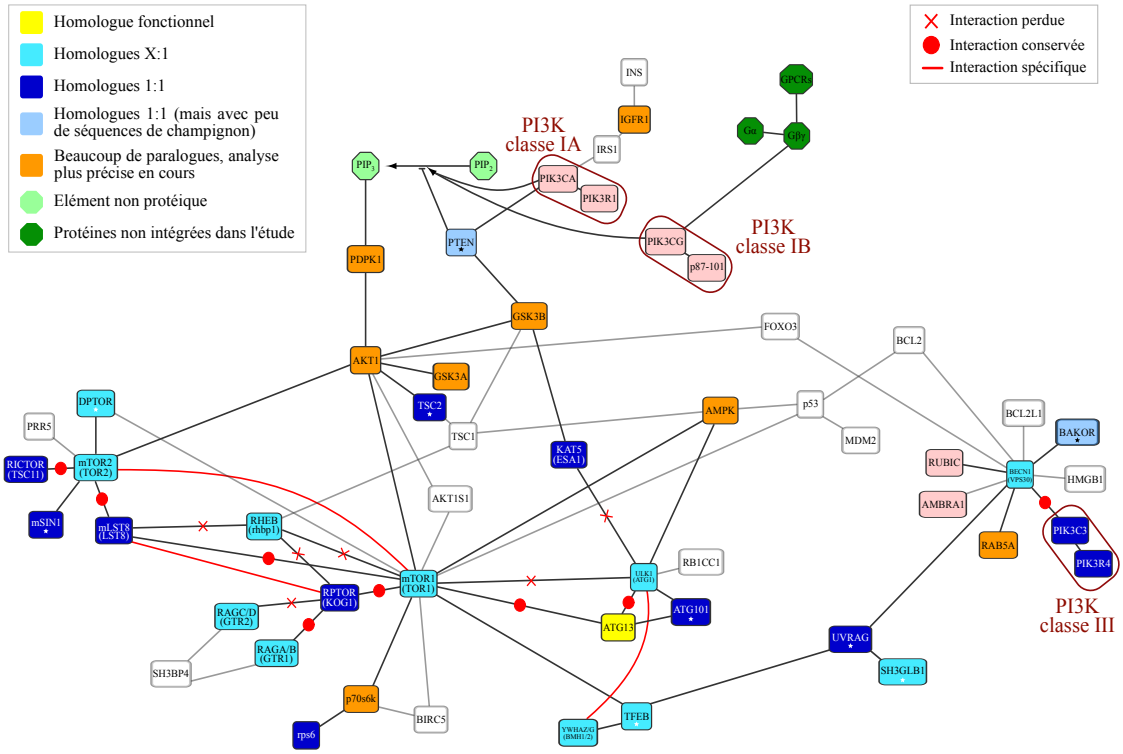

FIGURE 5.9 – ***Voie AKT/mTOR de l'ancêtre commun des Fungi.*** Le code couleur est le même que pour la Figure 5.8. Les étoiles indiquent les gènes possédant un homologue chez des champignons mais absents du génome de *S. cerevisiae*.

Les duplications récentes

La voie AKT/mTOR a connu deux expansions majeures. La première a eu lieu chez le dernier ancêtre commun des Metazoa. En effet, de 21 protéines chez les Opisthokonta, la voie est passée à 35 protéines au moment de la divergence des Choanoflagellida/Metazoa. La seconde expansion a eu lieu chez les Vertebrata dont le dernier ancêtre commun possédait déjà l'ensemble des protéines de la voie humaine.

La duplication RRAGC/RRAGD a eu lieu avant la divergence entre les Gnatostomata et les Chondrichthyes. Enfin, deux duplications, celles de Beclin1/Beclin2 et celle de RRAGA/RRAGB, sont plus récentes puisque spécifiques des Mammalia.

5.4 Conclusions

Dans cette étude, j'ai tout d'abord démontré qu'un peu plus d'un tiers des protéines de la voie AKT/mTOR étaient déjà présentes chez LECA, notamment les protéines clefs telles que les PI3K, mTOR et Beclin. J'ai également identifié plusieurs absences majeures chez les plantes, ce qui pourrait en partie expliquer les différences structurelles observées dans le processus d'autophagie de ce groupe eucaryote. L'analyse de l'interactome de la levure a quant à lui permis de mettre en évidence la conservation d'une partie importante des interactions existantes chez l'Homme. Ces résultats suggèrent donc que ces homologues pourraient exercer la même fonction biologique. De façon surprenante, il apparaît que les champignons *Allomyces macrogynus*, *Mortierella verticillata* et *Mucor circinelloides* possèdent un homologue de la protéine PTEN bien que j'ai précédemment montré que les PI3K de classe I, qui catalysent la réaction inverse de PTEN, ont été perdues dans les lignées de Fungi. Si l'analyse phylogénétique approfondie des protéines PDPK1 et AKT1 révèle la présence d'homologues dans ce groupe eucaryote, on pourra supposer que l'activation de l'autophagie chez les champignons s'effectue, comme chez l'Homme, à travers la présence dans le cytoplasme de PI(3,4,5)P_3 ; bien que ces derniers ne soient pas synthétisés grâce aux protéines PI3K de la classe I.

Enfin, si le pipeline EPINe, dans sa version actuelle, permet l'inférence automatique des arbres de gènes correspondants aux identifiants protéiques des composés d'une voie de signalisation d'intérêt, l'intégration de données interactomiques d'espèces modèles est encore en cours. L'utilisation de cet outil pour la découverte de nouvelles protéines impliquées dans un processus précis est également envisagée et détaillée dans le chapitre 6 de ce manuscrit.

Chapitre 6

Conclusion générale

De la famille génique au réseau métabolique

Le thème central de ma thèse était l'étude de l'évolution des voies de signalisation au moyen des méthodes de phylogénie moléculaire. Dans un premier temps, l'analyse de la famille des PI3K m'a permis d'approfondir mes connaissances théoriques sur les méthodologies existantes. Bien que deux phylogénies incomplètes de cette famille aient été publiées avant ma thèse, l'étude présentée dans ce manuscrit dresse un portrait plus détaillé de leur histoire évolutive. Tout d'abord, j'ai montré que LECA possédait déjà une sous-unité catalytique et une sous-unité régulatrice de la classe III, qui n'ont subi aucune duplication secondaire. Selon l'hypothèse admise, son génome codait également pour une ou deux protéines catalytiques de la classe I/II. Ensuite, de manière surprenante, l'analyse a montré que les protéines régulatrices de la classe I ont émergé beaucoup plus récemment que les sous-unités catalytiques qu'elles régulent chez l'Homme. On peut donc se demander, par exemple, quel est le mécanisme de régulation de ces protéines catalytiques dans un groupe tel que les Amoebozoa, dont les génomes ne codent pour aucune protéines régulatrices de classe I. Pour finir, l'analyse phylogénétique précise de la classe I des PI3K couplée à l'étude de leur composition en domaines suggère que la duplication à l'origine des sous-classes IA et IB catalytiques s'est accompagnée d'un changement fonctionnel. Chez l'Homme, la sous-unité catalytique IB est en effet impliquée dans la motilité de certaines cellules du système immunitaire tandis que les sous-unités IA sont impliquées dans différents autres processus. La copie pré-duplication présente chez les Amoebozoa a, quant à elle, été décrite comme impliquée dans la chimiotaxie de ces organismes dans de nombreuses études indépendantes. On peut

donc se demander si la duplication IA/IB survenue chez le dernier ancêtre commun des MIC et le changement fonctionnel l'accompagnant n'est pas directement lié au passage à la multi-cellularité au sein des Opisthokonta.

Cette première étude phylogénétique a mis en évidence une problématique spécifique des jeux de données eucaryotes, à savoir la gestion des transcrits alternatifs. Détectés lors de la phase de recherche de similarité de séquences, une sélection de ces transcrits alternatifs doit être effectuée avant l'étape d'inférence des arbres de gènes. En effet, une seule séquence protéique par locus génomique doit être conservée afin d'éviter des biais lors de l'alignement multiple de séquences ainsi qu'au moment de la sélection de blocs conservés. Pourtant, aucune méthodologie fondée sur des critères de similarités de séquences n'a encore été publiée pour répondre à ce problème. Dans ce contexte, j'ai développé BATfinder qui, bien qu'utilisant une méthodologie très proche de celle de GUIDANCE, présente un temps d'exécution plus faible que ce dernier, notamment en mode parallèle. Ce logiciel permet également d'obtenir en moyenne de meilleurs arbres phylogénétiques que la sélection systématique du transcrit alternatif le plus long.

Enfin, l'étude de la voie AKT/mTOR, activée par les PI3K, m'a donnée l'occasion de commencer le développement d'EPINe, un pipeline automatique dédié à l'étude de la mise en place des réseaux métaboliques eucaryotes. En effet, grâce à l'expérience acquise avec l'étude des PI3K, j'ai pu mettre en place une chaîne de traitement qui, à partir d'une liste d'identifiants UniProtKB, produit les arbres eucaryotes correspondants. Afin de faciliter leur interprétation par l'utilisateur, ces derniers sont annotés (fonction biologique) et coloriés selon leur appartenance taxonomique. Bien que toujours en cours, l'étude de cette voie m'a permis de montrer que les principales protéines de la voie AKT/mTOR telles que AKT, Beclin1 ou mTOR, étaient déjà présentes chez LECA. L'analyse des arbres phylogénétiques a également permis de mettre en évidence une expansion majeure qui a eu lieu avant le dernier ancêtre commun des Metazoa.

Cependant, au delà de l'étude de l'histoire évolutive de chaque composant d'une voie de signalisation, c'est l'analyse couplée de ces données phylogénétiques et de données interactomiques qui est pertinente pour l'étude de la mise en place des voies de signalisations au cours de l'Évolution. En effet, la plupart des protéines impliquées dans un processus cellulaire exercent leur fonction en interagissant physiquement et directement avec d'autres protéines présentes dans la cellule. Dans le cas de la voie AKT/mTOR, l'étude préliminaire des arbres inférés conjointement

aux données issues des interactomes de l'Homme et de la levure a ainsi permis de mettre en évidence un certain degré de conservation des interactions, supposant une conservation de la fonction des homologues.

Perspectives d'EPINe

Si l'analyse des 62 arbres phylogénétiques de la voie de signalisation AKT/mTOR a été effectuée manuellement, l'objectif du pipeline EPINe est, à terme, de gérer automatiquement cette étape. Dans le cadre ce projet, les principales informations à extraire des arbres phylogénétiques sont : i) la date de l'émergence de la protéine chez les Eucaryotes, ii) les événements de duplication que le gène a subi, et iii) la liste des organismes ou lignées ayant perdu le gène.

Dans un premier temps, l'identification automatique des événements évolutifs (duplications et pertes) subi par chaque gène sera inféré à l'aide de PhylDog [412], qui est un logiciel de réconciliation d'arbres de gènes et d'arbre d'espèces. Ainsi, cette étape implique de posséder un arbre d'espèces de référence composés de l'ensemble des espèces eucaryotes présentes dans mes banques de données (arbre présenté dans l'annexe C.4). Dans un second temps, le logiciel TPMS [413] sera utilisé afin de rechercher des motifs dans les arbres réconciliés et donc de dater les duplications identifiées par PhylDog. Les quelques tests préliminaires que j'ai effectués sur des protéines de la voie AKT/mTOR sont prometteurs et l'intégration de ces deux logiciels dans EPINe devrait être effectué assez rapidement.

Par ailleurs, son principal objectif étant de combiner informations phylogénétiques et interactomiques, il est nécessaire de proposer différents interactomes d'organismes modèles dans EPINe. Une perspective de ce travail est donc d'obtenir les interactomes complets du nématode, de la plante *A. thaliana* et si possible un interactome d'Excavata à partir des données issues des bases publiques d'interactions protéine-protéine. Enfin, le couplage automatique de ces deux types d'informations sera effectué à travers la visualisation de l'évolution de la voie de signalisation avec le logiciel Cytoscape [410]. En effet ce logiciel permet, entre autres, de visualiser les graphes ainsi que de colorier automatiquement les nœuds (*i.e.* les protéines) et les arrêtes (*i.e.* les interactions) en fonction d'un attribut donné (absence/présence d'homologue dans le groupe considéré, conservation de l'interaction, etc.). Néanmoins, la mise en place de cette approche est complexifiée par la nécessité de l'automatisation de la recherche de conservation d'interaction. En effet, elle implique de gérer automatiquement les pertes éventuelles d'interacteurs ainsi que

les duplications ayant eu lieu depuis la divergence des deux organismes modèles. L'implémentation de cette étape dépend donc de l'étape précédente permettant de détecter automatiquement ces événements évolutifs.

Enfin, un dernier objectif du projet EPINe est l'identification de nouvelles protéines potentiellement impliquées dans la voie de signalisation étudiée mais non décrites comme tel dans la littérature. Pour ce faire, un algorithme de clustering sera tout d'abord utilisé sur les interactomes des organismes modèles afin de mettre en évidence des sous-réseaux de protéines fortement connectées. Dans un second temps, les sous-réseaux dont une portion significative des nœuds sont des protéines de la voie d'intérêt seront retenus. En effet, puisqu'en interaction avec ces dernières, les protéines composant ces sous-réseaux sont potentiellement impliquées dans la régulation du processus biologique étudié. Pour tester cette hypothèse, la dernière étape consiste à inférer l'histoire évolutive de ces protéines candidates. Si, dans les organismes modèles, des interologues de ces protéines existent ; alors il peut être intéressant de tester expérimentalement l'impact de ces dernières sur le processus étudié. Dans ce but, j'ai commencé à utiliser l'algorithme OCG (***O****verlapping* ***C****luster* ***G****enerator*) [414] qui présente l'avantage de partitionner les graphes en modules chevauchants (Figure 6.1). Ainsi, un nœud peut appartenir à plusieurs sous-réseaux ou *cliques* ; ce qui est particulièrement intéressant puisqu'une protéine peut être impliquée dans plusieurs processus biologiques.

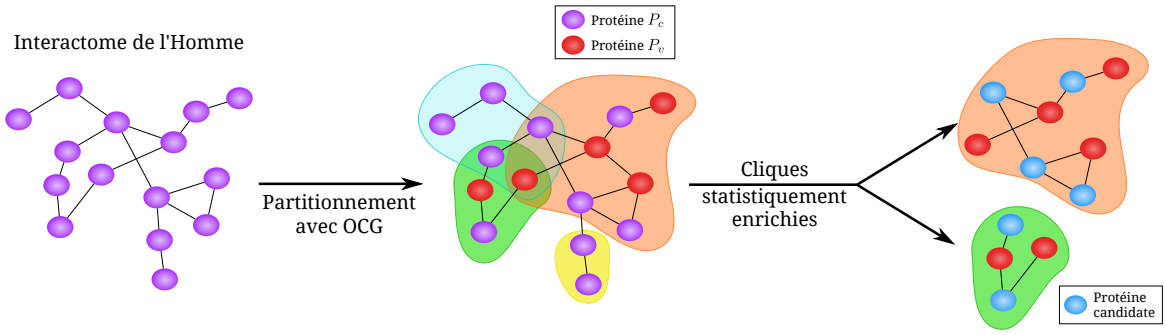

FIGURE *6.1* – ***Principe de l'identification de nouvelles protéines potentiellement impliquées dans la voie de signalisation étudiée.***

Soient P_v les protéines initialement identifiées comme impliquées dans la voie de signalisation considérée et P_c les autres protéines de l'interactome. Une fois le programme OCG exécuté, un test de Fisher exact est réalisé afin de déterminer si le nombre de protéines P_v du sous-réseaux est significatif. Ce test étant réalisé sur l'ensemble des réseaux dont au moins l'une des protéines fait partie de la liste

d'entrée d'EPINe, une correction des tests multiples par la méthode de Benjamini et Hochberg [415] est effectuée.

Dans le cadre de l'étude de la voie AKT/mTOR, j'ai utilisé l'interactome humain généré par Christine Brun[1] à partir des PPI répertoriées dans les principales banques de données d'interaction (voir Table 1.3 du chapitre 1.2.1). Celui-ci est composé de 12 614 protéines formant 67 973 interactions. L'exécution d'OCG a conduit à la génération de 923 cliques chevauchantes dont 326 possèdent une des 62 protéines de la voie AKT/mTOR. Avec un seuil de sélection $P < 10^{-5}$, seuls onze sous-réseaux sont identifiés comme statistiquement enrichis en protéines de cette voie. L'analyse détaillée de ces derniers ainsi que l'analyse phylogénétique des protéines candidates est en cours.

Pour conclure, si le pipeline EPINe est actuellement développé pour l'étude de l'histoire évolutive du processus d'autophagie par la voie AKT/mTOR, il est théoriquement généralisable à l'étude de la mise en place de n'importe quel réseau métabolique eucaryote.

[1]Laboratoire TAGC, Inserm U1090, Université Aix-Marseille

Bibliographie

[1] C. von Linne. *Systema naturae per regna tria naturæ, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis.* Leyde, 1735.

[2] G. Lecointre and H. Le Guyader. *The Tree of Life : A Phylogenetic Classification.* Harvard University Press, Harvard, 2006.

[3] J. B. Lamarck. *Philosophie Zoologique.* Dentu, Paris, 1809.

[4] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* John Murray, London, 1859.

[5] J. R. Porter. Antony van Leeuwenhoek : tercentenary of his discovery of bacteria. *Bacteriol. Rev.*, 40(2) :260–269, 1976.

[6] W. Hennig. *Phylogenetic Systematics.* University of Illinois Press, Champaign, 1965.

[7] C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a natural system of organisms : proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA*, 87(12) :4576–4579, 1990.

[8] S. M. Adl, A. G. Simpson, C. E. Lane, et al. The revised classification of eukaryotes. *J. Eukaryot. Microbiol.*, 59(5) :429–493, 2012.

[9] F. Burki. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harbor Perspect. Biol.*, 6(5) :a016147, 2014.

[10] D. Graur and W. Martin. Reading the entrails of chickens : molecular timescales of evolution and the illusion of precision. *Trends Genet.*, 20(2) :80–86, 2004.

[11] S. B. Hedges and S. Kumar. Precision of molecular time estimates. *Trends Genet.*, 20(5) : 242–247, 2004.

[12] A. J. Roger and L. A. Hug. The origin and diversification of eukaryotes : problems with molecular phylogenetics and molecular clock estimation. *Phil. Trans. R. Soc. B*, 361(1470) :1039–1054, 2006.

[13] S. B. Hedges, J. E. Blair, M. L. Venturi, and J. L. Shoe. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol.*, 4 :2, 2004.

[14] E. J. P. Douzery, E. A. Snell, E. Bapteste, F. Delsuc, and H. Philippe. The timing of eukaryotic evolution : does a relaxed molecular clock reconcile proteins and fossils ? *Proc. Natl. Acad. Sci. USA*, 101(43) :15386–15391, 2004.

[15] L. Eme, S. C. Sharpe, M. W. Brown, and A. J. Roger. On the age of eukaryotes : evaluating evidence from fossils and molecular clocks. *Cold Spring Harbor Perspect. Biol.*, 6(8), 2014.

[16] H. Philippe, A. Germot, and D. Moreira. The new phylogeny of eukaryotes. *Curr. Opin. Genet. Dev.*, 10(6) :596–601, 2000.

[17] R. Derelle and B. F. Lang. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol. Biol. Evol.*, 29(4) :1277–1289, 2012.

[18] Eukaryotic Cells. http://techhydra.com/science/biology/cellular-biology/.

[19] Proteomics/Protein Primary Structure/Alternative Splicing. https://en.wikibooks.org/wiki/Proteomics/Protein_Primary_Structure/Alternative_Splicing.

[20] W. B. Barbazuk, Y. Fu, and K. M. McGinnis. Genome-wide analyses of alternative splicing in plants : opportunities and challenges. *Genome Res.*, 18(9) :1381–1392, 2008.

[21] E. T. Wang, R. Sandberg, S. Luo, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221) :470–476, 2008.

[22] C. W. Sugnet, W. J. Kent, M. Ares, Jr, and D. Haussler. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.*, pages 66–77, 2004.

[23] Les différents types d'epissage. http://www.crcl.fr/547-Grand-public.crcl.aspx?language=fr-FR.

[24] G. Edwalds-Gilbert. Regulation of mRNA splicing by signal transduction. *Nature Educ.*, 3(9) : 43, 2010.

[25] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid. *Nature*, 171(4356) :737–738, 1953.

[26] R. Dahm. Friedrich Miescher and the discovery of DNA. *Dev. Biol.*, 278(2) :274–288, 2005.

[27] G. B. Blackburn, M. J. Gait, D. Loakes, and D. M. Williams. *Nucleic Acids in Chemistry and Biology*. The Royal Society of Chemistry, Cambridge, 2006.

[28] P. A. Levene. The structure of Yeast nucleic acid. IV. Ammonia hydrolysis. *J. Biol. Chem.*, 40 : 415–424, 1919.

[29] F. Sanger and H. Tuppy. The amino-acid sequence in the phenylalanyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. *Biochem. J.*, 49(4) :481–490, 1951.

[30] F. Sanger and H. Tuppy. The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem. J.*, 49(4) :463–481, 1951.

[31] R. Wu. Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage lambda and 186 DNA. *J. Mol. Biol.*, 51(3) :501–521, 1970.

[32] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94(3) :441–448, 1975.

[33] F. Sanger, G. M. Air, B. G. Barrell, et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596) :687–695, 1977.

[34] R. D. Fleischmann, M. D. Adams, O. White, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223) :496–512, 1995.

[35] A. Goffeau, B. G. Barrell, H. Bussey, et al. Life with 6000 genes. *Science*, 274(5287) :546, 563–546, 567, 1996.

[36] C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans* : a platform for investigating biology. *Science*, 282(5396) :2012–2018, 1998.

[37] E. S. Lander, L. M. Linton, B. Birren, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921, 2001.

[38] J. C. Venter, M. D. Adams, E. W. Myers, et al. The sequence of the human genome. *Science*, 291(5507) :1304–1351, 2001.

[39] J. D. McPherson, M. Marra, L. Hillier, et al. A physical map of the human genome. *Nature*, 409(6822) :934–941, 2001.

[40] M. Margulies, M. Egholm, W. E. Altman, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057) :376–380, 2005.

[41] 454Roche Whole Genome Sequencing. http://454.com/applications/whole-genome-sequencing/index.asp.

[42] History of Illumina Sequencing. http://www.illumina.com/technology/next-generation-sequencing/solexa-technology.html.

[43] J. Shendure, G. J. Porreca, N. B. Reppas, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741) :1728–1732, 2005.

[44] M. Kircher and J. Kelso. High-throughput DNA sequencing–concepts and limitations. *Bioessays*, 32(6) :524–536, 2010.

[45] 5500 Series Genetic Analysis Systems. http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html.

[46] L. Rensing. Periodic geophysical and biological signals as Zeitgeber and exogenous inducers in animal organisms. *Int. J. Biometeorol.*, 16 Suppl :113–125, 1972.

[47] M. Rodbell, L. Birnbaumer, S. L. Pohl, and H. M. Krans. The glucagon-sensitive adenyl cyclase system in plasma membranes of rat liver. V. An obligatory role of guanylnucleotides in glucagon action. *J. Biol. Chem.*, 246(6) :1877–1882, 1971.

[48] M. Rodbell. The role of hormone receptors and GTP-regulatory proteins in membrane transduction. *Nature*, 284(5751) :17–22, 1980.

[49] C. Sanchez, C. Lachaize, F. Janody, et al. Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database. *Nucleic Acids Res.*, 27(1) :89–94, 1999.

[50] S. V. Rajagopala, P. Sikorski, A. Kumar, et al. The binary protein-protein interaction landscape of *Escherichia coli*. *Nat. Biotechnol.*, 32(3) :285–290, 2014.

[51] P. A. Elkins, J. M. Watts, M. Zalacain, et al. Insights into catalysis by a knotted TrmD tRNA methyltransferase. *J. Mol. Biol.*, 333(5) :931–949, 2003.

[52] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230) :245–246, 1989.

[53] M. Fromont-Racine, J.C. Rain, and P. Legrain. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.*, 16 :277–282, 1997.

[54] P. Uetz, L. Giot, G. Cagney, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770) :623–627, 2000.

[55] A. J. Walhout, R. Sordella, X. Lu, et al. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, 287(5450) :116–122, 2000.

[56] S. Li, C. M. Armstrong, N. Bertin, et al. A map of the interactome network of the metazoan *C. elegans*. *Science*, 303(5657) :540–543, 2004.

[57] L. Giot, J. S. Bader, C. Brouwer, et al. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651) :1727–1736, 2003.

[58] J.F. Rual, K. Venkatesan, T. Hao, et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062) :1173–1178, 2005.

[59] U. Stelzl, U. Worm, M. Lalowski, et al. A human protein-protein interaction network : a resource for annotating the proteome. *Cell*, 122(6) :957–968, 2005.

[60] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, 18(12) :1257–1261, 2000.

[61] E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data ? *J. Mol. Biol.*, 327(5) :919–923, 2003.

[62] V. S. Rao, K. Srinivas, G. N. Sujini, and G. N. S. Kumar. Protein-protein interaction detection : methods and analysis. *Int. J. Proteomics*, 2014 :147648, 2014.

[63] G. Neubauer, A. Gottschalk, P. Fabrizio, et al. Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl. Acad. Sci. USA*, 94 (2) :385–390, 1997.

[64] Y. Ho, A. Gruhler, A. Heilbut, et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868) :180–183, 2002.

[65] W. H. Dunham, M. Mullin, and A.C. Gingras. Affinity-purification coupled to mass spectrometry : basic principles and strategies. *Proteomics*, 12(10) :1576–1590, 2012.

[66] B. A. Shoemaker and A. R. Panchenko. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, 3 (4) :e43, 2007.

[67] E. M. Marcotte, M. Pellegrini, H. L. Ng, et al. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428) :751–753, 1999.

[68] P. E. Hodges, A. H. McKee, B. P. Davis, W. E. Payne, and J. I. Garrels. The Yeast Proteome Database (YPD) : a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.*, 27(1) :69–73, 1999.

[69] N. Simonis, J. F. Rual, A. R. Carvunis, et al. Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. *Nat. Methods*, 6(1) :47–54, 2009.

[70] M. C. Costanzo, J. D. Hogan, M. E. Cusick, et al. The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD) : comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.*, 28(1) :73–76, 2000.

[71] Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an *Arabidopsis* interactome map. *Science*, 333 :601–606, 2011.

[72] K. Venkatesan, J.F. Rual, A. Vazquez, et al. An empirical framework for binary interactome mapping. *Nat. Methods*, 6(1) :83–90, 2009.

[73] M. P. H. Stumpf, T. Thorne, E. de Silva, et al. Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA*, 105(19) :6959–6964, 2008.

[74] J. Andreani and R. Guerois. Evolution of protein interactions : from interactomes to interfaces. *Arch. Biochem. Biophys.*, 554 :65–75, 2014.

[75] Human Interactome Project. http://interactome.dfci.harvard.edu/H_sapiens/index.php.

[76] T. Rolland, M. Tasan, B. Charloteaux, et al. A proteome-scale map of the human interactome network. *Cell*, 159(5) :1212–1226, 2014.

[77] M. Tyagi, R. R. Thangudu, D. Zhang, et al. Homology inference of protein-protein interactions via conserved binding sites. *PLoS One*, 7(1) :e28896, 2012.

[78] V. Janjić, R. Sharan, and N. Pržulj. Modelling the yeast interactome. *Sci. Rep.*, 4 :4273, 2014.

[79] T. Ito, T. Chiba, R. Ozawa, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, 98(8) :4569–4574, 2001.

[80] A.-C. Gavin, M. Bösche, R. Krause, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868) :141–147, 2002.

[81] C. A. Stanyon, G. Liu, B. A. Mangiola, et al. A Drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biol.*, 5(12) :R96, 2004.

[82] E. Formstecher, S. Aresta, V. Collura, et al. Protein interaction mapping : a Drosophila case study. *Genome Res.*, 15(3) :376–384, 2005.

[83] S. Orchard, M. Ammari, B. Aranda, et al. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, 42(Database issue) :D358–D363, 2014.

[84] A. Chatr-Aryamontri, B.-J. Breitkreutz, R. Oughtred, et al. The BioGRID interaction database : 2015 update. *Nucleic Acids Res.*, 43(Database issue) :D470–D478, 2015.

[85] L. Licata, L. Briganti, D. Peluso, et al. MINT, the molecular interaction database : 2012 update. *Nucleic Acids Res.*, 40(Database issue) :D857–D861, 2012.

[86] D. Szklarczyk, A. Franceschini, M. Kuhn, et al. The STRING database in 2011 : functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, 39(Database issue) :D561–D568, 2011.

[87] S. Kerrien, B. Aranda, L. Breuza, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, 40(Database issue) :D841–D846, 2012.

[88] H. W. Mewes, A. Ruepp, F. Theis, et al. MIPS : curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.*, 39(Database issue) :D220–D224, 2011.

[89] R. Goel, H. C. Harsha, A. Pandey, and T S K. Prasad. Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. *Mol Biosyst*, 8(2) :453–463, 2012.

[90] I. H. Moal and J. Fernández-Recio. SKEMPI : a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics*, 28(20) :2600–2607, 2012.

[91] R. Mosca, A. Céol, A. Stein, R. Olivella, and P. Aloy. 3did : a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, 42(Database issue) :D374–D379, 2014.

[92] B. A. Shoemaker, D. Zhang, M. Tyagi, et al. IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Res.*, 40(Database issue) :D834–D840, 2012.

[93] A. Baspinar, E. Cukuroglu, R. Nussinov, O. Keskin, and A. Gursoy. PRISM : a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. *Nucleic Acids Res.*, 42(Web Server issue) :W285–W289, 2014.

[94] G. Faure, J. Andreani, and R. Guerois. InterEvol database : exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Res.*, 40(Database issue) :D847–D856, 2012.

[95] S. Orchard, H. Hermjakob, and R. Apweiler. The proteomics standards initiative. *Proteomics*, 3(7) :1374–1376, 2003.

[96] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, et al. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, 22(2) :177–183, 2004.

[97] S. Kerrien, S. Orchard, L. Montecchi-Palazzi, et al. Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, 5 :44, 2007.

[98] B. Aranda, H. Blankenburg, S. Kerrien, et al. PSICQUIC and PSISCORE : accessing and scoring molecular interactions. *Nat. Methods*, 8(7) :528–529, 2011.

[99] L. Salwinski, C. S. Miller, A. J. Smith, et al. The Database of Interacting Proteins : 2004 update. *Nucleic Acids Res.*, 32(Database issue) :D449–D451, 2004.

[100] T. V. Vo, J. Das, M. J. Meyer, et al. A Proteome-wide fission yeast interactome reveals network evolution principles from yeasts to Human. *Cell*, 164(1-2) :310–323, 2016.

[101] L. R. Matthews, P. Vaglio, J. Reboul, et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res.*, 11(12) :2120–2126, 2001.

[102] H. Yu, N. M. Luscombe, H. X. Lu, et al. Annotation transfer between genomes : protein-protein interologs and protein-DNA regulogs. *Genome Res.*, 14(6) :1107–1118, 2004.

[103] T. K. B. Gandhi, J. Zhong, S. Mathivanan, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.*, 38(3) :285–293, 2006.

[104] K. R. Brown and I. Jurisica. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.*, 8(5) :R95, 2007.

[105] T. J. P. van Dam and B. Snel. Protein complex evolution does not involve extensive network rewiring. *PLoS Comput. Biol.*, 4(7) :e1000132, 2008.

[106] W. Qian, X. He, E. Chan, H. Xu, and J. Zhang. Measuring the evolutionary rate of protein-protein interaction. *Proc. Natl. Acad. Sci. USA*, 108(21) :8725–8730, 2011.

[107] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.*, 18(7) :1283–1292, 2001.

[108] P. Beltrao and L. Serrano. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput. Biol.*, 3(2) :e25, 2007.

[109] H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. Evolutionary rate in the protein interaction network. *Science*, 296(5568) :750–752, 2002.

[110] S. Wuchty, A.-L. Barabási, and M. T. Ferdig. Stable evolutionary signal in a yeast protein interaction network. *BMC Evol. Biol.*, 6 :8, 2006.

[111] C. Shou, N. Bhardwaj, H. Y. Lam, et al. Measuring the evolutionary rewiring of biological networks. *PLoS Comput. Biol.*, 7(1) :e1001050, 2011.

[112] C. Liongue, T. Taznin, and A. C. Ward. Signaling via the CytoR/JAK/STAT/SOCS pathway : Emergence during evolution. *Mol. Immunol.*, 71 :166–175, 2016.

[113] M. Dergai, A. Iershov, O. Novokhatska, S. Pankivskyi, and A. Rynditch. Evolutionary changes on the way to clathrin-mediated endocytosis in animals. *Genome Biol. Evol.*, 2016.

[114] M. Freissmuth, P. J. Casey, and A. G. Gilman. G proteins control diverse pathways of transmembrane signaling. *FASEB J.*, 3(10) :2125–2131, 1989.

[115] V. Anantharaman, S. Abhiman, R. F. de Souza, and L. Aravind. Comparative genomics uncovers novel structural and functional features of the heterotrimeric GTPase signaling system. *Gene*, 475(2) :63–78, 2011.

[116] S. R. Neves, P. T. Ram, and R. Iyengar. G protein pathways. *Science*, 296(5573) :1636–1639, 2002.

[117] H. Cho and J. H. Kehrl. Regulation of immune function by G protein-coupled receptors, trimeric G proteins, and RGS proteins. *Prog. Mol. Biol. Transl. Sci.*, 86 :249–298, 2009.

[118] R. Fredriksson and H. B. Schiöth. The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol. Pharmacol.*, 67(5) :1414–1425, 2005.

[119] R. Strotmann, K. Schröck, I. Böselt, et al. Evolution of GPCR : change and continuity. *Mol. Cell. Endocrinol.*, 331(2) :170–178, 2011.

[120] A. de Mendoza, A. Sebé-Pedrós, and I. Ruiz-Trillo. The evolution of the GPCR signaling system in eukaryotes : modularity, conservation, and the transition to metazoan multicellularity. *Genome Biol. Evol.*, 6(3) :606–619, 2014.

[121] G. Perrière and C. Brochier-Armanet. *Concepts et Méthodes en Phylogénie Moléculaire*. Springer, Paris, 2010.

[122] R. V. Eck and M. O Dayhoff. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Springs, 1966.

[123] B. J. Strasser. Collecting, comparing, and computing sequences : the making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965. *J. Hist. Biol.*, 43(4) :623–660, 2010.

[124] J. Mashima, Y. Kodama, T. Kosuge, et al. DNA data bank of Japan (DDBJ) progress report. *Nucleic Acids Res.*, 44(D1) :D51–57, 2016.

[125] R. Gibson, B. Alako, C. Amid, et al. Biocuration of functional annotation at the European nucleotide archive. *Nucleic Acids Res.*, 44(D1) :D58–66, 2016.

[126] K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Res.*, 44(D1) :D67–72, 2016.

[127] G. Cochrane, I. Karsch-Mizrachi, T. Takagi, and International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database collaboration. *Nucleic Acids Res.*, 44(D1) :D48–50, 2016.

[128] UniProt Consortium. UniProt : a hub for protein information. *Nucleic Acids Res.*, 43(Database issue) :D204–212, 2014.

[129] 2015 NAR Database Summary Paper Alphabetic List. http://www.oxfordjournals.org/our_journals/nar/database/a/.

[130] X. M. Fernández-Suárez and M. Y. Galperin. The 2013 Nucleic Acids Resaearch Database Issue and the online molecular biology database collection. *Nucleic Acids Res.*, 41(Database issue) : D1–7, 2013.

[131] M. Y. Galperin, D. J. Rigden, and X. M. Fernández-Suárez. The 2015 Nucleic Acids Research Database Issue and molecular biology database collection. *Nucleic Acids Res.*, 43(Database issue) :D1–5, 2015.

[132] Growth of GenBank and WGS. http://www.ncbi.nlm.nih.gov/genbank/statistics.

[133] J. del Campo, M. E. Sieracki, R. Molestina, et al. The others : our biased perspective of eukaryotic genomes. *Trends Ecol. Evol.*, 29(5) :252–259, 2014.

[134] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3) :403–410, 1990.

[135] Blast Specifics, Peking university. http://www.cbi.pku.edu.cn/docs/faq/BlastSpecifics.html.

[136] S. F. Altschul, T. L. Madden, A. A. Schäffer, et al. Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17) :3389–3402, 1997.

[137] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, 87(6) :2264–2268, 1990.

[138] Blast FAQ. http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#expect.

[139] M. Bhagwat and L. Aravind. PSI-BLAST tutorial. *Methods Mol. Biol.*, 395 :177–186, 2007.

[140] A. Torres, A. Cabada, and J. J. Nieto. An exact formula for the number of alignments between two DNA sequences. *DNA Seq.*, 14(6) :427–430, 2003.

[141] D. Sankoff, R. J. Cedergren, and G. Lapalme. Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. *J. Mol. Evol.*, 7(2) :133–149, 1976.

[142] S. Pascarella and P. Argos. Analysis of insertions/deletions in protein structures. *J. Mol. Biol.*, 224(2) :461–471, 1992.

[143] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22) :4673–4680, 1994.

[144] W. M. Fitch. Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *J. Mol. Biol.*, 26(3) :499–507, 1967.

[145] T. Gojobori, W. H. Li, and D. Graur. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.*, 18(5) :360–369, 1982.

[146] T. D. Kocher and Wilson A. C. Sequence Evolution of Mitochondrial DNA in Humans and Chimpanzees : Control Region and a Protein-Coding Region. In S. Osawa and T. Honjo, editors, *Evolution of Life : Fossils, Molecules and Culture*, pages 391–413. Springer, New York, 1991.

[147] M. O. Dayhoff, R. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume Suppl. 3, pages 345–358. National Biomedical Research Foundation, Washington DC, 1978.

[148] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89(22) :10915–10919, 1992.

[149] A. A. Schäffer, L. Aravind, T. L. Madden, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, 29(14) :2994–3005, 2001.

[150] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón. TrimAl : a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15) :1972–1973, 2009.

[151] A. Criscuolo and S. Gribaldo. BMGE (Block Mapping and Gathering with Entropy) : a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.*, 10 :210, 2010.

[152] K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7 : improvements in performance and usability. *Mol. Biol. Evol.*, 30(4) :772–780, 2013.

[153] L. Duret. *Evolution des Séquences Non-codantes Chez les Vertébrés.* PhD thesis, Université Claude Bernard Lyon 1, 1995.

[154] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3) :443–453, 1970.

[155] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1) :195–197, 1981.

[156] P. Hogeweg and B. Hesper. The alignment of sets of sequences and the construction of phyletic trees : an integrated method. *J. Mol. Evol.*, 20(2) :175–186, 1984.

[157] F. Corpet. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.*, 16 (22) :10881–10890, 1988.

[158] H. M. Martinez. A flexible multiple sequence alignment program. *Nucleic Acids Res.*, 16(5) : 1683–1691, 1988.

[159] D. F. Feng and R. F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, 25(4) :351–360, 1987.

[160] D. G. Higgins and P. M. Sharp. CLUSTAL : a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1) :237–244, 1988.

[161] W. J. Wilbus and D. J. Lipman. The context dependent comparison of biological sequences. *SIAM J. Appl. Math.*, 44(3) :557–567, 1984.

[162] P. H. Sneath and R. R. Sokal. *Numerical Taxonomy.* W.H. Freeman and Company, San Francisco, 1973.

[163] F. Sievers, A. Wilm, D. Dineen, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7 :539, 2011.

[164] G. Blackshields, F. Sievers, W. Shi, A. Wilm, and D. G. Higgins. Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms Mol. Biol.*, 5 :21, 2010.

[165] S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2) :129–136, 1982.

[166] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematics Statistics and Probability*, volume 1, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[167] C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee : A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302(1) :205–217, 2000.

[168] X. Huang and W. Miller. A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.*, 12(3) :337–357, 1991.

[169] M. O. Dayhoff. *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Silver Springs, 1972.

[170] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, 30(14) :3059–3066, 2002.

[171] D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, 8(3) :275–282, 1992.

[172] K. Katoh, K. Kuma, H. Toh, and T. Miyata. MAFFT version 5 : improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, 33(2) :511–518, 2005.

[173] K. Katoh and H. Toh. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics*, 26(15) :1899–1900, 2010.

[174] Robert C. Edgar. MUSCLE : a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5 :113, 2004.

[175] Robert C. Edgar. MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5) :1792–1797, 2004.

[176] M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, 1983.

[177] A. Löytynoja and N. Goldman. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. USA*, 102(30) :10557–10562, 2005.

[178] W. Duchemin, V. Daubin, and E. Tannier. Reconstruction of an ancestral *Yersinia pestis* genome and comparison with an ancient sequence. *BMC Genomics*, 16 Suppl 10 :S9, 2015.

[179] M. Groussin, J. K. Hobbs, G. J. Szöll ?si, et al. Toward more accurate ancestral protein genotype-phenotype reconstructions with the use of species tree-aware gene trees. *Mol. Biol. Evol.*, 32(1) : 13–22, 2015.

[180] G. C. Finnigan, V. Hanson-Smith, B. D. Houser, H. J. Park, and T. H. Stevens. The reconstructed ancestral subunit a functions as both V-ATPase isoforms Vph1p and Stv1p in *Saccharomyces cerevisiae*. *Mol. Biol. Cell*, 22(17) :3176–3191, 2011.

[181] R. K. Bradley, A. Roberts, M. Smoot, et al. Fast statistical alignment. *PLoS Comput. Biol.*, 5 (5) :e1000392, May 2009.

[182] V. Ranwez, S. Harispe, F. Delsuc, and E. J P. Douzery. MACSE : Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One*, 6(9) :e22594, 2011.

[183] J. D. Thompson, F. Plewniak, and O. Poch. BAliBASE : a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1) :87–88, 1999.

[184] U. Roshan and D. R. Livesay. ProbAlign : multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, 22(22) :2715–2721, 2006.

[185] Y. Liu, B. Schmidt, and D. L. Maskell. MSAProbs : multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics*, 26(16) : 1958–1964, 2010.

[186] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou. ProbCons : Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, 15(2) :330–340, 2005.

[187] A. Löytynoja and N. Goldman. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320(5883) :1632–1635, 2008.

[188] J. D. Thompson, B. Linard, O. Lecompte, and O. Poch. A comprehensive benchmark study of multiple sequence alignment methods : current challenges and future perspectives. *PLoS One*, 6(3) :e18093, 2011.

[189] F. S. Pais, P. D. E. C. Ruy, G. Oliveira, and R. S. Coimbra. Assessing the efficiency of multiple sequence alignment programs. *Algorithms Mol. Biol.*, 9(1) :4, 2014.

[190] M. T. Pervez, M. E. Babar, A. Nadeem, et al. Evaluating the accuracy and efficiency of multiple sequence alignment methods. *Evol. Bioinform.*, 10 :205–217, 2014.

[191] K. Liu, T. J. Warnow, M. T. Holder, et al. SATe-II : very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst. Biol.*, 61(1) :90–106, 2012.

[192] J. D. Thompson, P. Koehl, R. Ripp, and O. Poch. BAliBASE 3.0 : latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1) :127–136, 2005.

[193] H. Brinkmann and H. Philippe. Archaea sister group of Bacteria ? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.*, 16(6) :817–825, 1999.

[194] J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, 17(4) :540–552, 2000.

[195] O. Penn, E. Privman, H. Ashkenazy, et al. GUIDANCE : a web server for assessing alignment confidence scores. *Nucleic Acids Res.*, 38(Web Server issue) :W23–W28, 2010.

[196] J. Von Neumann. *Mathematische Grundlagen der Quantenmechanik*. Springer, 1932.

[197] A. W M. Dress, C. Flamm, G. Fritzsch, et al. Noisy : identification of problematic columns in multiple sequence alignments. *Algorithms Mol. Biol.*, 3 :7, 2008.

[198] Ensembl Species. http://www.ensembl.org/info/about/species.html.

[199] E. Zuckerkandl and L. B. Pauling. Molecular disease, evolution, and genetic heterogeneity. In M. Kasha and B. Pullman, editors, *Horizons in Biochemistry*, pages 189–225. Academic Press, 1962.

[200] E. Margoliash. Primary structure and evolution of cytochrome C. *Proc. Natl. Acad. Sci. USA*, 50 :672–679, 1963.

[201] G. G. Simpson. Organisms and Molecules in Evolution. *Science*, 146(3651) :1535–1538, 1964.

[202] C. D. Laird, B. L. McConaughy, and B. J. McCarthy. Rate of fixation of nucleotide substitutions in evolution. *Nature*, 224(5215) :149–154, 1969.

[203] X. Gu and W. H. Li. Higher rates of amino acid substitution in rodents than in humans. *Mol. Phylogenet Evol.*, 1(3) :211–214, 1992.

[204] S. Kumar. Molecular clocks : four decades of evolution. *Nat. Rev. Genet.*, 6(8) :654–662, 2005.

[205] L. L. Cavalli-Sforza and A. W. Edwards. Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.*, 19(3 Pt 1) :233–257, 1967.

[206] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Math. Biosci.*, 53 :131–147, 1981.

[207] M. K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, 11(3) :459–468, 1994.

[208] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.*, 20(1) :86–93, 1984.

[209] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[210] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16(2) :111–120, 1980.

[211] J. Felsenstein. Evolutionary trees from DNA sequences : a maximum likelihood approach. *J. Mol. Evol.*, 17(6) :368–376, 1981.

[212] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22(2) :160–174, 1985.

[213] K. Tamura. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol. Biol. Evol.*, 9(4) :678–687, 1992.

[214] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10(3) :512–526, 1993.

[215] M. Schöniger and A. von Haeseler. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.*, 3(3) :240–247, 1994.

[216] S. V. Muse. Evolutionary analyses when nucleotides do not evolve independently. In M. Nei and N. Takahata, editors, *Current Topics on Molecular Evolution*, pages 115–124. Penn State University Press, Penn State, 1995.

[217] Z. Yang and D. Roberts. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.*, 12(3) :451–458, 1995.

[218] P. G. Foster. Modeling compositional heterogeneity. *Syst. Biol.*, 53(3) :485–495, 2004.

[219] V. Jayaswal, L. S. Jermiin, L. Poladian, and J. Robinson. Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. *Syst. Biol.*, 60(1) :74–86, 2011.

[220] J. A. Lake. Reconstructing evolutionary trees from DNA and protein sequences : paralinear distances. *Proc. Natl. Acad. Sci. USA*, 91(4) :1455–1459, 1994.

[221] S. Blanquart and N. Lartillot. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.*, 23(11) :2058–2071, 2006.

[222] J. R. Lobry. Properties of a general model of DNA evolution under no-strand-bias conditions. *J. Mol. Evol.*, 40(3) :326–330, 1995.

[223] N. Galtier and M. Gouy. Inferring pattern and process : maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.*, 15(7) :871–879, 1998.

[224] S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, 18(5) :691–699, 2001.

[225] S. Q. Le and O. Gascuel. An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, 25(7) :1307–1320, 2008.

[226] N. Lartillot and H. Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, 21(6) :1095–1109, 2004.

[227] J. Adachi and M. Hasegawa. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, 42(4) :459–468, 1996.

[228] L. S. Quang, O. Gascuel, and N. Lartillot. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24(20) :2317–2323, 2008.

[229] S. Q. Le, N. Lartillot, and O. Gascuel. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. B*, 363(1512) :3965–3976, 2008.

[230] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19 (6) :716–723, 1974.

[231] Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites : approximate methods. *J. Mol. Evol.*, 39(3) :306–314, 1994.

[232] W. Hennig. *Phylogenetic Systematics.* University of Illinois Press, 1966.

[233] W. M. Fitch. Toward defining the course of evolution : Minimum change for a specific tree topology . *Syst. Zool.*, 20(4) :406–416, 1971.

[234] E. N. Adams. Consensus techniques and the comparison of taxonomic trees. *Syst. Zool.*, 21(4) : 390–397, 1972.

[235] T. Margush and F. R. McMorris. Consensus n-trees. *Bull. Math. Biol.*, 43(2) :239–244, 1981.

[236] D. Sankoff. Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, 28(1) :35–42, 1975.

[237] D. S. Hochbaum and A. Pathria. Path costs in evolutionary tree reconstruction. *J. Comput. Biol.*, 4(2) :163–175, 1997.

[238] I. Agnarsson and J. A. Miller. Is ACCTRAN better than DELTRAN ? *Cladistics*, 24(6) : 1032–1038, 2008.

[239] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(3760) :279–284, 1967.

[240] K. K. Kidd and L. A. Sgaramella-Zonta. Phylogenetic analysis : Concepts and methods. *Am. J. Hum. Genet.*, 23(3) :235–252, 1971.

[241] A. Rzhetsky and M. Nei. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *J. Mol. Evol.*, 35(4) :367–375, 1992.

[242] A. Rzhetsky and M. Nei. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.*, 10(5) :1073–1095, 1993.

[243] D. Bryant and P. Waddell. Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Mol. Biol. Evol.*, 15 :1346–1359, 1998.

[244] N. Saitou and M. Nei. The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4) :406–425, 1987.

[245] J. A. Studier and K. J. Keppler. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.*, 5(6) :729–731, 1988.

[246] O. Gascuel. BIONJ : an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, 14(7) :685–695, 1997.

[247] J. Aldrich. R.A. Fisher and the making of maximum likelihood 1912-1922. *Stat. Sci.*, 12(3) : 162–176, 1997.

[248] J. Neyman. Molecular studies of evolution : a source of novel statistical problems. In J. Gupta, S. S. et Yackel, editor, *Statistical Decision Theory and Related Topics*, pages 1–17. Academic Press, New York, 1971.

[249] J. Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.*, 25(5) :471–492, 1973.

[250] Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences : a Markov Chain Monte Carlo method. *Mol. Biol. Evol.*, 14(7) :717–724, 1997.

[251] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21 :1087–1092, 1953.

[252] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109, 1970.

[253] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Stat. Sci.*, 7(4) :457–511, 1992.

[254] J. A. Eisen. Phylogenomics : improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, 8(3) :163–167, 1998.

[255] G. H Y. He, C. C. Helbing, M. J. Wagner, C. W. Sensen, and K. Riabowol. Phylogenetic analysis of the ING family of PHD finger proteins. *Mol Biol Evol*, 22(1) :104–116, 2005.

[256] H. Kidron, S. Repo, M. S. Johnson, and T. A. Salminen. Functional classification of amino acid decarboxylases from the alanine racemase structural family by phylogenetic studies. *Mol Biol Evol*, 24(1) :79–89, 2007.

[257] C. Afrasiabi, B. Samad, D. Dineen, C. Meacham, and K. Sjölander. The PhyloFacts FAT-CAT web server : ortholog identification and function prediction using fast approximate tree classification. *Nucleic Acids Res*, 41(Web Server issue) :W242–W248, 2013.

[258] P. Gaudet, M. S. Livstone, S. E. Lewis, and P. D. Thomas. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief Bioinform*, 12(5) :449–462, 2011.

[259] S. M. Sahraeian, K. R. Luo, and S. E. Brenner. SIFTER search : a web server for accurate phylogeny-based protein function prediction. *Nucleic Acids Res*, 43(W1) :W141–W147, 2015.

[260] D. Liu, W. Shi, Y. Shi, et al. Origin and diversity of novel avian influenza A H7N9 viruses causing human infection : phylogenetic, structural, and coalescent analyses. *Lancet*, 381(9881) : 1926–1932, 2013.

[261] O. G. Pybus and J. Thézé. Hepacivirus cross-species transmission and the origins of the hepatitis C virus. *Curr Opin Virol*, 16 :1–7, 2016.

[262] M. Worobey, M. L. Santiago, B. F. Keele, et al. Origin of AIDS : contaminated polio vaccine theory refuted. *Nature*, 428(6985) :820, 2004.

[263] E. Hooper. *The River : A Journey Back to the Source of HIV and AIDS*. Penguin Press, London, 1999.

[264] A. L. Hughes and R. Friedman. The effect of branch lengths on phylogeny : an empirical study using highly conserved orthologs from mammalian genomes. *Mol. Phylogenet. Evol.*, 45(1) : 81–88, 2007.

[265] Z. Zou and J. Zhang. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations ? *Mol. Biol. Evol.*, 32(8) :2085–2096, 2015.

[266] M. A. Bakewell, P. Shi, and J. Zhang. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl. Acad. Sci. USA*, 104(18) :7489–7494, 2007.

[267] M. Carneiro, F. W. Albert, J. Melo-Ferreira, et al. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. *Mol. Biol. Evol.*, 29(7) : 1837–1849, 2012.

[268] J. L. Villanueva-Cañas, S. Laurie, and M. M. Albà. Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol. Evol.*, 5(2) :457–467, 2013.

[269] O. Penn, E. Privman, G. Landan, D. Graur, and T. Pupko. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.*, 27(8) :1759–1767, 2010.

[270] Guidance : Running time. http://guidance.tau.ac.il/ver2/overview.php.

[271] M. Nei and J. Zhang. Evolutionary distance : estimation. *Encyclopaedia Life Sci.*, pages 1–3, 2006.

[272] Eigen template library for linear algebra. http://eigen.tuxfamily.org/.

[273] The openmp® api specification for parallel programming. http://openmp.org/.

[274] M. Gouy, S. Guindon, and O. Gascuel. SeaView version 4 : A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, 27(2) :221–224, 2010.

[275] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, 27 :401–410, 1978.

[276] H. Philippe, P. Lopez, H. Brinkmann, et al. Early-branching or fast-evolving eukaryotes ? An answer based on slowly evolving positions. *Proc. Biol. Sci.*, 267(1449) :1213–1221, 2000.

[277] S. Djordjevic and P. C. Driscoll. Structural insight into substrate specificity and regulatory mechanisms of phosphoinositide 3-kinases. *Trends Biochem. Sci.*, 27(8) :426–432, 2002.

[278] S. Koyasu. The role of PI3K in immune cells. *Nat. Immunol.*, 4(4) :313–319, 2003.

[279] C. Burman and N.T. Ktistakis. Regulation of autophagy by phosphatidylinositol 3-phosphate. *FEBS Lett.*, 584(7) :1302–1312, 2010.

[280] M. Graupera and M. Potente. Regulation of angiogenesis by PI3K signaling networks. *Exp. Cell Res.*, 319(9) :1348–1355, 2013.

[281] C. Sadhu, B. Masinovsky, K. Dick, C G. Sowell, and D. E. Staunton. Essential role of phosphoinositide 3-kinase δ in neutrophil directional movement. *J. Immunol.*, 170(5) :2647–2654, 2003.

[282] P. A. Iglesias. Spatial regulation of PI3K signaling during chemotaxis. *Wiley Interdiscip. Rev. Syst. Biol.*, 1(2) :247–253, 2009.

[283] P. V. Afonso and C. A. Parent. PI3K and chemotaxis : a priming issue ? *Sci. Signal.*, 4(170) : pe22, 2011.

[284] L. C. Foukas, I. M. Berenjeno, A. Gray, A. Khwaja, and B. Vanhaesebroeck. Activity of any class IA PI3K isoform can sustain cell proliferation and survival. *Proc. Natl. Acad. Sci. USA*, 107(25) :11381–11386, 2010.

[285] A. Kumar, J. Redondo-Muñoz, V. Perez-García, et al. Nuclear but not cytosolic phosphoinositide 3-kinase beta has an essential function in cell survival. *Mol. Cell. Biol.*, 31(10) :2122–2133, 2011.

[286] J. Domin and M. D. Waterfield. Using structure to define the function of phosphoinositide 3-kinase family members. *FEBS Lett.*, 410(1) :91–95, 1997.

[287] M. P. Wymann and L. Pirola. Structure and function of phosphoinositide 3-kinases. *Biochim. Biophys. Acta*, 1436(1-2) :127–150, 1998.

[288] T. Maffucci and M. Falasca. New insight into the intracellular roles of class II phosphoinositide 3-kinases. *Biochem. Soc. Trans.*, 42(5) :1378–1382, 2014.

[289] J. R. Brown and K. R. Auger. Phylogenomics of phosphoinositide lipid kinases : perspectives on the evolution of second messenger signaling and drug discovery. *BMC Evol. Biol.*, 11 :4, 2011.

[290] J. Yu, Y. Zhang, J. McIlroy, et al. Regulation of the p85/p110 phosphatidylinositol 3'-kinase : stabilization and inhibition of the p110α catalytic subunit by the p85 regulatory subunit. *Mol. Cell. Biol.*, 18(3) :1379–1387, 1998.

[291] B. Geering, P. R. Cutillas, G. Nock, S. I. Gharbi, and B. Vanhaesebroeck. Class IA phosphoinositide 3-kinases are obligate p85-p110 heterodimers. *Proc. Natl. Acad. Sci. USA*, 104(19) : 7809–7814, 2007.

[292] O. Vadas, J. E. Burke, X. Zhang, A. Berndt, and R. L. Williams. Structural basis for activation and inhibition of class I phosphoinositide 3-kinases. *Sci. Signal*, 4(195) :re2, 2011.

[293] A. Shymanets, Prajwal, K. Bucher, et al. p87 and p101 subunits are distinct regulators determining class IB phosphoinositide 3-kinase (PI3K) specificity. *J. Biol. Chem.*, 288(43) :31059–31068, 2013.

[294] E. Hirsch, L. Braccini, E. Ciraolo, F. Morello, and A. Perino. Twice upon a time : PI3K's secret double life exposed. *Trends Biochem. Sci.*, 34(5) :244–248, 2009.

[295] H. A., 3rd Burris. Overcoming acquired resistance to anticancer therapy : focus on the PI3K/AKT/mTOR pathway. *Cancer Chemother. Pharmacol.*, 71(4) :829–842, 2013.

[296] J. A. Engelman, J. Luo, and L. C. Cantley. The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. *Nat. Rev. Genet.*, 7(8) :606–619, 2006.

[297] F. O Farrell, T. E. Rusten, and H. Stenmark. Phosphoinositide 3-kinases as accelerators and brakes of autophagy. *FEBS J.*, 280(24) :6322–6337, 2013.

[298] Z. A. Knight, B. Gonzalez, M. E. Feldman, et al. A pharmacological map of the PI3-K family defines a role for p110α in insulin signaling. *Cell*, 125(4) :733–747, 2006.

[299] P. Manna and S. K. Jain. Phosphatidylinositol-3,4,5-triphosphate and cellular signaling : implications for obesity and diabetes. *Cell. Physiol. Biochem.*, 35(4) :1253–1275, 2015.

[300] W. P. Fung-Leung. Phosphoinositide 3-kinase delta (PI3Kδ) in leukocyte signaling and function. *Cell Signal*, 23(4) :603–608, 2011.

[301] A. Khwaja. PI3K as a target for therapy in haematological malignancies. *Curr. Top. Microbiol. Immunol.*, 347 :169–188, 2010.

[302] S. E. M. Herman, A. L. Gordon, A. J. Wagner, et al. Phosphatidylinositol 3-kinase-δ inhibitor CAL-101 shows promising preclinical activity in chronic lymphocytic leukemia by antagonizing intrinsic and extrinsic cellular survival signals. *Blood*, 116(12) :2078–2088, 2010.

[303] J. R. Brown, J. C. Byrd, S. E. Coutre, et al. Idelalisib, an inhibitor of phosphatidylinositol 3-kinase p110δ, for relapsed/refractory chronic lymphocytic leukemia. *Blood*, 123(22) :3390–3397, 2014.

[304] P. Sujobert, V. Bardet, P. Cornillet-Lefebvre, et al. Essential role for the p110δ isoform in phosphoinositide 3-kinase activation and cell proliferation in acute myeloid leukemia. *Blood*, 106 (3) :1063–1066, 2005.

[305] C. Billottet, V. L. Grandage, R. E. Gale, et al. A selective inhibitor of the p110δ isoform of PI 3-kinase inhibits AML cell proliferation and survival and increases the cytotoxic effects of VP16. *Oncogene*, 25(50) :6648–6659, 2006.

[306] N. F. Smirnova, S. Gayral, C. Pedros, et al. Targeting PI3Kγ activity decreases vascular trauma-induced intimal hyperplasia through modulation of the Th1 response. *J. Exp. Med.*, 211(9) : 1779–1792, 2014.

[307] A. Fougerat, S. Gayral, P. Gourdy, et al. Genetic and pharmacological targeting of phosphoinositide 3-kinase-γ reduces atherosclerosis and favors plaque stability by modulating inflammatory processes. *Circulation*, 117(10) :1310–1317, 2008.

[308] Y. YXie, P. W. Abel, J. K. Kirui, et al. Identification of upregulated phosphoinositide 3-kinase γ as a target to suppress breast cancer cell migration and invasion. *Biochem. Pharmacol.*, 85 (10) :1454–1462, 2013.

[309] P. Rickert, O. D. Weiner, F. Wang, H. R. Bourne, and G. Servant. Leukocytes navigate by compass : roles of PI3Kγ and its lipid products. *Trends Cell Biol.*, 10(11) :466–473, 2000.

[310] C. Gu and S. Park. The p110γ PI-3 kinase is required for EphA8-stimulated cell migration. *FEBS Lett.*, 540(1-3) :65–70, 2003.

[311] Molly S. Thomas, Jason S. Mitchell, Christopher C. DeNucci, Amanda L. Martin, and Yoji Shimizu. The p110γ isoform of phosphatidylinositol 3-kinase regulates migration of effector CD4 T lymphocytes into peripheral inflammatory sites. *J. Leukoc. Biol.*, 84(3) :814–823, 2008.

[312] A. M. Hasan, M. Mourtada-Maarabouni, M. S. Hameed, G. T. Williams, and G. Dent. Phosphoinositide 3-kinase γ mediates chemotactic responses of human eosinophils to platelet-activating factor. *Int. Immunopharmacol.*, 10(9) :1017–1021, 2010.

[313] L. H. Saal, K. Holm, M. Maurer, et al. PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. *Cancer Res.*, 65(7) :2554–2559, 2005.

[314] I. G. Campbell, S. E. Russell, D. Y. Choong, et al. Mutation of the PIK3CA gene in ovarian and breast cancer. *Cancer Res.*, 64(21) :7678–7681, 2004.

[315] S. Y. Li, M. Rong, F. Grieu, and B. Iacopetta. PIK3CA mutations in breast cancer are associated with poor outcome. *Breast Cancer Res. Treat.*, 96(1) :91–95, 2006.

[316] S. Ogino, P. Lochhead, E. Giovannucci, et al. Discovery of colorectal cancer PIK3CA mutation as potential predictive biomarker : power and promise of molecular pathological epidemiology. *Oncogene*, 33(23) :2949–2955, 2014.

[317] X. Liao, T. Morikawa, P. Lochhead, et al. Prognostic role of PIK3CA mutation in colorectal cancer : cohort study and literature review. *Clin. Cancer Res.*, 18(8) :2257–2268, 2012.

[318] Y. Samuels, Z. Wang, A. Bardelli, et al. High frequency of mutations of the PIK3CA gene in human cancers. *Science*, 304(5670) :554, 2004.

[319] Z. Ming, D. Jiang, Q. Hu, et al. Diagnostic application of PIK3CA mutation analysis in Chinese esophageal cancer patients. *Diagn. Pathol.*, 9 :153, 2014.

[320] H. Shigaki, Y. Baba, M. Watanabe, et al. PIK3CA mutation is associated with a favorable prognosis among patients with curatively resected esophageal squamous cell carcinoma. *Clin. Cancer Res.*, 19(9) :2451–2459, 2013.

[321] D. Kong and T. Yamori. Phosphatidylinositol 3-kinase inhibitors : promising drug candidates for cancer therapy. *Cancer Sci.*, 99(9) :1734–1740, 2008.

[322] A. Arcaro and M. P. Wymann. Wortmannin is a potent phosphatidylinositol 3-kinase inhibitor : the role of phosphatidylinositol 3,4,5-trisphosphate in neutrophil responses. *Biochem. J.*, 296 (Pt 2) :297–301, 1993.

[323] C. J. Vlahos, W. F. Matter, K. Y. Hui, and R. F. Brown. A specific inhibitor of phosphatidylinositol 3-kinase, 2-(4-morpholinyl)-8-phenyl-4H-1-benzopyran-4-one (LY294002). *J. Biol. Chem.*, 269(7) :5241–5248, 1994.

[324] K. D. Puri and M. R. Gold. Selective inhibitors of phosphoinositide 3-kinase delta : modulators of B-cell function with potential for treating autoimmune inflammatory diseases and B-cell malignancies. *Front. Immunol.*, 3 :256, 2012.

[325] Y. Posor, M. Eichhorn-Gruenig, D. Puchkov, et al. Spatiotemporal control of endocytosis by phosphatidylinositol-3,4-bisphosphate. *Nature*, 499(7457) :233–237, 2013.

[326] S. J. Turner, J. Domin, M. D. Waterfield, S. G. Ward, and J. Westwick. The CC chemokine monocyte chemotactic peptide-1 activates both the class I p85/p110 phosphatidylinositol 3-kinase and the class II PI3K-C2α. *J. Biol. Chem.*, 273(40) :25987–25995, 1998.

[327] C. Ktori, P. R. Shepherd, and L. O'Rourke. TNF-α and leptin activate the α-isoform of class II phosphoinositide 3-kinase. *Biochem. Biophys. Res. Comm.*, 306(1) :139–143, 2003.

[328] T. Maffucci, F. T. Cooke, F. M. Foster, et al. Class II phosphoinositide 3-kinase defines a novel signaling pathway in cell migration. *J. Cell Biol.*, 169(5) :789–799, 2005.

[329] K. Yoshioka, K. Yoshida, H. Cui, et al. Endothelial PI3K-C2α, a class II PI3K, has an essential role in angiogenesis and vascular barrier function. *Nat. Med.*, 18(10) :1560–1569, 2012.

[330] I. Franco, F. Gulluni, C. C. Campa, et al. PI3K class II α controls spatially restricted endosomal PtdIns3P and Rab11 activation to promote primary cilium function. *Dev. Cell*, 28(6) :647–658, 2014.

[331] A. Russo, M. Nazir Okur, M. Bosland, and J. P. O'Bryan. Phosphatidylinositol 3-kinase, class 2 beta (PI3KC2β) isoform contributes to neuroblastoma tumorigenesis. *Cancer Lett.*, 359(2) : 262–268, 2015.

[332] L. Braccini, E. Ciraolo, C. C. Campa, et al. PI3K-C2γ is a Rab5 effector selectively controlling endosomal Akt2 activation downstream of insulin signalling. *Nat. Commun.*, 6 :7400, 2015.

[333] B. Ravikumar, S. Sarkar, J. E. Davies, et al. Regulation of mammalian autophagy in physiology and pathophysiology. *Physiol. Rev.*, 90(4) :1383–1435, 2010.

[334] S. Kongara and V. Karantza. The interplay between autophagy and ROS in tumorigenesis. *Front. Oncol.*, 2 :171, 2012.

[335] W. Martina, J. Justin, and A. T. Sharon. Autophagosome formation–the role of ULK1 and Beclin1-PI3KC3 complexes in setting the stage. *Semin. Cancer Biol.*, 23(5) :301–309, 2013.

[336] P. V. Schu, K. Takegawa, M. J. Fry, et al. Phosphatidylinositol 3-kinase encoded by yeast VPS34 gene essential for protein sorting. *Science*, 260(5104) :88–91, 1993.

[337] K. Takegawa, D. B. DeWald, and S. D. Emr. *Schizosaccharomyces pombe* Vps34p, a phosphatidylinositol-specific PI 3-kinase essential for normal cell growth and vacuole morphology. *J. Cell Sci.*, 108 (Pt 12) :3745–3756, 1995.

[338] R. Eck, A. Bruckmann, R. Wetzker, and W. Künkel. A phosphatidylinositol 3-kinase of Candida albicans : molecular cloning and characterization. *Yeast*, 16(10) :933–944, 2000.

[339] K. Zhou, K. Takegawa, S. D. Emr, and R. A. Firtel. A phosphatidylinositol (PI) kinase gene family in *Dictyostelium discoideum* : biological roles of putative mammalian p110 and yeast Vps34p PI 3-kinase homologs during growth and development. *Mol. Cell. Biol.*, 15(10) :5645–5656, 1995.

[340] P. Welters, K. Takegawa, S. D. Emr, and M. J. Chrispeels. AtVPS34, a phosphatidylinositol 3-kinase of *Arabidopsis thaliana*, is an essential protein with homology to a calcium-dependent lipid binding domain. *Proc. Natl. Acad. Sci. USA*, 91(24) :11398–11402, 1994.

[341] Q. Jiang, L. Zhao, J. Dai, and Q. Wu. Analysis of autophagy genes in microalgae : Chlorella as a potential model to study mechanism of autophagy. *PLoS One*, 7(7) :e41826, 2012.

[342] J. M. Backer. The regulation and function of Class III PI3Ks : novel roles for Vps34. *Biochem. J.*, 410(1) :1–17, 2008.

[343] L. Roggo, V. Bernard, A. L. Kovacs, et al. Membrane transport in *Caenorhabditis elegans* : an essential role for VPS34 at the nuclear membrane. *EMBO J.*, 21(7) :1673–1683, 2002.

[344] C. Linassier, L. K. MacDougall, J. Domin, and M. D. Waterfield. Molecular cloning and biochemical characterization of a Drosophila phosphatidylinositol-specific phosphoinositide 3-kinase. *Biochem. J.*, 321 (Pt 3) :849–856, 1997.

[345] S. Jean and A. A. Kiger. Classes of phosphoinositide 3-kinases at a glance. *J. Cell Sci.*, 127(Pt 5) :923–928, 2014.

[346] B. Vanhaesebroeck, L. Stephens, and P. Hawkins. PI3K signalling : the path to discovery and understanding. *Nat. Rev. Mol. Cell Biol.*, 13(3) :195–203, 2012.

[347] S. E. Wilkowsky, M. A. Barbieri, P. Stahl, and E.L. Isola. *Trypanosoma cruzi* : phosphatidylinositol 3-kinase and protein kinase B activation is associated with parasite invasion. *Exp. Cell Res.*, 264(2) :211–218, 2001.

[348] J.H. Quan, G. H. Cha, W. Zhou, et al. Involvement of PI 3 kinase/Akt-dependent Bad phosphorylation in *Toxoplasma gondii*-mediated inhibition of host cell apoptosis. *Exp. Parasitol.*, 133(4) :462–471, 2013.

[349] W. Daher, J. Morlon-Guyot, L. Sheiner, et al. Lipid kinases are essential for apicoplast homeostasis in *Toxoplasma gondii*. *Cell. Microbiol.*, 17(4) :559–578, 2015.

[350] S. Merlot and R. A. Firtel. Leading the way : Directional sensing through phosphatidylinositol 3-kinase and other signaling pathways. *J. Cell Sci.*, 116(Pt 17) :3471–3478, 2003.

[351] C. A. Parent, B. J. Blacklock, W. M. Froehlich, D. B. Murphy, and P. N. Devreotes. G protein signaling events are activated at the leading edge of chemotactic cells. *Cell*, 95(1) :81–91, 1998.

[352] A. Levchenko and P. A. Iglesias. Models of eukaryotic gradient sensing : application to chemotaxis of amoebae and neutrophils. *Biophys. J.*, 82(1 Pt 1) :50–63, 2002.

[353] J. H. Stack, D. B. DeWald, K. Takegawa, and S. D. Emr. Vesicle-mediated protein transport : regulatory interactions between the Vps15 protein kinase and the Vps34 PtdIns 3-kinase essential for protein sorting to the vacuole in yeast. *J. Cell Biol.*, 129(2) :321–334, 1995.

[354] K. Picazarri, K.and Nakada-Tsukui and T. Nozaki. Autophagy during proliferation and encystation in the protozoan parasite Entamoeba invadens. *Infect. Immun.*, 76(1) :278–288, 2008.

[355] J. Cheng, A. Fujita, H. Yamamoto, et al. Yeast and mammalian autophagosomes exhibit distinct phosphatidylinositol 3-phosphate asymmetries. *Nat. Commun.*, 5 :3207, 2014.

[356] X. Zhuang and L. Jiang. Autophagosome biogenesis in plants : roles of SH3P2. *Autophagy*, 10 (4) :704–705, 2014.

[357] M. N. Rai, V. Sharma, S. Balusu, and R. Kaur. An essential role for phosphatidylinositol 3-kinase in the inhibition of phagosomal maturation, intracellular survival and virulence in Candida glabrata. *Cell. Microbiol.*, 17(2) :269–287, 2015.

[358] T. Kawashima, M. Tokuoka, S. Awazu, N. Satoh, and Y. Satou. A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. VIII. Genes for PI3K signaling and cell cycle. *Dev. Genes Evol.*, 213(5-6) :284–290, 2003.

[359] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 44(D1) :D7–D19, 2016.

[360] F. Ronquist and J. P. Huelsenbeck. MrBayes 3 : Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12) :1572–1574, 2003.

[361] H. Philippon, C. Brochier-Armanet, and G. Perrière. Evolutionary history of phosphatidylinositol-3-kinases : ancestral origin in eukaryotes and complex duplication patterns. *BMC Evol. Biol.*, 15 :226, 2015.

[362] UniProt Consortium. UniProt : a hub for protein information. *Nucleic Acids Res.*, 43(Database issue) :D204–D212, 2015.

[363] F. Cunningham, M. R. Amode, D. Barrell, et al. Ensembl 2015. *Nucleic Acids Res.*, 43(Database issue) :D662–669, 2015.

[364] M. Gouy and S. Delmotte. Remote access to ACNUC nucleotide and protein sequence databases at PBIL. *Biochimie*, 90(4) :555–562, 2008.

[365] E. W. Sayers, T. Barrett, D. A. Benson, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 38(Database issue) :D5–16, 2010.

[366] J. D. Thompson, F. Plewniak, R. Ripp, J. C. Thierry, and O. Poch. Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.*, 314(4) :937–951, 2001.

[367] D. Darriba, G. L. Taboada, R. Doallo, and D. Posada. ProtTest 3 : fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8) :1164–1165, 2011.

[368] G. Schwartz. Estimating the dimension of a model. *Ann. Stat.*, 6(2) :461–464, 1978.

[369] S. Guindon, J. F. Dufayard, V. Lefort, et al. New algorithms and methods to estimate maximum-likelihood phylogenies : assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3) :307–321, 2010.

[370] R. D. Finn, A. Bateman, J. Clements, et al. Pfam : the protein families database. *Nucleic Acids Res.*, 42(Database issue) :D222–D230, 2014.

[371] Sean R. Eddy. Accelerated profile HMM searches. *PLoS Comput. Biol.*, 7(10) :e1002195, 2011.

[372] A. Marchler-Bauer, S. Lu, J. B. Anderson, et al. CDD : a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*, 39(Database issue) :D225–D229, 2011.

[373] B. Vanhaesebroeck, S. J. Leevers, G. Panayotou, and M. D. Waterfield. Phosphoinositide 3-kinases : a conserved family of signal transducers. *Trends Biochem. Sci.*, 22(7) :267–272, 1997.

[374] M. Cully, H. You, A. J. Levine, and T. W. Mak. Beyond PTEN mutations : the PI3K pathway as an integrator of multiple inputs during tumorigenesis. *Nat. Rev. Cancer*, 6(3) :184–192, 2006.

[375] R. Fritsch, I. de Krijger, K. Fritsch, et al. RAS and RHO families of GTPases directly regulate distinct phosphoinositide 3-kinase isoforms. *Cell*, 153(5) :1050–1063, 2013.

[376] R. Dhand, K. Hara, I. Hiles, et al. PI 3-kinase : structural and functional analysis of intersubunit interactions. *EMBO J.*, 13(3) :511–521, 1994.

[377] A. Klippel, J. A. Escobedo, Q. Hu, and L. T. Williams. A region of the 85-kilodalton (kDa) subunit of phosphatidylinositol 3-kinase binds the 110-kDa catalytic subunit in vivo. *Mol. Cell. Biol.*, 13(9) :5560–5566, 1993.

[378] K. H. Holt, L. Olson, W. S. Moye-Rowley, and J. E. Pessin. Phosphatidylinositol 3-kinase activation is mediated by high-affinity interactions between distinct domains within the p110 and p85 subunits. *Mol. Cell. Biol.*, 14(1) :42–49, 1994.

[379] F. Delsuc, G. Tsagkogeorga, N. Lartillot, and H. Philippe. Additional molecular support for the new chordate phylogeny. *Genesis*, 46(11) :592–604, 2008.

[380] J. Collén, B. Porcel, W. Carré, et al. Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. *Proc. Natl. Acad. Sci. USA*, 110(13) :5247–5252, 2013.

[381] H. Nozaki, H. Takano, O. Misumi, et al. A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. *BMC Biol.*, 5 :28, 2007.

[382] K. Jain, K. Krause, F. Grewe, et al. Extreme features of the *Galdieria sulphuraria* organellar genomes : a consequence of polyextremophily ? *Genome Biol. Evol.*, 7(1) :367–380, 2015.

[383] E. Hirsch, V. L. Katanaev, C. Garlanda, et al. Central role for G protein-coupled phosphoinositide 3-kinase γ in inflammation. *Science*, 287(5455) :1049–1053, 2000.

[384] T. Sasaki, J. Irie-Sasaki, R. G. Jones, et al. Function of PI3Kγ in thymocyte development, T cell activation, and neutrophil migration. *Science*, 287(5455) :1040–1046, 2000.

[385] V. Silió, J. Redondo-Muñoz, and A. C. Carrera. Phosphoinositide 3-kinase β regulates chromosome segregation in mitosis. *Mol. Biol. Cell*, 23(23) :4526–4542, 2012.

[386] D. J. Klionsky. Autophagy : from phenomenology to molecular understanding in less than a decade. *Nat. Rev. Mol. Cell. Biol.*, 8(11) :931–937, 2007.

[387] D. J. Klionsky. Autophagy revisited : a conversation with Christian de Duve. *Autophagy*, 4(6) : 740–743, 2008.

[388] R. Ojha, S. Bhattacharyya, and S. K. Singh. Autophagy in cancer stem cells : A potential link between chemoresistance, recurrence, and metastasis. *Biores. Open Access*, 4(1) :97–108, 2015.

[389] N. N. Shan, L. L. Dong, X. M. Zhang, X. Liu, and Y. Li. Targeting autophagy as a potential therapeutic approach for immune thrombocytopenia therapy. *Crit. Rev. Oncol. Hematol.*, 100 : 11–15, 2016.

[390] K. Wang. Autophagy and apoptosis in liver injury. *Cell Cycle*, 14(11) :1631–1642, 2015.

[391] S. Dash, S. Chava, P. K. Chandra, et al. Autophagy in hepatocellular carcinomas : from pathophysiology to therapeutic response. *Hepat. Med.*, 8 :9–20, 2016.

[392] B. Hoang, A. Benavides, Y. Shi, P. Frost, and A. Lichtenstein. Effect of autophagy on multiple myeloma cell viability. *Mol. Cancer Ther.*, 8(7) :1974–1984, 2009.

[393] H. A. Ekiz, G. Can, and Y. Baran. Role of autophagy in the progression and suppression of leukemias. *Crit. Rev. Oncol. Hematol.*, 81(3) :275–285, 2012.

[394] F. Burada, E. R. Nicoli, M. E. Ciurea, et al. Autophagy in colorectal cancer : An important switch from physiology to pathology. *World J. Gastrointest. Oncol.*, 7(11) :271–284, 2015.

[395] B. Cosway and P. Lovat. The role of autophagy in squamous cell carcinoma of the head and neck. *Oral Oncol.*, 54 :1–6, 2016.

[396] P. Maycotte and A. Thorburn. Targeting autophagy in breast cancer. *World J. Clin. Oncol.*, 5 (3) :224–240, 2014.

[397] S. Fulda and D. Kögel. Cell death by autophagy : emerging molecular mechanisms and implications for cancer therapy. *Oncogene*, 34(40) :5105–5113, 2015.

[398] T. A. Eyre, G. P. Collins, A. H. Goldstone, and K. Cwynarski. Time now to TORC the TORC ? New developments in mTOR pathway inhibition in lymphoid malignancies. *Br. J. Haematol.*, 166(3) :336–351, 2014.

[399] S. Wesselborg and B. Stork. Autophagy signal transduction by ATG proteins : from hierarchies to networks. *Cell. Mol. Life Sci.*, 72(24) :4721–4757, 2015.

[400] J. L. Schneider and A. M. Cuervo. Autophagy and human disease : emerging themes. *Curr. Opin. Genet. Dev.*, 26 :16–23, 2014.

[401] M. Tsukada and Y. Ohsumi. Isolation and characterization of autophagy-defective mutants of *Saccharomyces cerevisiae*. *FEBS Lett.*, 333(1-2) :169–174, 1993.

[402] D. J. Klionsky, J. M. Cregg, W. A Dunn, Jr, et al. A unified nomenclature for yeast autophagy-related genes. *Dev. Cell*, 5(4) :539–545, 2003.

[403] C. Ward, N. Martinez-Lopez, E. G. Otten, et al. Autophagy, lipophagy and lysosomal lipid storage disorders. *Biochim. Biophys. Acta*, 1864(4) :269–284, 2016.

[404] W. G. van Doorn and A. Papini. Ultrastructure of autophagy in plant cells : a review. *Autophagy*, 9(12) :1922–1936, 2013.

[405] S. Michaeli, G. Galili, P. Genschik, A. R. Fernie, and T. Avin-Wittenberg. Autophagy in Plants - What's New on the Menu ? *Trends Plant Sci.*, 21(2) :134–144, 2016.

[406] N. N. Noda and Y. Fujioka. Atg1 family kinases in autophagy initiation. *Cell. Mol. Life Sci.*, 72(16) :3083–3096, 2015.

[407] T. Hara and N. Mizushima. Role of ULK-FIP200 complex in mammalian autophagy : FIP200, a counterpart of yeast Atg17 ? *Autophagy*, 5(1) :85–87, 2009.

[408] H. Zhang, J. T. Chang, B. Guo, et al. Guidelines for monitoring autophagy in *Caenorhabditis elegans*. *Autophagy*, 11(1) :9–27, 2015.

[409] B. C. Stöver and K. F. Müller. TreeGraph 2 : combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics*, 11 :7, 2010.

[410] P. Shannon, A. Markiel, O. Ozier, et al. Cytoscape : a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11) :2498–2504, 2003.

[411] J. Heitman, N. R. Movva, and M. N. Hall. Targets for cell cycle arrest by the immunosuppressant rapamycin in yeast. *Science*, 253(5022) :905–909, 1991.

[412] B. Boussau, G. J. Szöllosi, L. Duret, et al. Genome-scale coestimation of species and gene trees. *Genome Res.*, 23(2) :323–330, 2013.

[413] T. Bigot, V. Daubin, F. Lassalle, and G. Perrière. TPMS : a set of utilities for querying collections of gene trees. *BMC Bioinformatics*, 14 :109, 2013.

[414] E. Becker, B. Robisson, C. S. E. Chapple, A. Guénoche, and C. Brun. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, 28(1) :84–90, 2012.

[415] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, 57(1) :289–300, 1995.

[416] PI3K / Akt Signaling Pathway. http://www.cellsignal.com/common/content/content.jsp?id=pathways-akt-signaling.

[417] S. Y. Lin, T. Y. Li, Q. Liu, et al. Protein phosphorylation-acetylation cascade connects growth factor deprivation to autophagy. *Autophagy*, 8(9) :1385–1386, 2012.

[418] P. Sini, D. James, C. Chresta, and S. Guichard. Simultaneous inhibition of mTORC1 and mTORC2 by mTOR kinase inhibitor AZD8055 induces autophagy and cell death in cancer cells. *Autophagy*, 6(4) :553–554, 2010.

[419] N. Oshiro, J. Rapley, and J. Avruch. Amino acids activate mammalian target of rapamycin (mTOR) complex 1 without changing Rag GTPase guanyl nucleotide charging. *J. Biol. Chem.*, 289(5) :2658–2674, 2014.

[420] J. A. Martina, Y. Chen, M. Gucek, and R. Puertollano. MTORC1 functions as a transcriptional regulator of autophagy by preventing nuclear transport of TFEB. *Autophagy*, 8(6) :903–914, 2012.

[421] T. B. Huber, G. Walz, and E. W. Kuehn. mTOR and rapamycin in the kidney : signaling and therapeutic implications beyond immunosuppression. *Kidney Int.*, 79(5) :502–511, 2011.

[422] H. Roca, Z. S. Varsos, K. Mizutani, and K. J. Pienta. CCL2, survivin and autophagy : new links with implications in human cancer. *Autophagy*, 4(7) :969–971, 2008.

[423] H. Roca, Z. Varsos, and K. J. Pienta. CCL2 protects prostate cancer PC3 cells from autophagic death via phosphatidylinositol 3-kinase/AKT-dependent survivin up-regulation. *J. Biol. Chem.*, 283(36) :25057–25073, 2008.

[424] M C. Maiuri, E. Zalckvar, A. Kimchi, and G. Kroemer. Self-eating and self-killing : crosstalk between autophagy and apoptosis. *Nat. Rev. Mol. Cell. Biol.*, 8(9) :741–752, 2007.

[425] E. Rozengurt. Mechanistic target of rapamycin (mTOR) : a point of convergence in the action of insulin/IGF-1 and G protein-coupled receptor agonists in pancreatic cancer cells. *Front. Physiol.*, 5 :357, 2014.

[426] C. Muñoz-Pinedo, N. El Mjiyad, and J. E. Ricci. Cancer metabolism : current perspectives and future directions. *Cell Death Dis.*, 3 :e248, 2012.

[427] J. A. McCubrey, L. S. Steelman, F. E. Bertrand, et al. Multifaceted roles of GSK-3 and Wnt/β-catenin in hematopoiesis and leukemogenesis : opportunities for therapeutic intervention. *Leukemia*, 28(1) :15–33, 2014.

[428] D. Tang, R. Kang, K. M. Livesey, et al. Endogenous hmgb1 regulates autophagy. *J. Cell Biol.*, 190(5) :881–892, 2010.

[429] N. Chen and J. Debnath. Autophagy and tumorigenesis. *FEBS Lett.*, 584(7) :1427–1435, 2010.

[430] H. Cheong, C. Lu, T. Lindsten, and C. B. Thompson. Therapeutic targets in cancer cell metabolism and autophagy. *Nat. Biotechnol.*, 30(7) :671–678, 2012.

[431] S. Alers, A. S. Löffler, S. Wesselborg, and B. Stork. Role of AMPK-mTOR-Ulk1/2 in the regulation of autophagy : cross talk, shortcuts, and feedbacks. *Mol. Cell. Biol.*, 32(1) :2–11, 2012.

[432] Y. M. Kim and D. H. Kim. dRAGging amino acid-mTORC1 signaling by SH3BP4. *Mol. Cells*, 35(1) :1–6, 2013.

[433] Y. M. Kim, M. Stone, T. H. Hwang, et al. SH3BP4 is a negative regulator of amino acid-Rag GTPase-mTORC1 signaling. *Mol. Cell*, 46(6) :833–846, 2012.

[434] F. Nazio and F. Cecconi. mTOR, AMBRA1, and autophagy : an intricate relationship. *Cell Cycle*, 12(16) :2524–2525, 2013.

[435] C. A. Mercer, A. Kaliappan, and P. B. Dennis. A novel, human Atg13 binding protein, Atg101, interacts with ULK1 and is essential for macroautophagy. *Autophagy*, 5(5) :649–662, 2009.

Annexes

Annexe A

Article de BATfinder soumis

SOFTWARE

BATfinder: alternative transcript selection in multiple sequence alignments

Héloïse Philippon[1], Alexia Souvane[1] and Guy Perrière[1*]

*Correspondence: guy.perriere@univ-lyon1.fr
[1]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard – Lyon 1, 43 bd. du 11 Novembre 1918, 69622 Villeurbanne, France
Full list of author information is available at the end of the article

Abstract

Background: The reliability of a molecular phylogeny strongly depends on the quality of the multiple sequence alignment used. In the case of analyses involving sequences in which alternative transcripts are available, it is necessary to select the transcripts that are the most suited for phylogenetic reconstruction, *i.e.* the ones allowing to obtain the best possible multiple sequence alignment. Indeed, selecting the longest isoform is usually not an optimal solution and only a handful of programs allowing to perform this operation automatically are available.

Results: In this paper we present BATfinder (Best Aligned Transcript finder), a software allowing to select alternative transcripts in a protein sequence dataset. The selection is based on the computation of a sum-of-pairs score, the set of transcripts allowing to obtain the best scores being the one retained in the alignment. Computation of this score involves a bootstrap procedure followed by the construction of a set of guide trees that are then used to build perturbed alignments.

Conclusions: Implemented in C/C++ and easy to install, BATfinder is a freely available software that runs on Linux and MacOSX operating systems. It provides an alternative to the usual selection of the longest transcript and gives better results in term of alignment quality in the framework of phylogenetic reconstruction. Optimized for parallel computing it is also faster than its only direct equivalent when multi-threading is enabled.

Keywords: Alternative transcripts; bootstrap; sum-of-pairs score; phylogeny

Background

The selection of homologous sequences is a crucial step in molecular phylogeny. For eukaryotic species, standard search for homologs often leads to the obtention of datasets containing a lot of alternative transcripts. About 20% of plant genes are affected by alternative splicing [1] and Metazoa are even more concerned [2]. For instance, about 90% of human genes have more than one transcript [3]. Therefore, the selection of one transcript per gene is mandatory to avoid phylogenetic redundancy and biases during the multiple alignment trimming step. Two simple approaches consist in performing a random selection [4, 5] or keeping the longest transcript [6, 7]. The problem is that there is no justification for the former and the latter usually leads to the introduction of many gaps in the alignment, which can be problematic when considering sites homology.

There are presently two tools dedicated to transcript selection: PALO [8] and Guidance [9]. The first one uses a sequence length criteria and the second one alignment scores. Nevertheless, PALO is specifically designed to be used on protein

sequences from the Ensembl database [10] while Guidance is a rather general tool devoted to Multiple Sequence Alignments (MSAs) quality analysis.

In that context, we developed BATfinder in order to provide a tool specifically devoted to the selection of alternate transcripts. BATfinder uses the same scoring function as the one implemented in Guidance but is faster and allows introducing special options to penalize gaps and/or short transcripts. Indeed, it appears that the choice of an appropriate alternative transcript requires to take into account three criteria: i) the isoform selected must have the highest possible similarity with sequences from closely related species; ii) it must minimize the number of gaps introduced in the alignment; and iii) it must be long enough [11]. The options implemented in BATfinder allows to tune the balance between those three criteria.

Implementation

Algorithm

BATfinder is based on the sum-of-pairs score introduced by Thompson *et al.* [12]. The program is divided into three main steps. The first one consists in the alignment of the input protein dataset to generate a *reference alignment*. The second step is the generation of a set of *perturbed alignments* built through a bootstrap procedure applied on the reference alignment. Finally, the third step is the computation of the sum-of-pairs score itself using those perturbed alignments. Due to the scoring methodology, the reference alignment and the perturbed alignments must be computed with the same program. For each sequence, BATfinder returns a score comprised between 0 and 1 representing how well this sequence is aligned relatively to the others in the reference alignment.

Perturbed alignments computation

The dataset provided by the user is first aligned with one of the three multiple alignment programs integrated in the BATfinder distribution: ClustalO [13], Mafft [14] or Muscle [15] (Fig. 1). Those programs have been chosen because they allow the input of an user-provided guide tree when building a MSA. Integration of other programs having this functionality is theoretically possible but would require the modification of the BATfinder source code and its recompilation. By default, the program used is Muscle.

The resulting reference alignment is then bootstrapped and an evolutionary distances matrix is computed for each bootstrap replicate. This computation is realized with the FastDist program which has been designed for parallel computing thanks to the use of OpenMP [16] and Eigen libraries [17]. It is therefore much faster than equivalent programs such as ProtDist [18] or BppDist [19]. FastDist allows to compute evolutionary distances using the most common amino acid substitution models: Poisson, PAM (or its Kimura approximation), JTT, BLOSUM62, WAG and LG. FastDist output uses the standard Phylip format for distance matrices. By default, the substitution model used by BATfinder is LG and the number of bootstrap replicates is 20. This default number has been chosen because it usually allows to reach a stable sequence selection on the datasets used to test the program (data not shown).

After the computation of the bootstrap distance matrices, BioNJ [20] is used to infer the corresponding guide trees. Those trees are then automatically rooted with

SeaView [21] (mid-point rooting). Perturbed alignments are then computed using the initial dataset and the guide trees.

Sum-of-pairs core computation

Figure 2 summarizes the procedure we used for the sum-of-pairs score computation. Let a be the reference alignment generated at the first step and containing ℓ sites for m sequences. Now let b ($1 \leq b \leq n$) be one of the n perturbed alignment generated from a. For each sequence i ($1 \leq i \leq m$) of length l_i from b, we can compute the pair score of the amino acid at position k ($1 \leq k \leq l_i$) as:

$$R_{ik}^{(b)} = \frac{1}{m_k - 1} \sum_{j=1, j \neq i}^{m} p_{ijk} \tag{1}$$

where p_{ijk} is equal to 1 or 0 whether or not the amino acid at position k of sequence i is facing the same amino acid of the position k' ($1 \leq k' \leq l_j$) of sequence j in the reference and perturbed alignment. Also, m_k is the number of sequences having a residue (*i.e.* not a gap) at this position in a.

From (1), it is possible to compute the average residue score over all bootstrap replicates as:

$$R_{ik} = \frac{1}{n} \sum_{b=1}^{n} R_{ik}^{(b)} \tag{2}$$

Finally, the sum-of-pairs score for sequence i is calculated by averaging the residues scores:

$$S_i = \frac{1}{l_i} \sum_{k=1}^{l_i} R_{ik} \tag{3}$$

Scores penalizations

BATfinder has an option (`-gap`) allowing to penalize the sequences introducing gaps in the reference alignment. With this option, p_{ijk} is weighted by the number of gaps present at position k in the reference alignment:

$$R_{ik}^{(b)} = \frac{1}{m_k - 1} \sum_{j=1, j \neq i}^{m} \frac{p_{ijk}}{m_g} \tag{4}$$

where $m_g = m - m_k$ is the number of gaps at this position in a. If there is no gap, then $m_g = 1$. Here, p_{ijk} is the same as before except it is fixed at -1 if there is no other residue at this position in the reference alignment.

There is also an option (`-short`) designed to penalize transcripts that are too short, even if they are well aligned. In this case, S_i is not longer calculated by dividing the sum of residue scores by the length of the sequence, but by the length of the reference alignment:

$$S_i = \frac{1}{\ell} \sum_{r=1}^{l_i} R_{ik} \tag{5}$$

By construction, those two weighting implies much smaller scores per sequences compared to the default.

Input and output files

BATfinder minimal input requirement is an unaligned protein sequence dataset in Fasta format (Fig. 3). The output is a text file containing the scores for each input sequence. Optionally, the user can provide a file in which the information on transcripts locus tag is given. In this case, BATfinder will also create a file in Fasta format containing the filtered dataset (*i.e.* in which only the transcript having the best score for a given gene is kept).

Datasets

An example dataset is included in the program distribution. The first file from this dataset (`example.fasta`) contains 59 homologs of the human AKTS1 protein taken from Ensembl and a local database of complete eukaryotic proteomes. Some of those homologs correspond to alternative transcripts and the second file (`transcript_file.txt`), contains information on the alternative transcripts provenance. In this file, the first column corresponds to the sequence names from the sequence file and the second column to an identifier allowing to associate a set of transcripts to a given gene. For example, the three lines below:

```
>ENSP00000375710|Homo_sapiens ENSG00000204673
>ENSP00000375711|Homo_sapiens ENSG00000204673
>ENSP00000375706|Homo_sapiens ENSG00000204673
```

allows to specify that sequences ENSP00000375710, ENSP00000375711 and ENSP-00000375706 are in fact alternative transcripts that originate from a single gene, identified as ENSG00000204673.

To test the performances of BATfinder, we built datasets containing homologs of the different proteins involved in the human autophagy pathway. We used the procedure and databases described in [11] for gathering the different sets of homologs. Among the 45 sets of homologs collected, four were very large and we were unable to align them with Mafft on our 16 Gbytes Virtual Machine (VM) used for testing, so we decided to discard them. The remaining 41 datasets contained from 42 to 1484 sequences and the corresponding reference MSAs built with Mafft ranged from 280 to 11175 sites. Alternative transcripts represented from 2.62% to 33.54% of those datasets content (see Additional File 1 for details). The complete 41 datasets can be downloaded from the program web site.

Results and discussion

Example dataset

Among the 59 proteins from the example dataset, 26 are alternative transcripts that come from 8 different genes. Filtering with BATfinder results in the selection of 8 transcripts among those 26 but the sequences selected are different depending on the parameters used. Table 1 shows the eight transcripts selected when using BATfinder with Mafft, JTT substitution model and: i) the default parameters; ii) the `-gap` option; iii) the `-short` option; and iv) both `-gap` and `-short` options.

For example, the best transcript selected for the atlantic cod *Gadus morhua* is either ENSGMOP00000007742 (257 AA), ENSGMOP00000007705 (353 AA) or ENSGMOP00000007715 (314 AA), depending on the parameters used. A visual inspection of the alignment shows that the first one is well aligned but also the shorter of the three alternative transcripts, this is why it is selected when using the default parameters and the `-gap` option. On the other hand, the ENSGMOP00000007705 sequence introduces two gap regions in the majority of other sequences at the N-terminal region; but it is also the longest one and is therefore selected when using the `-short` option. Finally, the ENSGMOP00000007715 transcript falls between the both previous sequences with no and some residues in the first and second gap region, respectively. It is therefore selected when using the combination of `-gap` and `-short` options.

For the mouse, four different transcripts are selected depending on the parameters chosen. All are pretty well aligned, but ENSMUSP00000103517 (328 AA) introduces a long gap in the N-terminal region (with no residue in other sequences) and corresponds to the longest alternative transcript of ENSMUSG00000011096. ENSMUSP00000116541 is very well aligned but is also the shorter transcript (only 95 AA) and lacks a lot of residues in the C-terminal region. Finally, ENSMUSP00000103514 and ENSMUSP00000049764 shared an homologous segment of 257 AA but the second transcript has 27 additional well aligned residues at the beginning of the MSA. This is why the ENSMUSP00000049764 transcript is selected when using both `-gap` and `-short` options.

To compare the results obtained with the different settings we used the tree length criteria described in the **Phylogenetic quality** subsection. For this example dataset, we can see that the `-gap` option gives the shortest tree (*i.e.* the best result) while the selection of the longest transcripts gives the longest tree.

Alignments quality

In order to test the efficiency of BATfinder in terms of alignments quality, we used a phylogenetic approach. For that purpose, we inferred the phylogenetic trees using the filtered alignments generated by BATfinder from the 41 test datasets and we compared them to the trees inferred on the alignments generated when the longest transcripts were used. All alignments were computed with Mafft and the phylogenetic reconstructions were performed with SeaView. Trees were built using observed divergences and the BioNJ algorithm.

To compare the quality of the alignments generated after BATfinder filtering to those obtained on the alignments generated by using the longest transcripts, we used the length of the trees built with BioNJ, *i.e.* the sum of all their respective branch length. Our hypothesis is that better alignments will result in trees having a smaller sum of branch lengths, especially when using observed divergence as the measure of evolutionary distance. On average, BATfinder filtering allows to obtain alignments resulting in significantly shorter trees than the ones obtained with the longest transcript selection. Indeed, Wilcoxon signed rank test for means comparison gives significant results when using default parameters ($P < 10^{-4}$) and the `-gap` option ($P = 1.19 \times 10^{-3}$). Differences are not statistically significant for the `-short` option and the combination of `-gap` and `-short` options.

When looking closely at the results, we found that a shorter tree is obtained with BATfinder for 38 datasets among 41. The parameters allowing to obtain the shortest tree with BATfinder were: i) the default parameters for 18 datasets; ii) the `-gap` option for 13 datasets; iii) the combination of the `-gap` and `-short` options for five datasets; and iv) the `-short` option for two datasets. Optimal parameters setting thus strongly depends on dataset characteristics and we recommend to try the different possibilities before making a choice.

Comparison with Guidance

As they use the same scoring scheme, we wanted to compare the performances of BATfinder and Guidance in terms of speed only. For both programs we used Mafft for computing the MSAs and JTT as the amino acid substitution model (Guidance default options). Also, the number of bootstrap replicates was set to 20. Computation were performed on an eight CPU Linux CentOS VM cadenced at 2.6 GHz and having 16 Gbytes of RAM. As both programs are multithreaded, we used an increasing number of CPU to see the effect of parallelization level on their relative performances. When only one or two CPUs were used, Guidance outperforms BATfinder (Fig. 4). On the other hand, BATfinder is usually faster than Guidance when the number of CPUs is larger or equal to three and the difference increases with this number. This is due to the fact that BATfinder has been optimized for parallel computing, especially the part devoted to evolutionary distances computation (FastDist program).

As an illustration, BATfinder is slower than Guidance when using one or two CPUs for the dataset containing the ATG13 homologs but, with the increasing number of CPUs, it becomes up to two times faster. For the dataset containing the SH3B4 homologs, BATfinder is from 3.6 (one CPU) to 13.3 (eight CPUs) times faster than Guidance. With four CPUs, this represents a gain in computation time of about eight hours.

Conclusion

BATfinder is a command line software designed for the automatic selection of alternative transcripts in the framework of phylogenetic analyses. Based on a sum-of-pairs score, it allows to obtain datasets that are optimized for tree reconstruction. The only other software that can be compared to BATfinder is Guidance but this program presents some limitations. First, it requires the independent installation of a broad range of tools (namely Perl, BioPerl and Ruby) while BATfinder is self-sufficient and is distributed with all the binaries required for its functioning. Then, it can only be run with JTT while BATfinder allows the use of all standard amino acids substitution models. On a practical point of view, Guidance is usually slower than BATfinder when multithreading is used. This point is probably linked to the fact that Guidance was not designed for alternative transcript selection but rather as a general tool for assessing MSA quality. With this broader purpose, Guidance has to compute many scores in addition to the sum-of-pairs, which diminish its performances in terms of speed relatively to BATfinder.

Availability and requirements

- **Project name:** BATfinder.
- **Project home page:** http://doua.prabi.fr/software/batfinder/
- **Operating systems:** Linux and MacOSX.
- **Programming language:** C/C++.
- **Shell commands:** cut, grep, sed and awk.
- **Other requirements:** Sequence alignment programs ClustalO v1.2.0, Muscle v3.6 and Mafft v7.266. Sequence alignment editor SeaView v4.4.1. Eigen v3.2 and (facultatively) OpenMP libraries. Binaries of sequence alignment programs and SeaView are included in Linux and MacOSX distributions, as well as the Eigen library.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
GP and HP conceived the software; AS, HP and GP implemented it. GP and HP wrote the manuscript. All authors read and approve the final manuscript.

Acknowledgements
We would like to thank the CNRS, the France Génomique consortium and the Région Rhône-Alpes, which was funding the Ph.D grant of HP. The different computations have been performed using the LBBE/PRABI cluster and the IFB cloud. We also thank Dominique Guyot for his advice and expertise on parallel computing.

Author details
[1]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard – Lyon 1, 43 bd. du 11 Novembre 1918, 69622 Villeurbanne, France. [2], , , .

References

1. Barbazuk, W.B., Fu, Y., McGinnis, K.M.: Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome Res. **18**, 1381–1392 (2008)
2. Gamazon, E.R., Stranger, B.E.: Genomics of alternative splicing: evolution, development and pathophysiology. Hum. Genet. **133**, 679–687 (2014)
3. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B.: Alternative isoform regulation in human tissue transcriptomes. Nature **456**, 470–476 (2008)
4. Hughes, A.L., Friedman, R.: The effect of branch lengths on phylogeny: an empirical study using highly conserved orthologs from mammalian genomes. Mol. Phylogenet. Evol. **45**, 81–88 (2007)
5. Zou, Z., Zhang, J.: Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol. Biol. Evol. **32**, 2085–2096 (2015)
6. Bakewell, M.A., Shi, P., Zhang, J.: More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc. Natl. Acad. Sci. USA **104**, 7489–7494 (2007)
7. Carneiro, M., Albert, F.W., Melo-Ferreira, J., Galtier, N., Gayral, P., Blanco-Aguiar, J.A., Villafuerte, R., Nachman, M.W., Ferrand, N.: Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. Mol. Biol. Evol. **29**, 1837–1849 (2012)
8. Villanueva-Cañas, J.L., Laurie, S., Albà, M.M.: Improving genome-wide scans of positive selection by using protein isoforms of similar length. Genome Biol. Evol. **5**, 457–467 (2013)
9. Penn, O., Privman, E., Landan, G., Graur, D., Pupko, T.: An alignment confidence score capturing robustness to guide tree uncertainty. Mol. Biol. Evol. **27**, 1759–1767 (2010)
10. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M.J., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R., Flicek, P.
11. Philippon, H., Brochier-Armanet, C., Perrière, G.: Evolutionary history of phosphatidylinositol-3-kinases: ancestral origin in eukaryotes and complex duplication patterns. BMC Evol. Biol. **15**, 226 (2015)
12. Thompson, J.D., Plewniak, F., Poch, O.: A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. **27**, 2682–2690 (1999)
13. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. **7**, 539 (2011)
14. Katoh, K., Standley, D.M.: MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30**, 772–780 (2013)
15. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**, 1792–1797 (2004)
16. The OpenMP® API Specification for Parallel Programming. http://openmp.org/

17. Eigen Template Library for Linear Algebra. http://eigen.tuxfamily.org/
18. Protdist – Program to Compute Distance Matrix from Protein Sequences. http://evolution.genetics.washington.edu/phylip/doc/protdist.html
19. Bio++ Program Suite. http://home.gna.org/bppsuite/
20. Gascuel, O.: BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**, 685–695 (1997)
21. Gouy, M., Guindon, S., Gascuel, O.: SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. **27**, 221–224 (2010)

Figures

Figure 1 BATfinder workflow. Schematic representation of the different steps performed by the BATfinder pipeline.

Figure 2 Score computation for two genes producing respectively three and two alternative transcripts. For a given residue k of a sequence i if, in the perturbed alignment b, this residue is facing the same residue k' of sequence j in the reference alignment, the score is set to 1 (light blue). If not, it is set to 0 (light red). The final score $R_{ik}^{(b)}$ is obtained by averaging all these pair scores (green). Then, the total score of a residue of a sequence corresponds to the mean of the $R_{ik}^{(b)}$ scores for all the n perturbed alignments (purple). Finally, the sequence score is computing by averaging the scores of all residues of the sequence (orange).

Figure 3 Input and output. Input files are represented by an unaligned protein sequence dataset (`example.fasta`) and (optionally) a file giving the correspondence between genes and their alternative transcripts (`transcript_file.txt`). Output files are the file containing the score of each input sequence (`Output.scores`), the reference (unfiltered) alignment in Fasta format (`Output.aln`) and (optionally) the filtered dataset with only one transcript per gene (`Output_filtered.fasta`).

Figure 4 Relative performance for BATfinder and Guidance. Proportions of faster runs for BATfinder and Guidance on the 41 protein datasets relatively the the number of CPUs used for computation.

Tables

Table 1 BATfinder results on the example dataset. This table lists the alternative transcript selected in each case. The last row corresponds to the total length of the tree inferred using observed divergences and BioNJ algorithm.

Additional Files

Additional file 1 – Characteristics of the 41 protein sequence datasets used to test BATfinder performances. In this table, UniProt identifiers correspond to the human proteins used as seed to gather eukaryotic homologs.

A.1 Vérification des corrélations

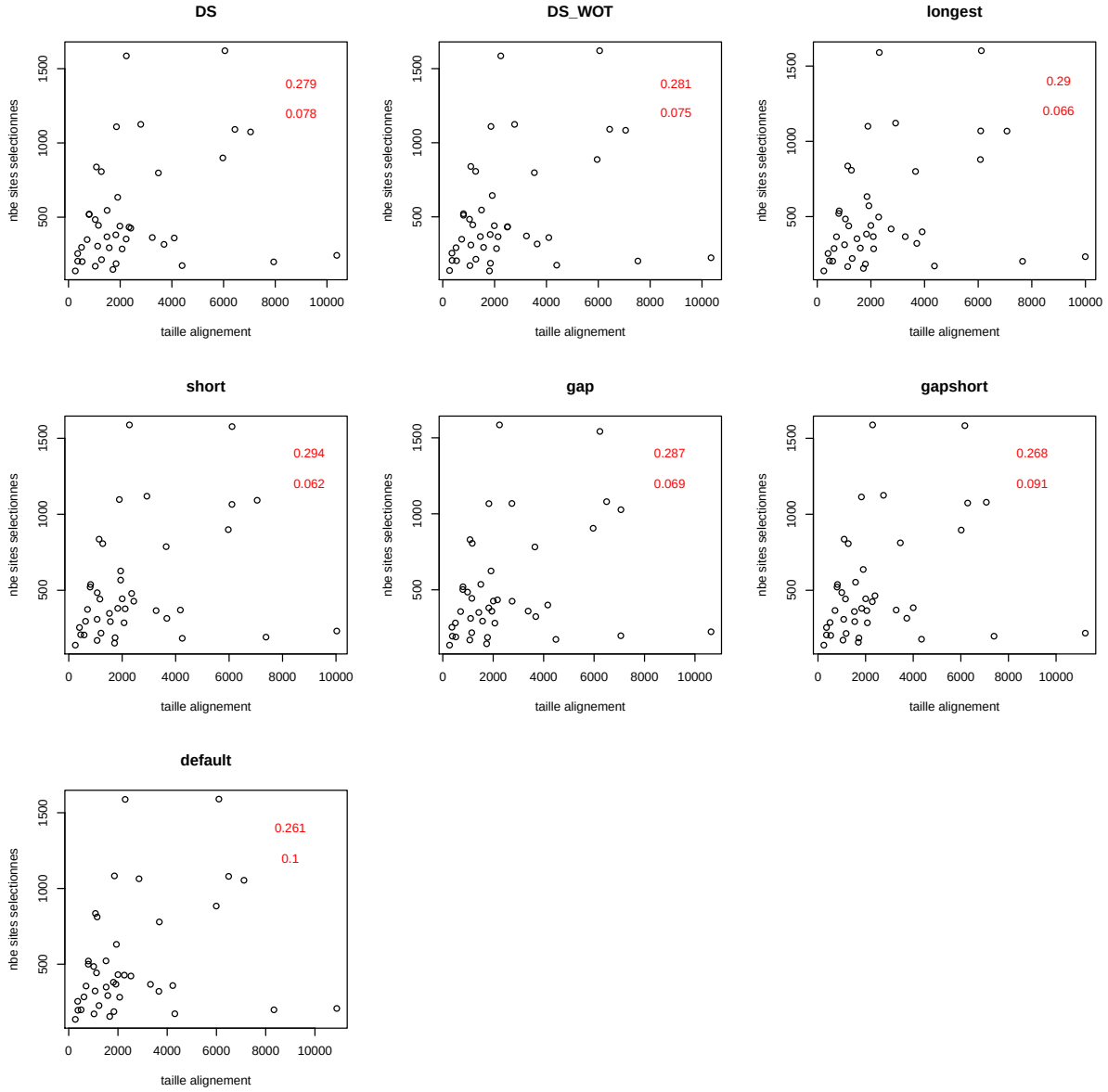

FIGURE *A.1* – ***Nombre de sites conservés par BMGE en fonction de la taille de l'alignement.*** Sélection de sites effectuée avec BMGE (options `-m BLOSUM30 -b 3 -g 0.4`). Pour chaque graphique, le coefficient de corrélation (première ligne) et la valeur p associée (seconde ligne) sont indiqués en rouge.

Annexe B

Article sur les PI3K

Philippon *et al. BMC Evolutionary Biology* (2015) 15:226
DOI 10.1186/s12862-015-0498-7

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Evolutionary history of phosphatidylinositol-3-kinases: ancestral origin in eukaryotes and complex duplication patterns

Héloïse Philippon, Céline Brochier-Armanet and Guy Perrière*

Abstract

Background: Phosphatidylinositol-3-kinases (PI3Ks) are a family of eukaryotic enzymes modifying phosphoinositides in phosphatidylinositols-3-phosphate. Located upstream of the AKT/mTOR signalling pathway, PI3Ks activate secondary messengers of extracellular signals. They are involved in many critical cellular processes such as cell survival, angiogenesis and autophagy. PI3K family is divided into three classes, including 14 human homologs. While class II enzymes are composed of a single catalytic subunit, class I and III also contain regulatory subunits. Here we present an in-depth phylogenetic analysis of all PI3K proteins.

Results: We confirmed that PI3K catalytic subunits form a monophyletic group, whereas regulatory subunits form three distinct groups. The phylogeny of the catalytic subunits indicates that they underwent two major duplications during their evolutionary history: the most ancient arose in the Last Eukaryotic Common Ancestor (LECA) and led to the emergence of class III and class I/II, while the second – that led to the separation between class I and II – occurred later, in the ancestor of Unikonta (*i.e.*, the clade grouping Amoebozoa, Fungi, and Metazoa). These two major events were followed by many lineage specific duplications in particular in vertebrates, but also in various protist lineages. Major loss events were also detected in Vidiriplantae and Fungi. For the regulatory subunits, we identified homologs of class III in all eukaryotic groups indicating that, for this class, both the catalytic and the regulatory subunits were presents in LECA. In contrast, homologs of the regulatory class I have a more recent origin.

Conclusions: The phylogenetic analysis of the PI3K shed a new light on the evolutionary history of these enzymes. We found that LECA already contained a PI3K class III composed of a catalytic and a regulatory subunit. Absence of class II regulatory subunits and the recent origin of class I regulatory subunits is puzzling given that the class I/II catalytic subunit was present in LECA and has been conserved in most present-day eukaryotic lineages. We also found surprising major loss and duplication events in various eukaryotic lineages. Given the functional specificity of PI3K proteins, this suggests dynamic adaptation during the diversification of eukaryotes.

Keywords: Phosphatidylinositol-3-kinases, phylogeny, signalling pathway, LECA

Background

Phosphatidylinositol-3-kinases (PI3Ks) are enzymes that phosphorylate the 3'-position of inositol ring to generate different phosphoinositides (PIs). They are involved in many critical cellular processes such as cell survival, angiogenesis [1] or autophagy[2] and are deregulated in many human disorders (see below). PI3Ks were discovered in the 1980's as a consequence of the growing interest for their products. Following their identification and first cDNA clones, their two main inhibitors, Wortmannin and LY294002, were discovered in 1993 and 1994 respectively [3, 4]. Domain organisation of PI3Ks was already partially discovered in 1997 [5] and the first three-dimensional protein structure was resolved two years later [6] (see [7] for a detailed review on the discovery of PI3Ks). They are divided into three classes depending on their substrate

*Correspondence: guy.perriere@univ-lyon1.fr
Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard – Lyon 1, 43 bd. du 11 Novembre 1918, 69622 Villeurbanne, France

(I, II and III), and 14 coding genes have been identified in human (see Table 1 for a complete nomenclature).

Class I proteins transform phosphatidylinositols-4,5-bisphosphate (PI(4,5)P_2) into phosphatidylinositols-3,4,5-triphosphate (PI(3,4,5)P_3). The reverse reaction is done by PTEN (Phosphatase and Tensin homolog), a well known tumour suppressor protein [8, 9]. Class I is subdivided into two groups called IA and IB, depending on whether or not they can bind p85-type regulatory proteins. In human, class IA catalytic subunits (p110α, p110β and p110δ) can bind the p85α (and its two alternatives forms p55α and p50α), p85β and p55γ regulatory subunits. In contrast, p110γ, the only catalytic human protein of class IB, can bind two regulatory subunits named p87 and p101. Class I is the most studied and its members are involved in a lot of human disorders like cancers. For instance, p110α expression is deregulated in more than 30 % of various solid tumours [10], and the corresponding gene is mutated in 25 % of breast tumour samples [11–13], in 15-20 % of colorectal cancers [14–17] and in 10 % of oesophagealgeal cancers [10, 18]. Proteins p110α and p110β are generally ubiquitously expressed, and no major difference in their functions have been discovered. The major activators of class IA are RTKs (Receptor Tyrosine Kinases) [19–21] and IGF1 (Insulin-like Growth Factor 1) [21], whereas class IB is principally activated by GPCRs (G Protein-Coupled Receptors) [19, 21].

Class II proteins (PI3K-C2α, PI3K-C2β and PI3K-C2γ) are the only ones without a regulatory subunit in human, and are the most poorly characterized. Their preferential phosphoinositide substrate is not yet clearly defined and can differ between *in vivo* and *in vitro* studies [22]. In terms of biological impact, it was proved in mouse that PI3K-C2α deficiency results in embryonic lethality caused by defects in vasculogenesis [23, 24]. Another study demonstrates a role in tumour angiogenesis in the context of Lewis lung carcinoma [23]. Activators of class II are chemokines like MCP-1[25], cytokines (TNF-α and leptin) [26] and Lysophosphatidic Acid (LPA) [27]. On the contrary, Tamoxifene seems to reduce its expression in mice [23].

Finally, class III proteins synthesize phosphatidylinositols-3-phosphate (PI(3)P) from phosphatidylinositide (PI). This class is made of one catalytic and one regulatory subunits named Vacuolar Protein Sorting 34 (VPS34 or PIK3C3) and Vacuolar Protein Sorting 15 (VPS15), respectively [28]. The role of class III PI3K is to regulate membrane trafficking [28] and autophagosome formation in human [29–31].

While PI3K proteins are well studied in human, little is known about these enzymes in other organisms. Homologs of class III have been reported in unicellular eukaryotes (*e.g.*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Candida albicans*, *Dictyostelium discoideum*), vertebrates, plants, *Caenorhabditis elegans*, *Drosophila melanogaster* [32] and in microalgae [33]. The yeast genome does not code for other classes of PI3K [34]. For the classes I and II, homologs were found in vertebrates, worm, fly and Amoebozoa but not in yeast [28]. From a functional point of view, little information is available in non-human organisms. For Excavata and SAR (Stramenopiles, Alveolata and Rhizaria [35]), studies

Table 1 Nomenclature of the 14 human PI3K proteins

	Common name	Gene name	Ensembl ID	UniProt ID	Length
Catalytic subunits					
IA	p110α	PIK3CA	ENSP00000263967	P42336	1068
	p110β	PIK3CB	ENSP00000418143	P42338	1070
	p110δ	PIK3CD	ENSP00000446444	O00329	1044
IB	p110γ	PIK3CG	ENSP00000392258	P48736	1102
II	PI3K-C2α	PIK3C2A	ENSP00000265970	O00443	1086
	PI3K-C2β	PIK3C2B	ENSP00000356155	O00750	1634
	PI3K-C2γ	PIK3C2G	ENSP00000266497	O75747	1445
III	VPS34	PIK3C3	ENSP00000262039	Q8NEB9	887
Regulatory subunits					
IA	p85α	PIK3R1	ENSP00000274335	P27986	724
	p85β	PIK3R2	ENSP00000222254	O00459	728
	p55γ	PIK3R3	ENSP00000361075	Q92569	461
IB	p101	PIK3R5	ENSP00000392812	Q8WYR1	880
	p87	PIK3R6	ENSP00000475670	Q5UE93	754
III	VPS15	PIK3R4	ENSP00000349205	Q99570	B3

generally focus on the pathogen impact on the host cell phosphatidylinositols quantity more than on the function of PI3K homologs [36, 37]. Nevertheless, it has been shown that in the apicomplexan *Toxoplasma gondii*, PI3Ks are involved in the shape and size of the apicoplast [38]. In the amoebozoan species *Dictyostelium*, the PI3K classes I and II are activated by GPCRs and are involved in chemotaxis [39–41]. In *Drosophila*, focus has been on EGRF-RAS or EGFR-TOR proteins [42–44], when PTEN deregulation was largely studied in yeast [45]. Finally, a recent study presents the impact of IGFR, a PI3K activator, on the arsenite-induced apoptosis in *C. elegans* [46].

Understanding the evolutionary history of PI3Ks can provide new information about their diversity and functions. More precisely, integrating functional information available from different species and phylogenetic history allows making predictions on the ancestral as well as present day protein functions [47, 48]. Despite their biological interest, only two incomplete phylogenies of PI3Ks have been published to date. The first one was published in 2003 by Kawashima *et al.* [49] and concerned the PI3K catalytic subunits and the class IA, III but not IB regulatory subunits in Opisthokonta species. The second, published in 2011 by Brown and Auger [50], focused on the catalytic subunits in eukaryotes. Both studies identified an ancient gene duplication event that lead to the separation of class III and I/II catalytic subunits that was followed by a more recent duplication at the origin of class I and II. They found homologs of class III catalytic subunit (VPS34) in all eukaryotic groups. Furthermore, the pattern of gene duplications in catalytic class II subunits was consistent between the two studies, but not the one of class I. Therefore, the evolutionary history of PI3K portrayed by those two studies is only partial. Especially, nothing is known about the evolutionary history of class IB regulatory subunit and the existence of non-Opisthokonta homologs of class IA regulatory protein.

Taking the opportunity of the rainfall of genomic data, we performed an in-depth phylogenetic analysis of the PI3K family. First we found that catalytic and regulatory class III proteins were already present in the Last Eukaryotic Common Ancestor (LECA). We inferred that the class I and class II catalytic subunits diverged from the ancestor of Unikonta, and we deciphered the pattern of duplications within classes I and II. We showed that class IA and IB regulatory proteins are of relative recent origin and emerge in the common ancestor shared by Metazoa, Ichthyosporea and Choanoflagellida (MIC) and in the Vertebrata lineage, respectively. Finally, the investigation of the domain composition of PI3K homologs allowed testing some hypotheses resulting from our phylogenetic analysis and provided information on protein functions.

Material and methods

The 14 human PI3K protein sequences were retrieved from UniProtKB [51] (Table 1). According to homology relationships, we built four phylogenies corresponding to: i) all catalytic subunits; ii) class IA regulatory subunits; iii) class IB regulatory subunits; and iv) the class III regulatory protein. Metazoan homologs were retrieved from Ensembl [52] and other eukaryotic homologs were retrieved from a local database of complete proteomes. Similarity searches were performed on the two databases using BLASTP [53] with default settings and a cut-off set to $E \leq 10^{-30}$. Because PI3K proteins have distant homologs, the retrieved homologs were used as the seed for new BLASTP runs with the same parameters (Additional file 1). To keep only one protein sequence per genomic locus, we grouped alternative transcripts together using the E-utilities [54] and the ACNUC sequence retrieval system [55]. Then, we manually selected the most conserved and/or the better aligned peptides.

In order to decrease noise and phylogenetic redundancy, we defined two taxonomic samplings. For the catalytic dataset, a subset of 44 representative species was selected for in-depth phylogenetic analysis. Among them, we kept only ten mammals over the 42 available in Ensembl. This choice was driven by three constraints: i) having a good representative diversity for the main eukaryotic groups; ii) limiting the number of fast-evolving sequences; and iii) including model organisms such as *S. cerevisiae* and *C. elegans*. For regulatory datasets, we only made a selection among mammals and kept all the other species.

For the multiple alignments we compared the results returned by PRANK [56] and MAFFT [57] using NorMD [58]. As its scores were consistently better, we chose to use the alignments computed by MAFFT. According to author recommendations, we set the maximum number of iterations at 100 and used the `localpair` options (equivalent to the `linsi` option). Alignments were then trimmed using BMGE [59]. Several sets of parameters were tested for this program in order to get a balance between the number of sites selected and the quality of the resulting multiple alignments (Additional file 2). Number of gaps per sequence after site selection are listed in Additional files 3 to 8.

The selection of evolutionary models used for the phylogenetic inference was carried out using ProtTest [60] and the Bayesian Information Criterion (BIC) [61]. In addition to the standard amino acids substitution models implemented in ProtTest we also performed the BIC test with UL3 [62] and CAT20 [63] models. The JTT+Γ_4 model [64] was proposed for the regulatory class IA and IB proteins, the subset for Opisthokonta homologs of class II and the complete catalytic subunits dataset. The UL3+Γ_4 model [62] was suggested for the regulatory class III, and for the

reduced catalytic subunits datasets. Finally, the subset for Opisthokonta homologs of class I was inferred using the LG+Γ_4 model [65].

Maximum likelihood trees were built with PhyML [66]. Shape parameter of the Gamma distribution was estimated by PhyML with four categories for substitution rates. Branch statistical supports were estimated by the Shimodaira-Hasegawa-like test (SH) and non-parametric bootstrap (BS) with 1000 replicates.

A Bayesian approach was also used to infer the phylogeny of the eukaryotic dataset of catalytic subunits. For that purpose we used MrBayes [67]. Default parameters were used with the exception of the substitution model for amino acids, which was set to mixed. Seven million MCMC (Markov Chain Monte-Carlo) iterations were required to reach convergence. Burn-in values were set at 50 % of the iterations and we built a 50 % majority rule consensus tree after sampling one thousand trees from the posterior distribution. This sample was also used to establish the clades posterior probabilities (PP).

Domain composition analysis of all sequences was performed using HMMScan from the HMMER package [68, 69]. We searched for domains in both PfamA and PfamB databases and used all default parameters. For class IA regulatory proteins we also used Batch CD-Search [70] to confirm the presence of the p110 binding domain in non-Euteleostomi sequences.

Results

Phylogeny of PI3K

We applied a two-step strategy to decipher the evolutionary history of PI3Ks catalytic and regulatory subunits. First we investigated the taxonomic distribution of PI3Ks in all eukaryotes and constructed the corresponding phylogenies in order to identify the major evolutionary events that have affected these proteins during the diversification of eukaryotes. Then we performed a detailed analysis in Metazoa, Choanoflagellida and Ichthyosporea in order to investigate the pattern of duplications that led to the great expansion of this protein family in these lineages, including human.

Catalytic subunits

For catalytic subunits, the multi seed similarity search performed allowed us to identify 1055 PI3K homologs. After a representative species selection (see Materials and methods), we reconstructed the maximum likelihood and Bayesian phylogenies of the 139 corresponding sequences. The resulting trees were congruent and in agreement with the phylogeny inferred with the complete set of 1055 sequences (Fig. 1 and Additional files 9 and 10). These trees showed two well supported clusters corresponding to class III and to classes I and II homologs, respectively (BS of 97 and 86 %, SH of 0.94 and 0.97, both PP of 1.0). Class III homologs are found in all major eukaryotic groups: SAR, Excavata, Archaeplastida, Amoebozoa and Opisthokonta (*i.e.*, Fungi, Metazoa and unicellular relatives). We also found sequences from Haptophyta (*Emiliana huxleyi)*, Cryptophyta (*Guillardia theta*) and Apusozoa (*Thecamonas trahens*). It is worth noting that no PI3K catalytic subunit was detected in red algae, while complete proteomes of three species were present in our database. The second cluster gathered sequences of classes I and II from all eukaryotic lineages with the exception of Fungi and most Archaeplastida (Fig. 1 and Additional files 9 and 10). Specific similarity searches in Archaeplastida and Fungi in the Non-Redundant NCBI database (NR) confirmed these absences (data not shown). Regarding other eukaryotic lineages, only one copy was present in Bikonta lineages (SAR, Excavata and Haptophyta), while two copies (corresponding to class I and class II) were found in Unikonta (Amoebozoan and Opisthokonta).

These results suggested that two successive gene duplication events occurred during the diversification of eukaryotes. The first one is very ancient and took place in an ancestor of all present day eukaryotes. It led to the separation of class III and classes I-II catalytic subunits. The second duplication event led to the separation of class I and II catalytic subunit. The grouping of Bikonta homologs with Unikonta class I proteins could suggest that this duplication event occurred also before LECA, but would imply that all Bikonta lineages have independently lost the gene coding for the class II catalytic subunit. However, the grouping of Bikonta sequences with Unikonta class I sequences was not significantly supported (BS $<$ 80 %, SH $<$ 0.95 and PP $<$ 0.5). This allows another interpretation, in which the duplication event occurred in the ancestor of Unikonta and is thus more recent (Additional file 9). This scenario is more parsimonious regarding the number of losses. Depending on the scenario, LECA had three or two PI3K catalytic coding genes. In any case, three major independent loss events occurred during the diversification of eukaryotes: the class I/II in Archaeplastida and classes I and II in Fungi.

While gene duplication of PI3K catalytic subunits have been documented in animals (especially in humans), we highlighted similar situations in major eukaryotic groups as Excavata, Alveolata, Stramenopiles or Amoebozoa (Fig. 1). This indicated that the expansion of this gene family was specific to neither Metazoa nor multicellular organisms. In order to decipher in detail the evolutionary origin of PI3K catalytic subunits in Metazoa, we performed a phylogenetic analysis focused on this lineage using Choanoflagellida and Ichthyosporea as outgroups. No duplication events were found for the class III (data not shown). For the class II, we detected three paralogs (PI3K-C2α, PI3K-C2β and PI3K-C2γ) in Vertebrata, as in Human, but only one copy in other metazoan species

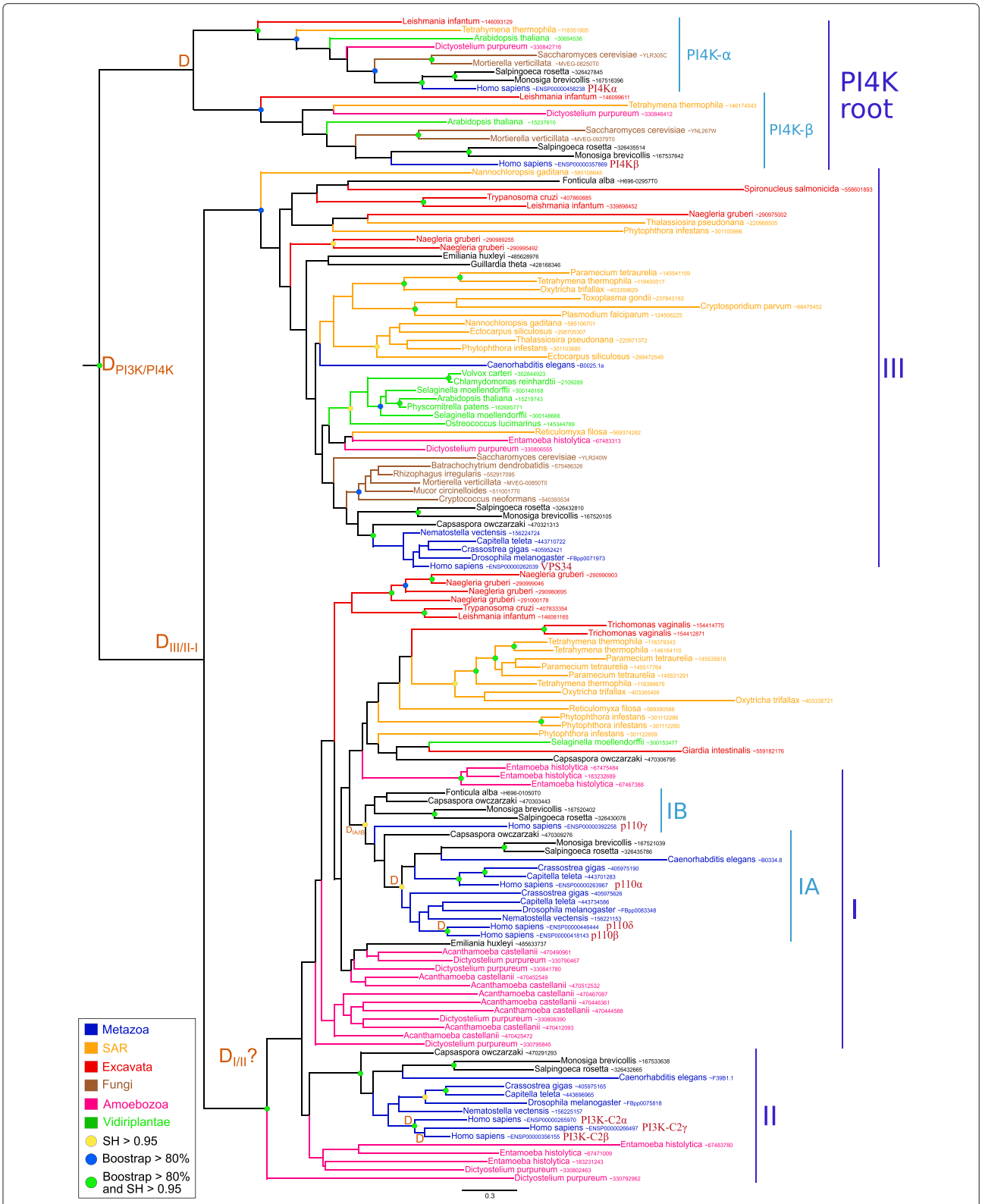

Fig. 1 Maximum likelihood phylogenetic tree of catalytic PI3Ks subunits. The tree was inferred with the UL3+Γ_4 model (398 sites, 139 sequences). Sequences are colored according to their taxonomic classification. Green circles correspond to nodes with SH > 0.95 and BS > 80 %. Blue and yellow circles correspond to nodes having either SH > 0.95 or BS > 80 %. Duplication events are indicated by an orange "D". The scale bar represents the average number of substitutions per site

like Mollusca, Cnidaria or Arthropoda (Additional file 11). This indicated that duplication events, at the origin of the three human paralogs, occurred in Vertebrata (both SH > 0.95). However, the three copies detected in *Petromyzon marinus* (*i.e.*, a Petromyzontidae) grouped with the PI3K-C2α proteins, while the two copies of *Callorhincus milii* (*i.e.*, a Chondrichthye) grouped with the PI3K-C2α and PI3K-C2β proteins, respectively. In that case the surprising location of the Petromyzontidae sequences could be due to a high evolutionary rate for PI3K-C2γ and PI3K-C2β coding genes.

An alternative hypothesis could be that the duplication event at the origin of PI3K-C2β and PI3K-C2γ occurred in Gnathostomata, suggesting a loss and two independent duplications of a PI3K-C2α coding gene in *P. marinus*. Theses two scenarios imply the loss of the PI3K-C2γ in the chondrichthyen species. Testing these hypotheses would require more data from Chondrichthyes and Petromyzontidae.

For class I, different taxonomic distributions are observed for the subclasses IA and IB. In fact, we identified homologs of classes IA and IB in most Metazoa, Choanoflagellida, Ichthyosporea and one sequence of Nucleariidae (Fig. 1 and Additional file 12). This suggests that the common ancestor of MIC possessed both classes IA and IB catalytic subunits (SH = 1 and BS > 80 %) and that the duplication event occurred before MIC. However, we could not date more precisely this duplication event due to weak statistical supports in this part of the tree, and due to the small number of proteomes available for protists related to MIC (Nucleariidae and Apusozoa). Within Metazoa, the taxonomic distribution of class IB was narrower compared to class IA. While the former was found only in the sponge Amphimedon and in Chordata, the latter was found in some protostomian lineages (Annelida, Mollusca, Platyhelminthes and Arthropoda). This indicated that secondary losses of class IB occurred during the diversification of Metazoa.

A careful examination of the phylogeny of classes IA and IB revealed also several important duplication events (Additional file 12). Within class IA, p110δ and p110β were more closely related, while p110α was more divergent. The tree suggested that the first duplication event occurred in the ancestor of Metazoa leading to the divergence of p110α, while the separation of p110β and p110δ happened in the ancestor of Vertebrata (SH = 1 and BS = 100 %). As a consequence, the absence of p110α in Arthropoda, Cnidaria and Placozoa (*Trichoplax adhaerens*) and the absence of the ancestral p110β/p110δ protein in Platyhelminthes should be interpreted as secondary losses. For class IB, the presence of two p110γ in Actinopterygii and Chondrichthyes indicates that a duplication event occurred in the ancestor of the Gnathostomata but one of the two resulting paralogs was subsequently lost in Sarcopterygii explaining why only one p110γ sequence is found in these lineages.

Regulatory subunits

For PI3K regulatory proteins, we found 117 homologs of the class III protein (VPS15), 126 homologs of the class IA and 67 homologs of the class IB protein. VPS15 homologs belonged to all the major eukaryotic groups including Fungi and Archaeplastida (Fig. 2). The taxonomic distribution and the maximum likelihood phylogeny of this protein were congruent with that of the catalytic class III subunit, indicating that both subunits were present in LECA and conserved along the diversification of present day eukaryotic lineages. The surprising grouping of sequences from *C. elegans* and *Fonticula alba* with Bikonta may be due to a long branch attraction artefact.

In contrast, regulatory subunits of class IA (p85α, p85β and p55γ) showed a more restricted taxonomic distribution, being present only in MIC (Fig. 3). While the three proteins were found in Euteleostomi, p85β and p85γ were found also in Chondrichthyes. In contrast, a single protein was found in the other metaozan lineages, Ichtyosporea and Choanoflagellida. The phylogenetic analysis of these proteins strongly supported the grouping of p85α and p85β (BS = 91 and SH = 0.98). This suggested that the three human proteins derived from two Gnathostomata specific duplications followed by loss of the p85α subunit in Chondrichthyes. However, as discussed before, because only one proteome was available for Chondrichthyes, we could not conclude with certainty about a loss in the whole taxonomic group. More surprisingly, we did not detect any p85β ortholog in Lepidosauria.

Finally, we determined that the two class IB regulatory subunits (p87 and p101) were homologous. Two paralogs were detected in Chondrichthyes, Sarcopterygii and Actinopterygii while only one sequence was present in Petromyzontidae. This strongly suggested that class IB emerged in the last common ancestor of Vertebrata and that a specific duplication underwent at the base of Gnathostomata (Fig. 4).

Domain composition evolution

Our phylogenetic analyses revealed that PI3K proteins have a complex evolutionary history involving many lineage specific duplications and, to a lesser extent, losses. To get insights on the putative function of the PI3K proteins in non-model eukaryotic species, we performed a survey of their domain composition.

Catalytic subunits

First, our results confirmed that all eukaryotic catalytic subunits shared three common domains in the same order: PI3KC2 (accession number PF00792), PI3KA

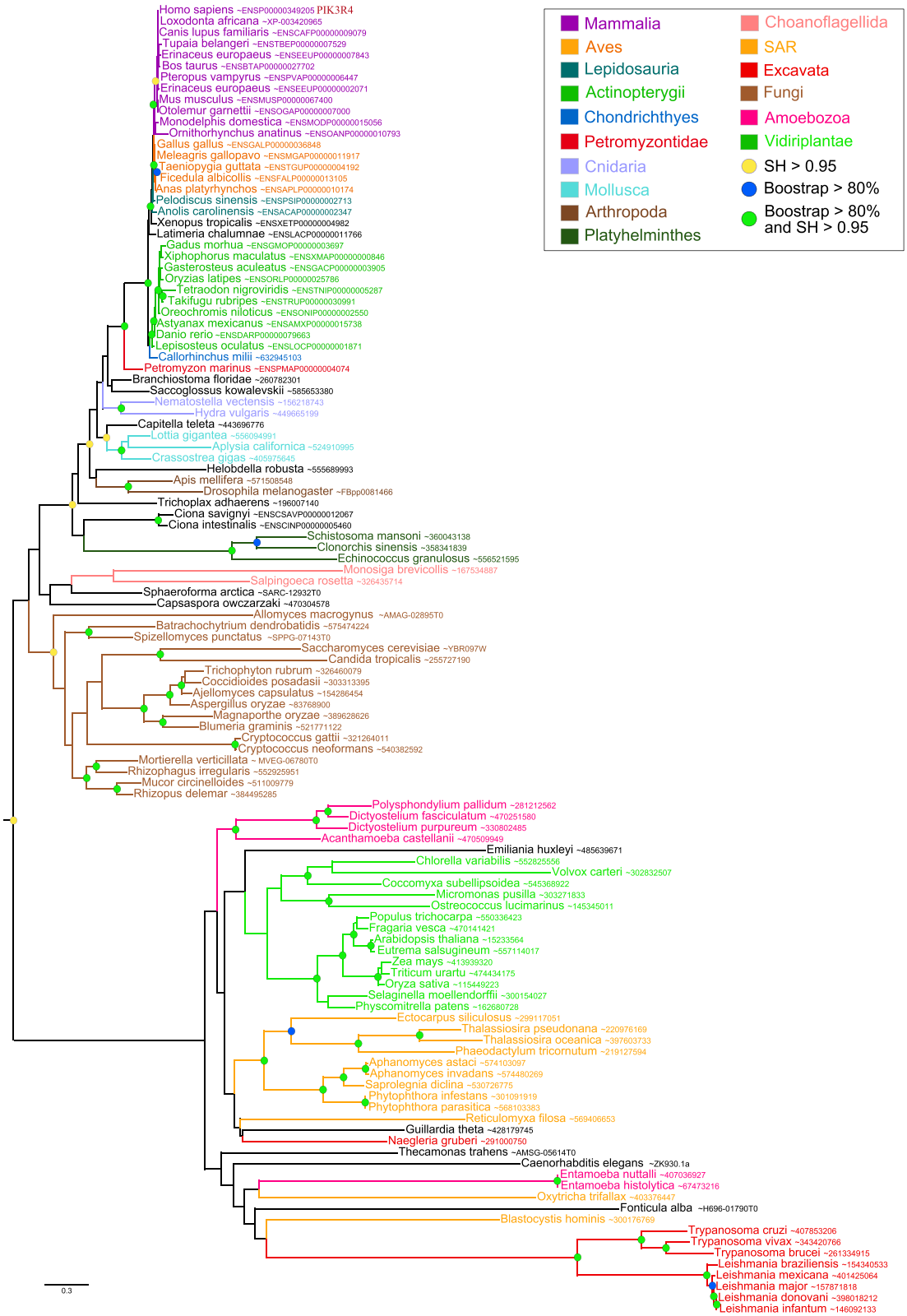

Fig. 2 Maximum likelihood phylogenetic tree of class III regulatory PI3Ks subunits. The tree was inferred with the UL3+Γ_4 model (839 sites, 117 sequences). Sequences are colored according to their taxonomic classification. Branch statistical supports and duplication events are shown using the same symbols as in Fig. 1. The scale bar represents the average number of substitutions per site

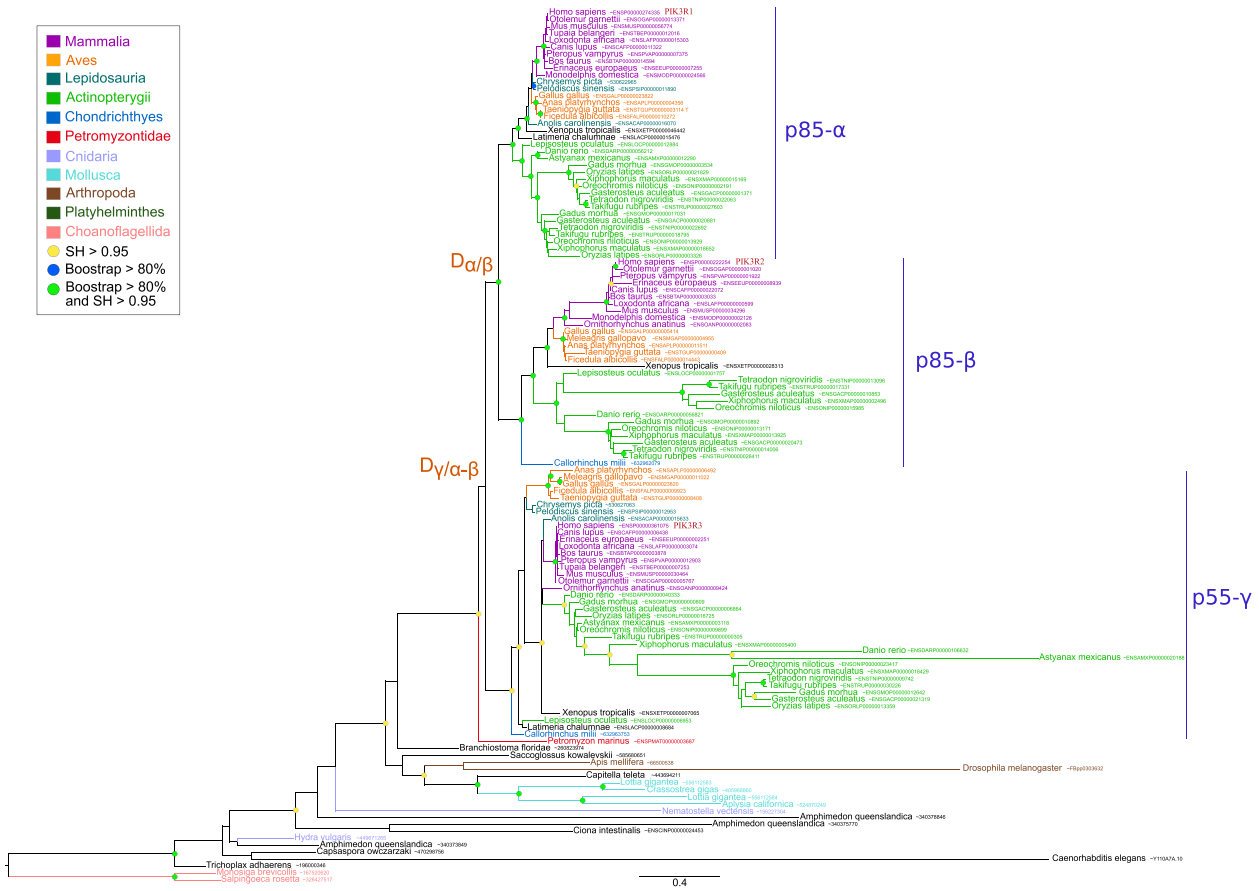

Fig. 3 Maximum likelihood phylogenetic tree of class IA regulatory PI3Ks subunits. The tree was inferred with the JTT+Γ_4 model (539 sites, 126 sequences). Branch statistical supports and duplication events are shown using the same symbols as in Fig. 1. The scale bar represents the average number of substitutions per site

(PF00613) and PI3K kinase (PF00454) (Fig. 5). Homologs of class III catalytic subunit did not harbour additional domains, while class I and II proteins possessed, in addition, the Ras Binding Domain (RBD, PF00794). This domain appeared after the class III catalytic diverged from the ancestral protein of classes I and II (*i.e.*, before LECA). It is essential for the activation of PI3K catalytic proteins by the Ras protein [71, 72]. This suggests that a functional change occurred after the duplication at the origin of classes III and I/II.

Amoebozoan class I and II proteins as well as Bikonta ancestor of class I/II protein shared exactly the same domain composition (*i.e.*, the four previously mentioned domains), precluding any conclusion regarding functional changes. In contrast the Opisthokonta subunits of classes I and II differed in their domain composition. First, we confirmed the specific presence of PX (PF00787) and C2 (PF00168) domains in class II catalytic proteins[73]. Furthermore, we detected two additional domains specific to class II PI3K-C2α and PI3K-C2β proteins: PB011861 was found at the N-terminal part of the PI3K-C2α homologs, whereas PB008942 was located in-between the RBD and PI3KC2 domains of PI3K-C2β homologs.

For class I, we confirmed the presence of the p85 binding domain (PF02192) in all IA homologs and its absence in IB homologs. Interestingly, the acquisition of the p85 binding domain by class IA proteins occurred in the last common ancestor of Metazoa, Ichthyosporea and Choanoflagellida, *i.e.*, while class IB diverged from class IA (see before). This coincided exactly with the origin of class IA regulatory proteins. This result was consistent with the fact that catalytic and regulatory subunits class IA form heterodimers through their p85 and p110 (or iSH2, PB011403) binding domains, respectively[74–78]. Among class IA catalytic proteins, p110β has a specific PfamB domain (PB000735) located in-between RBD and PI3KC2 while p110α and p110δ share exactly the same domain composition (Fig. 5).

Regulatory subunits

All class IA regulatory subunits harboured the same C-terminal domain composition, *i.e.*, a ρ-gap domain (PF00620) followed by two SH2 (PF00017) domains

Fig. 4 Maximum likelihood phylogenetic tree of class IB regulatory PI3Ks subunits. The tree was inferred with the JTT+Γ_4 model (599 sites, 67 sequences). Branch statistical supports and duplication events are shown using the same symbols as in Fig. 1. The scale bar represents the average number of substitutions per site

intercut by a PB011403 (p110-binding) domain (Fig. 6). Three exceptions were the p55γ homologs that lacked the ρ-gap domain but had a PB019399 domain, and the short Ichthyosporea and the Choanoflagellida proteins that lacked the p110-binding domain. This result is puzzling given that we detected class IA catalytic subunits in Ichthyosporea and Choanoflagellida (ancestor of p110α-β-δ). This suggested that in these species, the p110 binding domain is not required for the interaction between the regulatory and catalytic subunits. Because Ichthyosporea were represented by a single species in our databases, we could wonder if the very short protein detected is real or is artifactual because of sequencing errors. In addition to the three conserved domains (ρ-gap, SH2 and PB011403), additional domains are present in the N-terminal of some sequences (Fig. 6). For instance, the p85α and p85β proteins contained a PB000584 domain, while the copy present in Mollusca, Cnidaria and Ichtyosporea have a C1 (PF00130) domain at this location. We detected SAM_1 (PF00536) or SAM_2 (PF07647) domains in Mollusca, Arthropoda and Choanoflagellida, whereas Cnidaria and Ichtyosporea harboured an additional SH2 domain. Finally, we detected SH3 (PF00018) in Choanoflagellida and in p85α Actinopterygii homologs. Concerning class IB regulatory proteins, no domain was previously described. Our analysis detected only one PfamB domain named PI3K_1B_p101 (PF10486) in all dataset proteins (data not shown).

Finally, class III regulatory proteins showed a diverse domain composition (Fig. 7). The main information was that all proteins possessed a well-conserved Pkinase domain (PF00069) located at the N-terminal part of the proteins and two or more WD40 domains (PF00400) at

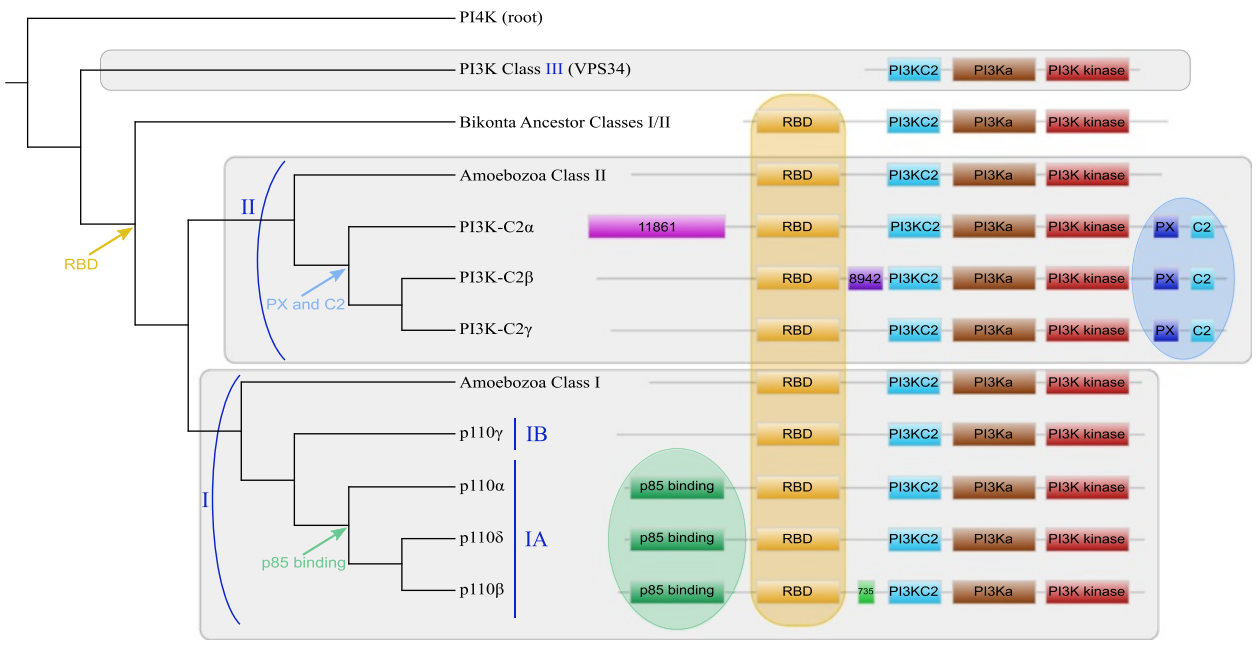

Fig. 5 Domain composition of catalytic proteins. The schematic phylogenetic tree on the left corresponds to the complete tree of Fig. 1. Light colored circles indicate domains specific to some class I and II subgroups. Box sizes are not related to the real length of the domains

their C-terminal part. Finally, a PB000285 domain, located between WD40 domains, was present in most of eukaryotic proteins except in Choanoflagellida and Excavata.

Discussion

PI3K proteins are key players of cell signalling pathways. These proteins form a very ancient protein family in eukaryotes that can be traced back to LECA. The evolutionary history of this protein family was complex and involved a lot of gene duplications and losses (Fig. 8). In addition, substantial functional changes likely occurred through gains or losses of functional domains.

Our analyses showed that two paralogous catalytic PI3K were present in LECA (class III and I/II). This indicates that the corresponding duplication is ancient and occurred during the eukaryogenesis. The regulatory subunit class III was also present in LECA meaning that, at this time, the two class III proteins were present and likely interacted together. In human and yeast, the main biological function of class III proteins is to induce autophagy

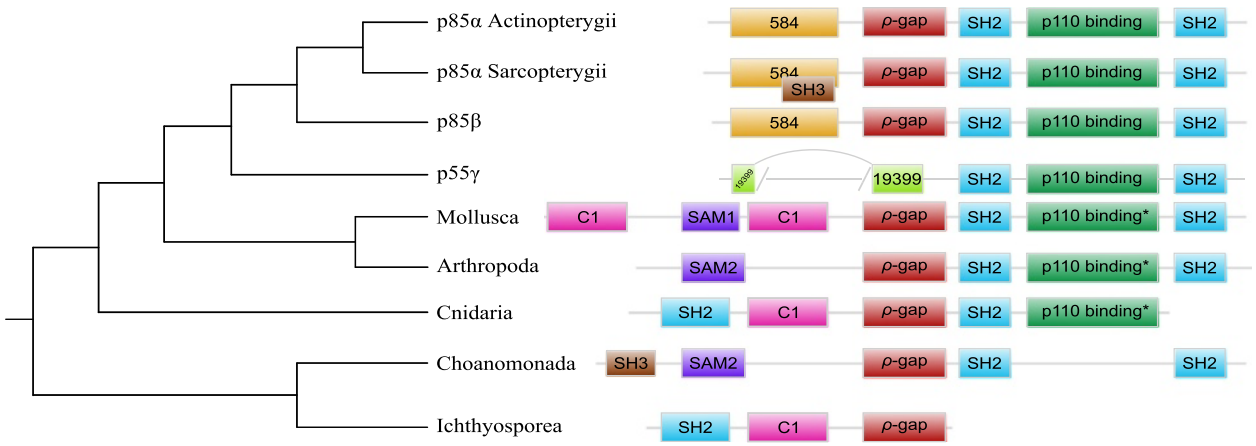

Fig. 6 Domain composition of class IA regulatory proteins. The schematic phylogenetic tree on the left corresponds to the complete tree of Fig. 2. The presence of the p110 binding domains in non Osteichthyes sequences was assessed using Batch CD-search and are indicated with asterisks (*). Box sizes are not relative to the real length of the domains

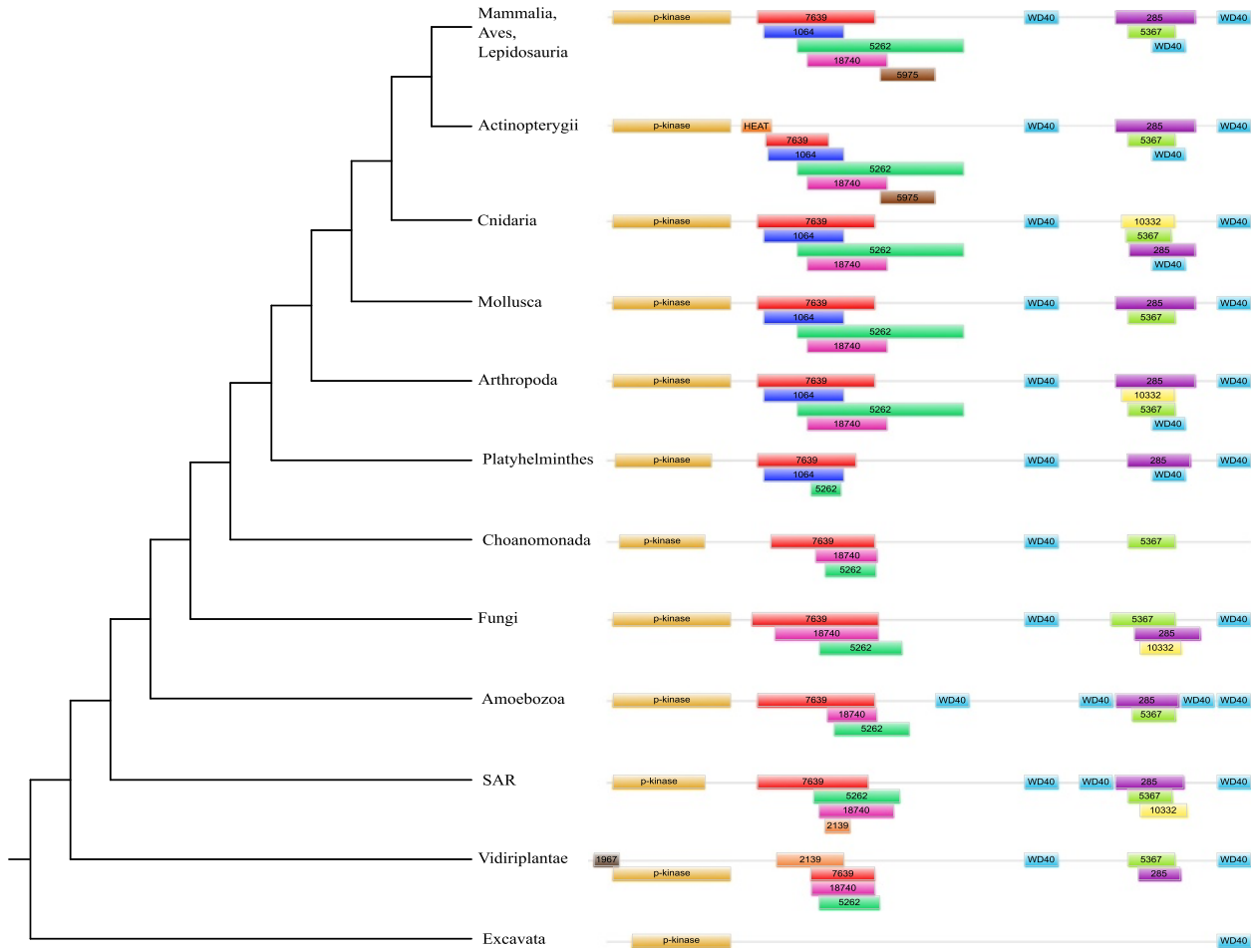

Fig. 7 Domain composition of class III regulatory proteins. The schematic phylogenetic tree on the left corresponds to the complete tree of Fig. 4. Box sizes are not related to the real length of the domains

by regulating autophagosome formation [29–31]. This suggested that these processes could have been already established in LECA, which could be further investigated by the phylogenetic study of the other proteins involved in this crucial function.

In agreement with the previous studies [49, 50], we found that two major duplication events affected the evolutionary history of catalytic subunits. As Brown and Auger, we inferred that the first duplication leading to the separation of class III and classes I/II proteins occurred before LECA. In the case of the Kawashima *et al.* study, the data set was only made of sequences that came from five Opisthokonta species (*Homo sapiens*, *D. melanogaster*, *C. elegans*, *Ciona intestinalis* and *S. pombe*), therefore, their taxonomic sample was too restricted to conclude precisely the timing of this duplication.

We detected neither catalytic nor regulatory PI3Ks proteins in red algae, suggesting three independent gene losses in this lineage. Among the three complete proteomes present in our database (*Chondrus crispus*, *Cyanidioschyzon merolae* and *Galdieria sulphuraria*) the first harbour an unusual structure [79], and the others are very small for eukaryotic genomes [80, 81]. Further analyses are needed to confirm and explain these absences.

The taxonomic distribution and the phylogeny of class I/II catalytic subunits suggested that these two classes originated in the Unikonta lineage. This implied the presence of an ancestral class I/II protein in Bikonta lineages. In our study, as in Brown and Auger, these sequences grouped with the class I homologs. But unlike them, we found that this branch of the tree was not significantly supported (BS = 7 % and SH = 0.87). All non-Opisthokonta proteins from classes I and II have the same domain composition which do not help to infer different biochemical or molecular functions for these paralogs. Moreover, the ancestral class I/II protein present in LECA might have the same biological function as the protein present in all present day Bikonta organisms.

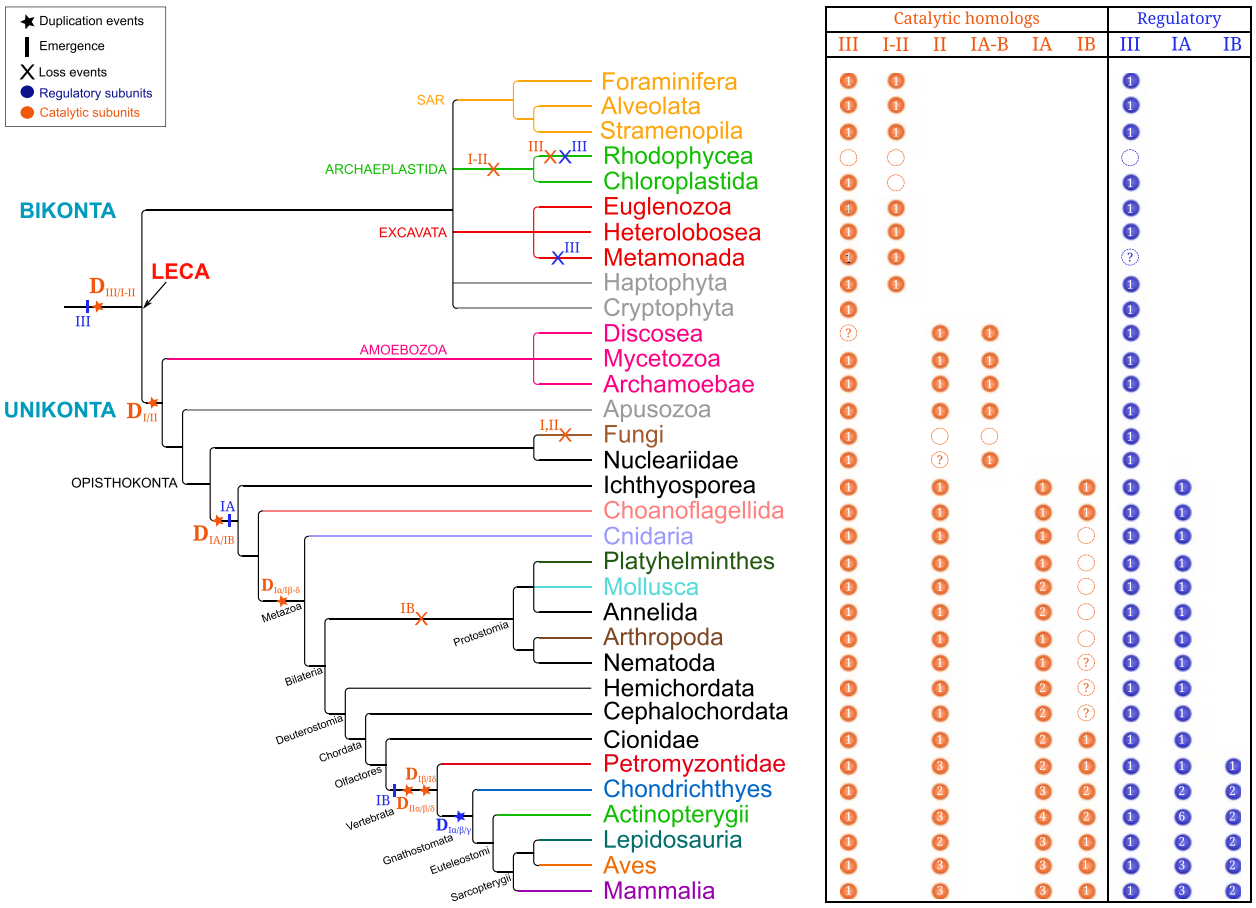

Fig. 8 Schematic representation of PI3K evolutionary history. The eukaryotic tree was built according to Adl *et al.* [35], Delsuc *et al.* [87] papers and Lecointre and Le Guyader book [88]. Information about catalytic and regulatory subunits are displayed in orange and dark blue, respectively. Taxonomic group colors are the same as in Fig. 1. Number of protein copies are indicated in the colored circles and the absence of homologs resulting from secondary losses are represented by a dotted circle. The question marks correspond to supposed lack of homologs for taxonomic groups where only one proteome is available in our databases. Duplication events are indicated by a star, losses by a cross and the emergence of regulatory subunit by a vertical line

Unfortunately, due to very few little functional studies available for Bikonta, we cannot make reasonable assumptions on the role of this protein in LECA. It would be interesting to study this protein in model organisms such as *Leishmania* and *Paramecium* species to infer its function in LECA.

In agreement with previous studies, we confirmed that *S. cerevisiae* possess only class III PI3K proteins [7, 50]. We extended this observation to all Fungi and showed that this resulted from independent losses of class I and II catalytic proteins in this lineage. Similarly, no Archaeplastida class I and II sequences were detected except in *Selaginella moellendorffii* in which a class I catalytic subunit was found. *S. moellendorffii* was not among the three Archaeplastida species of Brown and Auger dataset, so they did not find homologs in Archaeplastida in their study. Three hypotheses can explain the presence of a class I protein in *S. moellendorffii*: i) multiple and independent gene loss events occurred during the diversification of Archaeplastida except for this plant lineage; ii) an ancient gene loss event occurred in the common ancestor of Archaeplastida followed by a reacquisition by horizontal gene transfer in *S. moellendorffii*; iii) the sequence detected in this species was a contamination. In order to test the third hypothesis we performed a BLAST search to identify the homologs of the six protein genes surrounding the *S. moellendorffii* gene (data not shown). Best significant hits were all obtained with sequences from Viridiplantae, which invalidates this hypothesis of a contamination.

The evolutionary history of classes I and II was punctuated by gene duplications and losses. The two previous published phylogenies disagreed regarding the pattern of duplications. Brown and Auger found that p110α and p110γ (class IA and IB, respectively) grouped together, while Kawashima *et al.* found the three class IA proteins

(p110α, p110β and p110δ) in the same cluster. Our analyses agreed with the result of Kawashima *et al.*, but provided a more precise picture because we used 117 Opisthokonta complete proteomes, while they analyzed less than 30 Opisthokonta species. In fact, we found that a first duplication occurred before the last common ancestor of Metazoa, Ichthyosporea and Choanoflagellida and led to the separation of class IA and IB. Due to the low number of proteomes available for Nucleariidae and Apusozoa (only one of each in our database) and weak statistical branch supports, we could not date more precisely this gene duplication event. This question should be further addressed to conclude if the common ancestor of Opisthokonta possessed one or two copies of class I catalytic subunits. An Opisthokonta specific duplication would imply two independent gene losses in Fungi and one in Nucleariidae and Apusozoa, whereas a MIC specific duplication would only imply a gene loss in the fungal lineage and a misplacement of the Nucleariidae sequence. Then, two successive duplications occurred in class IA. The first one took place in the ancestor of Metazoa while the duplication leading to p110β and p110δ occurred in Vertebrata. The branch support of the subclasses duplication was not significant in the eukaryotic catalytic tree (SH = 0.58, BS = 38 % and PP = 0.68), but both BS and SH values supported this node in the catalytic tree built with Ichthyosporea, Choanoflagellida and Metazoa homologs (Additional file 12). Moreover, the domain composition of those proteins – and especially the apparition of the p85-binding domain in class IA – supports the conjecture of a first duplication leading to the separation of the two subclasses before secondary duplications in subclass IA.

For class IA regulatory proteins, our results differed from the only comparable phylogeny available [49]. In fact, we found p85α grouped with p85β whereas Kawashima *et al.* found p85α next to p55γ. Nevertheless, corresponding branches were supported neither in their study nor in our phylogeny. More precisely, the first duplication event was well supported in Kawashima *et al.* (BS = 100 %) but not in our study (SH = 0.43 and BS = 26 %). On the contrary, the second duplication event was supported by both values in our trees (SH = 0.99 and BS = 91 %), while the BS value was only equal to 82 % in the Kawashima *et al.* study. Interestingly, we detected class IA regulatory homologous proteins in Metazoa, Choanoflagellida and Ichtyosporea that exactly corresponds to the emergence of class IA catalytic subunits and the appearance of the p85-binding domain. In contrast, regulatory protein duplications occurred before the duplication of catalytic subunits (in Gnathostomata and Metazoa, respectively). So, in non-Gnathostomata organisms (*i.e.*, Mollusca, Annelida), there are two class IA catalytic subunits for only one class IA regulatory proteins. So we can hypothesize that the regulatory subunit of these organisms can regulate both p110α and the ancestor of p110β/p110δ proteins or that the regulation is done by another protein not yet characterised.

For the catalytic class II, the two previous phylogenies found PI3K-C2α grouped with the PI3K-C2β while, in our trees, PI3K-C2β is grouped with the PI3K-C2γ. We found that these three proteins resulted from two successive duplications that occurred in the Vertebrata or Gnathostomata lineage. The discrepancy can be the consequence of a restricted taxonomic sampling and of less efficient methods (*i.e.*, neighbour-joining *vs.* maximum likelihood and Bayesian approaches). In terms of domain composition, proteins of classes I and II shared four specific domains. We confirmed the presence of both PX and C2 terminal domains [73, 82, 83] in all Opisthokonta class II proteins. We discovered that PI3K-C2α and PI3K-C2β shared a specific domain located in the first half of the sequence. This new information about these poorly understood catalytic subunits suggests that they had specific molecular or biochemical functions.

Furthermore, we provided a detailed phylogenetic analysis of class III protein (VPS15). Where Kawashima *et al.* used only two Fungi, one *Drosophila* and one *Ciona* species, we detected 117 homologous sequences belonging to all eukaryotic groups. Note that this ubiquity among eukaryotes was previously partially shown in [32]. Interestingly, no duplication event in any organism occurred during eukaryotic evolution for this class. Our results suggest that both catalytic and regulatory class III subunits were already present in LECA and conserved in one copy in Opisthokonta and other present-day eukaryotes (excepted *Naegleria gruberi* and some SAR which possessed two or more catalytic class III subunits). This contrasted with classes I and II PI3Ks.

We provided the first phylogenetic analysis of class IB regulatory proteins. We found that p87 and p101 proteins appeared very recently (in Vertebrata) and result from a specific Gnathostomata duplication. But the catalytic class IB protein emerged in the last common ancestor of Opisthokonta. This raises the question of the regulation of IB catalytic protein in other animals, Choanoflagellida and Ichthyosporea organisms.

Finally, in terms of biological functions, a lot of studies demonstrated the implication of class I proteins in chemotaxis in *Dyctiostelium* [39–41]. Interestingly, in human, class IB is involved in the chemotaxis of different cell types like macrophages [84] or smooth muscle and $CD4^+$ T cells [85]. On the contrary, human class IA proteins are implicated in mitosis [86] and cell growth/proliferation through the AKT/mTOR signalling pathway regulation [7]. Accordingly, it is tempting to hypothesize that the ancestral function of class I was chemotaxy. Given that the duplication leading to classes IA and IB occurred in

the Opisthokonta lineage, we can wonder if there is a link between the duplication and the emergence of multicellularity in this taxon.

Conclusion

PI3Ks form a complex and very ancient protein family. This study allowed us to establish a much more accurate landscape of its evolutionary history thanks to the use of a broad set of completely sequenced eukaryotes. On the other hand, some parts of the trees we built for the different PI3K subunits are still poorly resolved. Especially we were unable to date precisely some duplication events (*e.g.*, duplication of the the three catalytic subunits of class II). This is mainly due to the lack of data for organisms such as Exacavates, SAR, Petromyzontidae and Chondrichthyes. Using the grounds provided by the approaches developed for this research, it will be possible to perform a broader study on the different proteins involved in the whole AKT/mTOR signaling pathway.

Availability of supporting data

The different data sets supporting the results of this article (multiple sequence alignments) are available at http://pbil.univ-lyon1.fr/datasets/Philippon2015/.

Additional file

Additional file 1: Organisms used as seeds for the second BLAST search. For detecting distant homologs we used sequences from 25 organisms as seeds for a second BLAST search. We choose organisms from different taxonomic groups in order to reach all eukaryotic homologs.

Additional file 2: Datasets characteristics and program parameters used. For each dataset, some information (number of human paralogs, number of homologs found, number of selected sites, etc.), as well as the parameters used for BMGE program and the substitution model selected are given. Also, the names of the supporting data files containing the corresponding trimmed multiple alignments are given.

Additional file 3: Number of gaps per sequence after site selection for the reduced catalytic dataset. Sequences are sorted by increased percentage of gaps.

Additional file 4: Number of gaps per sequence after site selection for the regulatory subunit class III dataset. Sequences are sorted by increased percentage of gaps.

Additional file 5: Number of gaps per sequence after site selection for the regulatory subunit class IA dataset. Sequences are sorted by increased percentage of gaps.

Additional file 6: Number of gaps per sequence after site selection for the the regulatory subunit class IB dataset. Sequences are sorted by increased percentage of gaps.

Additional file 7: Number of gaps per sequence after site selection for the MIC class II catalytic subunit dataset. Sequences are sorted by increased percentage of gaps.

Additional file 8: Number of gaps per sequence after site selection for the MIC class I catalytic subunit dataset. Sequences are sorted by increased percentage of gaps.

Additional file 9: Complete phylogenetic tree of catalytic subunits. The tree was inferred with the JTT+Γ_4 model (468 sites, 1055 sequences). Sequences are colored according to their taxonomic classification. SH support is indicated over the branches. Duplication events are shown by an orange"D". The scale bar represents the average number of substitutions per site.

Additional file 10: Bayesian phylogenetic tree of selected catalytic subunits. The tree was inferred using the MrBayes program and the same alignment as the one used to build the corresponding maximum likelihood tree (Fig. 1). Sequences are colored according to their taxonomic classification. Yellow and red circles correspond to PP > 0.90 and PP > 0.95, respectively. Duplication events are indicated by an orange "D". The scale bar represents the average number of substitutions per site.

Additional file 11: Phylogenetic tree of Metazoa, Ichthyosporea and Choanoflagellida homologs of class II catalytic proteins. The tree was inferred with the JTT+Γ_4 model (1113 sites, 108 sequences). Sequences are colored according to their taxonomic classification. Branch statistical supports and duplication events are shown using the same symbols as in Fig. 1. As described in the Material and methods section, we selected all non-mammal species from Ensembl, and kept ten representative mammal organisms and all Ichtyosporea, Choanoflagellida and Metazoa species from our local database.

Additional file 12: Phylogenetic tree of Metazoa, Ichtyosporea and Choanoflagellida homologs of PI3K class I catalytic proteins. The tree was inferred with the LG+Γ_4 model (828 sites, 185 sequences). Sequences are colored according to their taxonomic classification. Branch statistical supports and duplication events are shown using the same symbols as in Fig. 1.

Abbreviations

BIC: Bayesian Information Criterion; BS: Bootstrap support; CAT: Categories model; GPCR: G Protein-Coupled Receptors; IGF1: Insulin-like Growth Factor 1; JTT: Jones, Taylor and Thornton model; LECA: Last Eukaryotic Common Ancestor; LG: Le and Gascuel model; LPA: Lysophosphatidic Acid; MIC: Metazoa, Ichthyosporea and Choanoflagellida; NR: Non-Redundant (database); PI: phosphoinositides; PI3K: Phosphatidylinositol-3-kinases; PP: posterior probability; PTEN: Phosphatase and Tensin homolog; RBD: Ras Binding Domain; RTK: Receptor Tyrosine Kinase; SAR: Stramenopiles, Alveolata and Rhizaria; SH: Shimodaira-Hasegawa like support; UL3: Three-matrix unsupervised model; VPS15: Vacuolar Protein Sorting 15 (also named PIK3R4); VPS34: Vacuolar Protein Sorting 34 (also named PIK3C3).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CBA and GP conceived the research project; GP, CBA and HP defined analyses. HP did all computations and developments; HP, CBA and GP wrote the manuscript. All authors read and approve the final manuscript.

Acknowledgements

We would like to thank the Région Rhône-Alpes, which was funding this project and the Ph.D grant of HP. All computations have been performed using the LBBE/PRABI cluster. We also thank Murray Patterson for careful re-reading of the manuscript.

Received: 20 May 2015 Accepted: 28 September 2015
Published online: 19 October 2015

References

1. Graupera M, Potente M. Regulation of angiogenesis by PI3K signaling networks. Exp. Cell Res. 2013;319(9):1348–55.
2. Burman C, Ktistakis NT. Regulation of autophagy by phosphatidylinositol 3-phosphate. FEBS Lett. 2010;584(7):1302–12.
3. Arcaro A, Wymann MP. Wortmannin is a potent phosphatidylinositol 3-kinase inhibitor: the role of phosphatidylinositol 3,4,5-trisphosphate in neutrophil responses. Biochem. J. 1993;296 (Pt 2):297–301.
4. Vlahos CJ, Matter WF, Hui KY, Brown RF. A specific inhibitor of phosphatidylinositol 3-kinase, 2-(4-morpholinyl)-8-phenyl-4H-1-benzopyran-4-one (LY294002). J. Biol. Chem. 1994;269(7):5241–8.

5. Vanhaesebroeck B, Leevers SJ, Panayotou G, Waterfield MD. Phosphoinositide 3-kinases: a conserved family of signal transducers. Trends Biochem. Sci. 1997;22(7):267–72.
6. Walker EH, Perisic O, Ried C, Stephens L, Williams RL. Structural insights into phosphoinositide 3-kinase catalysis and signalling. Nature. 1999;402(6759):313–20.
7. Vanhaesebroeck B, Stephens L, Hawkins P. PI3K signalling: the path to discovery and understanding. Nat. Rev. Mol. Cell Biol. 2012;13(3):195–203.
8. Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, et al. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science. 1997;275(5308):1943–7.
9. Hopkins BD, Parsons RE. Molecular pathways: intercellular PTEN and the potential of PTEN restoration therapy. Clin. Cancer Res. 2014;20(21): 5379–83.
10. Ming Z, Jiang D, Hu Q, Li X, Huang J, Xu Y, et al. Diagnostic application of PIK3CA mutation analysis in Chinese esophageal cancer patients. Diagn. Pathol. 2014;9:153.
11. Saal LH, Holm K, Maurer M, Memeo L, Su T, Wang X, et al. PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. Cancer Res. 2005;65(7):2554–9.
12. Campbell IG, Russell SE, Choong DY, Montgomery KG, Ciavarella ML, Hooi CS, et al. Mutation of the PIK3CA gene in ovarian and breast cancer. Cancer Res. 2004;64(21):7678–81.
13. Li SY, Rong M, Grieu F, Iacopetta B. PIK3CA mutations in breast cancer are associated with poor outcome. Breast Cancer Res. Treat. 2006;96(1):91–5.
14. Ogino S, Lochhead P, Giovannucci E, Meyerhardt JA, Fuchs CS, Chan AT. Discovery of colorectal cancer PIK3CA mutation as potential predictive biomarker: power and promise of molecular pathological epidemiology. Oncogene. 2014;33(23):2949–55.
15. Liao X, Morikawa T, Lochhead P, Imamura Y, Kuchiba A, Yamauchi M, et al. Prognostic role of PIK3CA mutation in colorectal cancer: cohort study and literature review. Clin. Cancer Res. 2012;18(8):2257–68.
16. De Roock W, Claes B, Bernasconi D, De Schutter J, Biesmans B, Fountzilas G, et al. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. Lancet Oncol. 2010;11(8):753–62.
17. Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S, et al. High frequency of mutations of the PIK3CA gene in human cancers. Science. 2004;304(5670):554.
18. Shigaki H, Baba Y, Watanabe M, Murata A, Ishimoto T, Iwatsuki M, et al. PIK3CA mutation is associated with a favorable prognosis among patients with curatively resected esophageal squamous cell carcinoma. Clin. Cancer Res. 2013;19(9):2451–9.
19. Hirsch E, Braccini L, Ciraolo E, Morello F, Perino A. Twice upon a time: PI3K's secret double life exposed. Trends Biochem. Sci. 2009;34(5):244–8.
20. Burris H. Overcoming acquired resistance to anticancer therapy: focus on the PI3K/AKT/mTOR pathway. Cancer Chemother. Pharmacol. 2013;71(4): 829–42.
21. Engelman JA, Luo J, Cantley LC. The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. Nat. Rev. Genet. 2006;7(8):606–19.
22. Maffucci T, Falasca M. New insight into the intracellular roles of class II phosphoinositide 3-kinases. Biochem. Soc. Trans. 2014;42(5):1378–82.
23. Yoshioka K, Yoshida K, Cui H, Wakayama T, Takuwa N, Okamoto Y, et al. Endothelial PI3K-C2α, a class II PI3K, has an essential role in angiogenesis and vascular barrier function. Nat. Med. 2012;18(10):1560–9.
24. Franco I, Gulluni F, Campa CC, Costa C, Margaria JP, Ciraolo E, et al. PI3K class II α controls spatially restricted endosomal PtdIns3P and Rab11 activation to promote primary cilium function. Dev. Cell. 2014;28(6): 647–58.
25. Turner SJ, Domin J, Waterfield MD, Ward SG, Westwick J. The CC chemokine monocyte chemotactic peptide-1 activates both the class I p85/p110 phosphatidylinositol 3-kinase and the class II PI3K-C2α. J. Biol. Chem. 1998;273(40):25987–95.
26. Ktori C, Shepherd PR, O'Rourke L. TNF-α and leptin activate the α-isoform of class II phosphoinositide 3-kinase. Biochem. Biophys. Res. Comm. 2003;306(1):139–43.
27. Maffucci T, Cooke FT, Foster FM, Traer CJ, Fry MJ, Falasca M. Class II phosphoinositide 3-kinase defines a novel signaling pathway in cell migration. J. Cell Biol. 2005;169(5):789–99.
28. Jean S, Kiger AA. Classes of phosphoinositide 3-kinases at a glance. J. Cell Sci. 2014;127(Pt 5):923–28.
29. Ravikumar B, Sarkar S, Davies JE, Futter M, Garcia-Arencibia M, Green-Thompson ZW, et al. Regulation of mammalian autophagy in physiology and pathophysiology. Physiol. Rev. 2010;90(4):1383–435.
30. Kongara S, Karantza V. The interplay between autophagy and ROS in tumorigenesis. Front. Oncol. 2012;2:171.
31. Wirth M, Joachim J, Tooze SA. Autophagosome formation–the role of ULK1 and Beclin1-PI3KC3 complexes in setting the stage. Semin. Cancer Biol. 2013;23(5):301–9.
32. Backer JM. The regulation and function of class III PI3Ks: novel roles for Vps34. Biochem. J. 2008;410(1):1–17.
33. Jiang Q, Zhao L, Dai J, Wu Q. Analysis of autophagy genes in microalgae: Chlorella as a potential model to study mechanism of autophagy. PLoS One. 2012;7(7):41826.
34. Schu PV, Takegawa K, Fry MJ, Stack JH, Waterfield MD, Emr SD. Phosphatidylinositol 3-kinase encoded by yeast VPS34 gene essential for protein sorting. Science. 1993;260(5104):88–91.
35. Adl SM, Simpson AG, Lane CE, Lukes J, Bass D, Bowser SS, et al. The revised classification of eukaryotes. J. Eukaryot. Microbiol. 2012;59(5): 429–93.
36. Wilkowsky SE, Barbieri MA, Stahl P, Isola EL. *Trypanosoma cruzi*: phosphatidylinositol 3-kinase and protein kinase B activation is associated with parasite invasion. Exp. Cell Res. 2001;264(2):211–8.
37. Quan JH, Cha GH, Zhou W, Chu JQ, Nishikawa Y, Lee YH. Involvement of PI 3 kinase/Akt-dependent bad phosphorylation in *Toxoplasma gondii*-mediated inhibition of host cell apoptosis. Exp. Parasitol. 2013;133(4):462–71.
38. Daher W, Morlon-Guyot J, Sheiner L, Lentini G, Berry L, Tawk L, et al. Lipid kinases are essential for apicoplast homeostasis in *Toxoplasma gondii*. Cell. Microbiol. 2014;17(4):559–78.
39. Merlot S, Firtel RA. Leading the way: Directional sensing through phosphatidylinositol 3-kinase and other signaling pathways. J. Cell Sci. 2003;116(Pt 17):3471–8.
40. Iglesias PA. Spatial regulation of PI3K signaling during chemotaxis. Wiley Interdiscip. Rev. Syst. Biol. 2009;1(2):247–53.
41. Afonso PV, Parent CA. PI3K and chemotaxis: a priming issue? Sci. Signal. 2011;4(170):22.
42. Read RD, Cavenee WK, Furnari FB, Thomas JB. A drosophila model for EGFR-Ras and PI3K-dependent human glioma. PLoS Genet. 2009;5(2): 1000374.
43. McNeill H, Craig GM, Bateman JM. Regulation of neurogenesis and epidermal growth factor receptor signaling by the insulin receptor/target of rapamycin pathway in Drosophila. Genetics. 2008;179(2):843–53.
44. Read RD. *Drosophila melanogaster* as a model system for human brain cancers. Glia. 2011;59(9):1364–76.
45. Cid VJ, Rodríguez-Escudero I, Andrés-Pons A, Romá-Mateo C, Gil A, den Hertog J, et al. Assessment of PTEN tumor suppressor activity in nonmammalian models: the year of the yeast. Oncogene. 2008;27(41): 5431–42.
46. Wang S, Teng X, Wang Y, Yu H-Q, Luo X, Xu A, et al. Molecular control of arsenite-induced apoptosis in *Caenorhabditis elegans*: roles of insulin-like growth factor-1 signaling pathway. Chemosphere. 2014;112:248–55.
47. Eme L, Moreira D, Talla E, Brochier-Armanet C. A complex cell division machinery was present in the last common ancestor of eukaryotes. PLoS One. 2009;4(4):5021.
48. Eme L, Trilles A, Moreira D, Brochier-Armanet C. The phylogenomic analysis of the anaphase promoting complex and its targets points to complex and modern-like control of the cell cycle in the last common ancestor of eukaryotes. BMC Evol. Biol. 2011;11:265.
49. Kawashima T, Tokuoka M, Awazu S, Satoh N, Satou Y. A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. VIII. Genes for PI3K signaling and cell cycle. Dev. Genes Evol. 2003;213(5-6):284–90.
50. Brown JR, Auger KR. Phylogenomics of phosphoinositide lipid kinases: perspectives on the evolution of second messenger signaling and drug discovery. BMC Evol. Biol. 2011;11:4.
51. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):204–12.
52. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):662–9.

53. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
54. Sayers E. A general introduction to the E-utilities. In: Entrez Programming Utilities Help [Internet]. Bethesda: National Center for Biotechnology Information; 2010.
55. Gouy M, Delmotte S. Remote access to ACNUC nucleotide and protein sequence databases at PBIL. Biochimie. 2008;90(4):555–62.
56. Loytynoja A. Phylogeny-aware alignment with PRANK. Methods Mol. Biol. 2014;1079:155–70.
57. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 2013;30(4):772–80.
58. Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O. Towards a reliable objective function for multiple sequence alignments. J. Mol. Biol. 2001;314(4):937–51.
59. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 2010;10:210.
60. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27(8):1164–5.
61. Schwartz G. Estimating the dimension of a model. Ann. Stat. 1978;6(2): 461–4.
62. Le SQ, Lartillot N, Gascuel O. Phylogenetic mixture models for proteins. Phil. Trans. R. Soc. B. 2008;363:3965–76.
63. Le SQ, Gascuel O, Lartillot N. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 2008;24(20):2317–23.
64. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput. Applic. Biosci. 1992;8(3): 275–82.
65. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol. Biol. Evol. 2008;25(7):1307–20.
66. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 2010;59(3):307–21.
67. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 2012;61(3):539–42.
68. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):222–30.
69. Eddy SR. PLoS Comput. Biol. 2011;7(10):1002195.
70. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 2011;39(Database issue):225–9.
71. Cully M, You H, Levine AJ, Mak TW. Beyond PTEN mutations: the PI3K pathway as an integrator of multiple inputs during tumorigenesis. Nat. Rev. Cancer. 2006;6(3):184–92.
72. Fritsch R, de Krijger I, Fritsch K, George R, Reason B, Kumar MS, et al. RAS and RHO families of GTPases directly regulate distinct phosphoinositide 3-kinase isoforms. Cell. 2013;153(5):1050–63.
73. F O, Rusten TE, Stenmark H. Phosphoinositide 3-kinases as accelerators and brakes of autophagy. FEBS J. 2013;280(24):6322–37.
74. Dhand R, Hara K, Hiles I, Bax B, Gout I, Panayotou G, et al. PI 3-kinase: structural and functional analysis of intersubunit interactions. EMBO J. 1994;13(3):511–21.
75. Klippel A, Escobedo JA, Hu Q, Williams LT. A region of the 85-kilodalton (kDa) subunit of phosphatidylinositol 3-kinase binds the 110-kDa catalytic subunit in vivo. Mol. Cell. Biol. 1993;13(9):5560–6.
76. Yu J, Wjasow C, Backer JM. Regulation of the p85/p110α phosphatidylinositol 3'-kinase. Distinct roles for the N-terminal and C-terminal SH2 domains. J. Biol. Chem. 1998;273(46):30199–203.
77. Holt KH, Olson L, Moye-Rowley WS, Pessin JE. Phosphatidylinositol 3-kinase activation is mediated by high-affinity interactions between distinct domains within the p110 and p85 subunits. Mol. Cell. Biol. 1994;14(1):42–49.
78. Geering B, Cutillas PR, Nock G, Gharbi SI, Vanhaesebroeck B. Class IA phosphoinositide 3-kinases are obligate p85-p110 heterodimers. Proc. Natl. Acad. Sci. USA. 2007;104(19):7809–14.
79. Collén J, Porcel B, Carré W, Ball SG, Chaparro C, Tonon T, et al. Proc. Natl. Acad. Sci. USA. 2013;110(13):5247–52.
80. Nozaki H, Takano H, Misumi O, Terasawa K, Matsuzaki M, Maruyama S, et al. A 100 %-complete sequence reveals unusually simple genomic features in the hot-spring red alga *Cyanidioschyzon merolae*. BMC Biol. 2007;5:28.
81. Jain K, Krause K, Grewe F, Nelson GF, Weber AP, Christensen AC, et al. Extreme features of the *Galdieria sulphuraria* organellar genomes: a consequence of polyextremophily? Genome Biol. Evol. 2015;7(1):367–80.
82. Falasca M, Maffucci T. Role of class II phosphoinositide 3-kinase in cell signalling. Biochem. Soc. Trans. 2007;35(Pt 2):211–4.
83. Djordjevic S, Driscoll PC. Structural insight into substrate specificity and regulatory mechanisms of phosphoinositide 3-kinases. Trends Biochem. Sci. 2002;27(8):426–32.
84. Hirsch E, Katanaev VL, Garlanda C, Azzolino O, Pirola L, Silengo L, et al. Central role for G protein-coupled phosphoinositide 3-kinase γ in inflammation. Science. 2000;287(5455):1049–53.
85. Smirnova NF, Gayral S, Pedros C, Loirand G, Vaillant N, Malet N, et al. Targeting PI3Kγ activity decreases vascular trauma-induced intimal hyperplasia through modulation of the Th1 response. J. Exp. Med. 2014;211(9):1779–92.
86. Silió V, Redondo-Muñoz J, Carrera AC. Phosphoinositide 3-kinase β regulates chromosome segregation in mitosis. Mol. Biol. Cell. 2012;23(23): 4526–42.
87. Delsuc F, Brinkmann H, Chourrout D, Philippe H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature. 2006;439(7079):965–8.
88. Lecointre G, Le Guyader H. The Tree of Life: A Phylogenetic Classification. Harvard: Harvard University Press; 2006.

B.1 Fichiers relatifs à l'article des PI3K

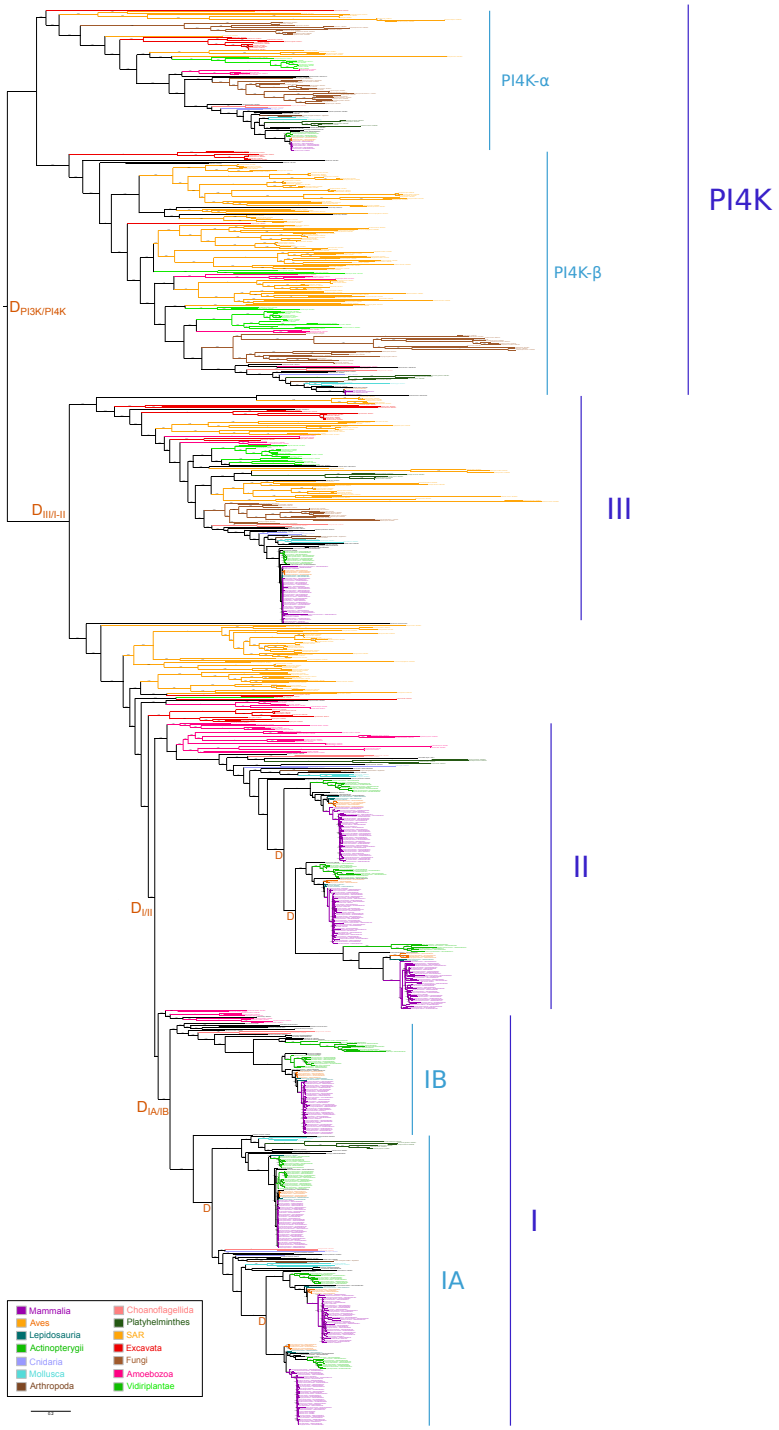

FIGURE *B.1* – ***Arbre phylogénétique des sous-unités catalytiques, raciné par les PI4K.*** L'arbre a été inféré par maximum de vraisemblance en utilisant le modèle évolutif JTT+Γ_4 (468 sites conservés, 1055 séquences). Le support SH est indiqué au dessus des branches. Les événements de duplication sont matérialisés par un D orange. La barre d'échelle représente le nombre moyen de substitutions par site.

Organism name and sequence ID	Number of gaps (percentage)
Homo sapiens ENSP00000418143	0 (0.0)
Homo sapiens ENSP00000392258	0 (0.0)
Phytophthora infestans 301122859	0 (0.0)
Trichomonas vaginalis 154414775	0 (0.0)
Dictyostelium purpureum 330806390	0 (0.0)
Dictyostelium purpureum 330795845	0 (0.0)
Acanthamoeba castellanii 470452549	0 (0.0)
Capsaspora owczarzaki 470309276	0 (0.0)
Drosophila melanogaster FBpp0083348	0 (0.0)
Paramecium tetraurelia 145531291	0 (0.0)
Tetrahymena thermophila 118379343	0 (0.0)
Homo sapiens ENSP00000263967	0 (0.0)
Entamoeba histolytica 67467388	0 (0.0)
Entamoeba histolytica 67471009	0 (0.0)
Oxytricha trifallax 403365406	0 (0.0)
Capsaspora owczarzaki 470303443	0 (0.0)
Nematostella vectensis 156221153	0 (0.0)
Entamoeba histolytica 67475484	0 (0.0)
Dictyostelium purpureum 330790467	0 (0.0)
Acanthamoeba castellanii 470490961	0 (0.0)
Tetrahymena thermophila 146164110	0 (0.0)
Salpingoeca rosetta 326430078	0 (0.0)
Paramecium tetraurelia 145517764	0 (0.0)
Dictyostelium purpureum 330841780	0 (0.0)
Paramecium tetraurelia 145535818	0 (0.0)
Phytophthora infestans 301112280	0 (0.0)
Phytophthora infestans 301112286	0 (0.0)
Capitella teleta 443734586	0 (0.0)
Capitella teleta 443701283	0 (0.0)
Fonticula alba H696-01050T0	0 (0.0)
Crassostrea gigas 405975190	0 (0.0)
Naegleria gruberi 290980695	0 (0.0)
Tetrahymena thermophila 118399676	0 (0.0)
Monosiga brevicollis 167521039	0 (0.0)
Acanthamoeba castellanii 470446361	0 (0.0)
Trichomonas vaginalis 154412871	0 (0.0)
Reticulomyxa filosa 569390586	1 (0.25)
Caenorhabditis elegans F39B1.1	1 (0.25)
Homo sapiens ENSP00000356155	1 (0.25)
Leishmania infantum 146081165	1 (0.25)
Homo sapiens ENSP00000265970	1 (0.25)
Nematostella vectensis 156225157	1 (0.25)
Dictyostelium purpureum 330802463	1 (0.25)
Trypanosoma cruzi 407833354	1 (0.25)
Entamoeba histolytica 183231243	1 (0.25)
Capsaspora owczarzaki 470291293	1 (0.25)
Drosophila melanogaster FBpp0075818	2 (0.5)
Caenorhabditis elegans B0334.8	2 (0.5)
Salpingoeca rosetta 326432665	2 (0.5)
Capsaspora owczarzaki 470306795	3 (0.75)
Oxytricha trifallax 403336721	4 (1.01)
Naegleria gruberi 290990903	7 (1.76)
Homo sapiens ENSP00000446444	8 (2.01)
Giardia intestinalis 559182176	10 (2.51)
Entamoeba histolytica 67483780	10 (2.51)
Monosiga brevicollis 167533638	11 (2.76)
Emiliania huxleyi 485633737	13 (3.27)
Acanthamoeba castellanii 470425472	13 (3.27)
Homo sapiens ENSP00000266497	14 (3.52)
Naegleria gruberi 290975002	15 (3.77)
Oxytricha trifallax 403359629	17 (4.27)
Crassostrea gigas 405975165	17 (4.27)
Acanthamoeba castellanii 470512532	19 (4.77)
Entamoeba histolytica 183232689	21 (5.28)
Acanthamoeba castellanii 470444588	21 (5.28)
Fonticula alba H696-02957T0	22 (5.53)
Batrachochytrium dendrobatidis 575486326	22 (5.53)
Dictyostelium purpureum 330806555	22 (5.53)
Phytophthora infestans 301103680	22 (5.53)
Toxoplasma gondii 237843183	22 (5.53)

Organism name and sequence ID	Number of gaps (percentage)
Salpingoeca rosetta 326432810	23 (5.78)
Caenorhabditis elegans B0025.1a	23 (5.78)
Emiliania huxleyi 485628976	23 (5.78)
Tetrahymena thermophila 118400017	23 (5.78)
Mucor circinelloides 511001770	23 (5.78)
Acanthamoeba castellanii 470412093	23 (5.78)
Paramecium tetraurelia 145541159	23 (5.78)
Physcomitrella patens 162685771	24 (6.03)
Arabidopsis thaliana 15219743	24 (6.03)
Crassostrea gigas 405952421	24 (6.03)
Rhizophagus irregularis 552917095	24 (6.03)
Capitella teleta 443710722	24 (6.03)
Nematostella vectensis 156224724	24 (6.03)
Homo sapiens ENSP00000262039	24 (6.03)
Saccharomyces cerevisiae YLR240W	24 (6.03)
Capsaspora owczarzaki 470321313	24 (6.03)
Mortierella verticillata MVEG-00850T0	24 (6.03)
Drosophila melanogaster FBpp0071973	24 (6.03)
Selaginella moellendorffii 300148168	25 (6.28)
Leishmania infantum 339898452	25 (6.28)
Plasmodium falciparum 124506225	25 (6.28)
Cryptosporidium parvum 66475452	25 (6.28)
Entamoeba histolytica 67483313	26 (6.53)
Selaginella moellendorffii 300148666	26 (6.53)
Trypanosoma cruzi 407860885	26 (6.53)
Cryptococcus neoformans 540393534	26 (6.53)
Nannochloropsis gaditana 585106701	27 (6.78)
Naegleria gruberi 290989255	27 (6.78)
Spironucleus salmonicida 558601893	28 (7.04)
Dictyostelium purpureum 330792962	29 (7.29)
Acanthamoeba castellanii 470467087	30 (7.54)
Phytophthora infestans 301100996	31 (7.79)
Ectocarpus siliculosus 299472549	32 (8.04)
Crassostrea gigas 405975626	33 (8.29)
Ectocarpus siliculosus 298705307	38 (9.55)
Dictyostelium purpureum 330842718	44 (11.06)
Reticulomyxa filosa 569374282	47 (11.81)
Mortierella verticillata MVEG-08250T0	48 (12.06)
Homo sapiens ENSP00000458238	49 (12.31)
Guillardia theta 428168346	49 (12.31)
Arabidopsis thaliana 30694536	50 (12.56)
Leishmania infantum 146093129	51 (12.81)
Tetrahymena thermophila 118351905	51 (12.81)
Salpingoeca rosetta 326427845	52 (13.07)
Naegleria gruberi 290999046	57 (14.32)
Naegleria gruberi 291000178	57 (14.32)
Ostreococcus lucimarinus 145344789	62 (15.58)
Monosiga brevicollis 167516396	64 (16.08)
Capitella teleta 443696965	64 (16.08)
Naegleria gruberi 290995492	68 (17.09)
Thalassiosira pseudonana 220971372	68 (17.09)
Chlamydomonas reinhardtii 2109289	69 (17.34)
Volvox carteri 302844923	69 (17.34)
Dictyostelium purpureum 330846412	77 (19.35)
Saccharomyces cerevisiae YNL267W	80 (20.1)
Mortierella verticillata MVEG-09379T0	84 (21.11)
Salpingoeca rosetta 326435514	85 (21.36)
Tetrahymena thermophila 146174543	86 (21.61)
Arabidopsis thaliana 15237610	88 (22.11)
Leishmania infantum 146099611	88 (22.11)
Monosiga brevicollis 167537642	88 (22.11)
Selaginella moellendorffii 300153477	88 (22.11)
Homo sapiens ENSP00000357869	89 (22.36)
Salpingoeca rosetta 326435786	89 (22.36)
Saccharomyces cerevisiae YLR305C	91 (22.86)
Nannochloropsis gaditana 585108648	148 (37.19)
Monosiga brevicollis 167520105	179 (44.97)
Thalassiosira pseudonana 220968505	204 (51.26)
Monosiga brevicollis 167520402	215 (54.02)

TABLE *B.1* – ***Nombre de gaps dans les séquences du jeu de données réduit des homologues catalytiques, après sélection des sites conservés.*** Les séquences sont triées par pourcentage croissant de gaps.

Organism name and sequence ID	Number of gaps (percentage)
Spizellomyces punctatus SPPG-07143T0	0 (0.0)
Danio rerio ENSDARP00000079663	0 (0.0)
Gasterosteus aculeatus ENSGACP00000003905	0 (0.0)
Takifugu rubripes ENSTRUP00000030991	0 (0.0)
Tetraodon nigroviridis ENSTNIP00000005287	0 (0.0)
Xiphophorus maculatus ENSXMAP00000000846	0 (0.0)
Latimeria chalumnae ENSLACP00000011766	0 (0.0)
Drosophila melanogaster FBpp0081466	0 (0.0)
Callorhinchus milii 632945103	0 (0.0)
Branchiostoma floridae 260782301	0 (0.0)
Capitella teleta 443696776	0 (0.0)
Apis mellifera 571508548	0 (0.0)
Thecamonas trahens AMSG-05614T0	0 (0.0)
Batrachochytrium dendrobatidis 575474224	0 (0.0)
Crassostrea gigas 405975645	0 (0.0)
Lottia gigantea 556094991	0 (0.0)
Saprolegnia diclina 530726775	0 (0.0)
Aphanomyces astaci 574103097	0 (0.0)
Canis lupus ENSCAFP00000009079	0 (0.0)
Ornithorhynchus anatinus ENSOANP00000010793	0 (0.0)
Bos taurus ENSBTAP00000027702	0 (0.0)
Pteropus vampyrus ENSPVAP00000006447	0 (0.0)
Homo sapiens ENSP00000349205	0 (0.0)
Otolemur garnettii ENSOGAP00000007000	0 (0.0)
Mus musculus ENSMUSP00000067400	0 (0.0)
Pelodiscus sinensis ENSPSIP00000002713	0 (0.0)
Anas platyrhynchos ENSAPLP00000010174	0 (0.0)
Ficedula albicollis ENSFALP00000013105	0 (0.0)
Taeniopygia guttata ENSTGUP00000004192	0 (0.0)
Xenopus tropicalis ENSXETP00000004982	0 (0.0)
Lepisosteus oculatus ENSLOCP00000001871	0 (0.0)
Oreochromis niloticus ENSONIP00000002550	1 (0.12)
Gadus morhua ENSGMOP00000003697	1 (0.12)
Ciona intestinalis ENSCINP00000005460	1 (0.12)
Nematostella vectensis 156218743	1 (0.12)
Aphanomyces invadans 574480269	1 (0.12)
Mucor circinelloides 511009779	2 (0.24)
Trichophyton rubrum 326460079	4 (0.48)
Aspergillus oryzae 83768900	4 (0.48)
Capsaspora owczarzaki 470304578	4 (0.48)
Coccidioides posadasii 303313395	4 (0.48)
Cryptococcus neoformans 540382592	5 (0.6)
Allomyces macrogynus AMAG-02895T0	6 (0.72)
Cryptococcus gattii 321264011	6 (0.72)
Helobdella robusta 555689993	6 (0.72)
Dictyostelium purpureum 330802485	7 (0.83)
Magnaporthe oryzae 389628626	7 (0.83)
Phaeodactylum tricornutum 219127594	7 (0.83)
Candida tropicalis 255727190	7 (0.83)
Phytophthora parasitica 568103383	9 (1.07)
Phytophthora infestans 301091919	9 (1.07)
Echinococcus granulosus 556521595	9 (1.07)
Blumeria graminis 521771122	9 (1.07)
Naegleria gruberi 291000750	9 (1.07)
Dictyostelium fasciculatum 470251580	10 (1.19)
Clonorchis sinensis 358341839	12 (1.43)
Fragaria vesca 470141421	13 (1.55)
Aplysia californica 524910995	14 (1.67)
Zea mays 413939320	14 (1.67)

Organism name and sequence ID	Number of gaps (percentage)
Physcomitrella patens 162680728	15 (1.79)
Selaginella moellendorffii 300154027	15 (1.79)
Eutrema salsugineum 557114017	15 (1.79)
Arabidopsis thaliana 15233564	15 (1.79)
Ostreococcus lucimarinus 145345011	16 (1.91)
Entamoeba histolytica 67473216	16 (1.91)
Entamoeba nuttalli 407036927	16 (1.91)
Monodelphis domestica ENSMODP00000015056	16 (1.91)
Acanthamoeba castellanii 470509949	17 (2.03)
Caenorhabditis elegans ZK930.1a	19 (2.26)
Triticum urartu 474434175	19 (2.26)
Micromonas pusilla 303271833	25 (2.98)
Leishmania braziliensis 154340533	26 (3.1)
Mortierella verticillata MVEG-06780T0	29 (3.46)
Thalassiosira pseudonana 220976169	30 (3.58)
Fonticula alba H696-01790T0	33 (3.93)
Chlorella variabilis 552825556	33 (3.93)
Monosiga brevicollis 167534887	36 (4.29)
Populus trichocarpa 550336423	37 (4.41)
Schistosoma mansoni 360043138	47 (5.6)
Saccharomyces cerevisiae YBR097W	50 (5.96)
Meleagris gallopavo ENSMGAP00000011917	51 (6.08)
Trypanosoma cruzi 407853206	54 (6.44)
Anolis carolinensis ENSACAP00000002347	59 (7.03)
Trypanosoma brucei 261334915	60 (7.15)
Tupaia belangeri ENSTBEP00000007529	64 (7.63)
Salpingoeca rosetta 326435714	69 (8.22)
Hydra vulgaris 449665199	72 (8.58)
Leishmania major 157871818	76 (9.06)
Leishmania infantum 146092133	76 (9.06)
Oxytricha trifallax 403376447	76 (9.06)
Leishmania donovani 398018212	76 (9.06)
Leishmania mexicana 401425064	76 (9.06)
Coccomyxa subellipsoidea 545368922	87 (10.37)
Loxodonta africana XP003420965	116 (13.83)
Ciona savignyi ENSCSAVP00000012067	124 (14.78)
Trichoplax adhaerens 196007140	127 (15.14)
Emiliania huxleyi 485639671	136 (16.21)
Astyanax mexicanus ENSAMXP00000015738	171 (20.38)
Saccoglossus kowalevskii 585653380	175 (20.86)
Ajellomyces capsulatus 154286454	175 (20.86)
Polysphondylium pallidum 281212562	179 (21.33)
Volvox carteri 302832507	183 (21.81)
Ectocarpus siliculosus 299117051	214 (25.51)
Rhizophagus irregularis 552925951	235 (28.01)
Erinaceus europaeus ENSEEUP00000007843	253 (30.15)
Gallus gallus ENSGALP00000036848	267 (31.82)
Rhizopus delemar 384495285	374 (44.58)
Erinaceus europaeus ENSEEUP00000002071	400 (47.68)
Blastocystis hominis 300176769	432 (51.49)
Reticulomyxa filosa 569406653	481 (57.33)
Petromyzon marinus ENSPMAP00000004074	534 (63.65)
Oryza sativa 115449223	588 (70.08)
Oryzias latipes ENSORLP00000025786	595 (70.92)
Guillardia theta 428179745	599 (71.39)
Thalassiosira oceanica 397603733	600 (71.51)
Sphaeroforma arctica SARC-12932T0	612 (72.94)
Trypanosoma vivax 343420766	672 (80.1)

TABLE *B.2* – ***Nombre de gaps dans les séquences du jeu de données des homologues de la protéine régulatrice de classe III, après sélection des sites conservés.*** Les séquences sont triées par pourcentage croissant de gaps.

Organism name and sequence ID	Number of gaps (percentage)
Callorhinchus milii 632962079	0 (0.0)
Xiphophorus maculatus ENSXMAP00000018652	0 (0.0)
Gasterosteus aculeatus ENSGACP00000020881	0 (0.0)
Takifugu rubripes ENSTRUP00000018795	0 (0.0)
Oreochromis niloticus ENSONIP00000013929	0 (0.0)
Takifugu rubripes ENSTRUP00000027603	0 (0.0)
Oryzias latipes ENSORLP00000021829	0 (0.0)
Danio rerio ENSDARP00000056212	0 (0.0)
Astyanax mexicanus ENSAMXP00000012290	0 (0.0)
Lepisosteus oculatus ENSLOCP00000012884	0 (0.0)
Ficedula albicollis ENSFALP00000009923	0 (0.0)
Anas platyrhynchos ENSAPLP00000006492	0 (0.0)
Meleagris gallopavo ENSMGAP00000011022	0 (0.0)
Gallus gallus ENSGALP00000023820	0 (0.0)
Lepisosteus oculatus ENSLOCP00000006953	0 (0.0)
Chrysemys picta 530627063	1 (0.19)
Callorhinchus milii 632963753	1 (0.19)
Oreochromis niloticus ENSONIP00000002191	1 (0.19)
Xiphophorus maculatus ENSXMAP00000015169	1 (0.19)
Tetraodon nigroviridis ENSTNIP00000022063	1 (0.19)
Gasterosteus aculeatus ENSGACP00000001371	1 (0.19)
Lepisosteus oculatus ENSLOCP00000001757	1 (0.19)
Otolemur garnettii ENSOGAP00000001020	1 (0.19)
Monodelphis domestica ENSMODP00000024566	2 (0.37)
Taeniopygia guttata ENSTGUP00000003114	2 (0.37)
Chrysemys picta 530622965	2 (0.37)
Ficedula albicollis ENSFALP00000010272	2 (0.37)
Gallus gallus ENSGALP00000023822	2 (0.37)
Pelodiscus sinensis ENSPSIP00000011890	2 (0.37)
Anolis carolinensis ENSACAP00000016070	2 (0.37)
Latimeria chalumnae ENSLACP00000015476	2 (0.37)
Homo sapiens ENSP00000274335	2 (0.37)
Gadus morhua ENSGMOP00000017031	2 (0.37)
Bos taurus ENSBTAP00000014594	2 (0.37)
Loxodonta africana ENSLAFP00000015303	2 (0.37)
Danio rerio ENSDARP00000056821	2 (0.37)
Oreochromis niloticus ENSONIP00000013171	2 (0.37)
Takifugu rubripes ENSTRUP00000026411	2 (0.37)
Pteropus vampyrus ENSPVAP00000007375	2 (0.37)
Gallus gallus ENSGALP00000005414	2 (0.37)
Ficedula albicollis ENSFALP00000014443	2 (0.37)
Canis lupus ENSCAFP00000011322	2 (0.37)
Mus musculus ENSMUSP00000056774	2 (0.37)
Tupaia belangeri ENSTBEP00000012016	3 (0.56)
Xenopus tropicalis ENSXETP00000046442	3 (0.56)
Oryzias latipes ENSORLP00000003326	3 (0.56)
Tetraodon nigroviridis ENSTNIP00000014006	3 (0.56)
Anas platyrhynchos ENSAPLP00000011511	3 (0.56)
Anas platyrhynchos ENSAPLP00000004356	4 (0.74)
Gadus morhua ENSGMOP00000003534	4 (0.74)
Gadus morhua ENSGMOP00000010892	5 (0.93)
Xiphophorus maculatus ENSXMAP00000013925	5 (0.93)
Ornithorhynchus anatinus ENSOANP00000002083	6 (1.11)
Monodelphis domestica ENSMODP00000002126	6 (1.11)
Mus musculus ENSMUSP00000034296	6 (1.11)
Homo sapiens ENSP00000222254	6 (1.11)
Loxodonta africana ENSLAFP00000000599	6 (1.11)
Canis lupus ENSCAFP00000022072	6 (1.11)
Taeniopygia guttata ENSTGUP00000008408	6 (1.11)
Saccoglossus kowalevskii 585680651	7 (1.3)
Pteropus vampyrus ENSPVAP00000001922	8 (1.48)
Bos taurus ENSBTAP00000003033	8 (1.48)
Branchiostoma floridae 260823974	9 (1.67)

Organism name and sequence ID	Number of gaps (percentage)
Xenopus tropicalis ENSXETP00000028313	9 (1.67)
Apis mellifera 66500538	10 (1.86)
Gasterosteus aculeatus ENSGACP00000020473	11 (2.04)
Taeniopygia guttata ENSTGUP00000000409	14 (2.6)
Crassostrea gigas 405968860	34 (6.31)
Capitella teleta 443694211	34 (6.31)
Amphimedon queenslandica 340378846	36 (6.68)
Otolemur garnettii ENSOGAP00000013371	57 (10.58)
Salpingoeca rosetta 326427517	63 (11.69)
Meleagris gallopavo ENSMGAP00000004955	74 (13.73)
Oreochromis niloticus ENSONIP00000023417	101 (18.74)
Erinaceus europaeus ENSEEUP00000007255	115 (21.34)
Petromyzon marinus ENSPMAT00000003687	121 (22.45)
Oryzias latipes ENSORLP00000013359	121 (22.45)
Gasterosteus aculeatus ENSGACP00000006864	122 (22.63)
Oreochromis niloticus ENSONIP00000009899	122 (22.63)
Oryzias latipes ENSORLP00000016725	122 (22.63)
Ornithorhynchus anatinus ENSOANP00000009424	124 (23.01)
Xenopus tropicalis ENSXETP00000007065	124 (23.01)
Anolis carolinensis ENSACAP00000015633	124 (23.01)
Bos taurus ENSBTAP00000003878	124 (23.01)
Loxodonta africana ENSLAFP00000003074	124 (23.01)
Homo sapiens ENSP00000361075	124 (23.01)
Canis lupus ENSCAFP00000006438	124 (23.01)
Mus musculus ENSMUSP00000030464	124 (23.01)
Gadus morhua ENSGMOP00000000809	124 (23.01)
Xiphophorus maculatus ENSXMAP00000018429	124 (23.01)
Amphimedon queenslandica 340375770	125 (23.19)
Tupaia belangeri ENSTBEP00000007253	125 (23.19)
Pteropus vampyrus ENSPVAP00000012903	125 (23.19)
Gasterosteus aculeatus ENSGACP00000021319	125 (23.19)
Danio rerio ENSDARP00000040333	127 (23.56)
Gadus morhua ENSGMOP00000012642	127 (23.56)
Tetraodon nigroviridis ENSTNIP00000009742	133 (24.68)
Takifugu rubripes ENSTRUP00000030226	133 (24.68)
Pelodiscus sinensis ENSPSIP00000012953	137 (25.42)
Erinaceus europaeus ENSEEUP00000008939	143 (26.53)
Takifugu rubripes ENSTRUP00000000305	143 (26.53)
Otolemur garnettii ENSOGAP00000005767	146 (27.09)
Drosophila melanogaster FBpp0303632	151 (28.01)
Xiphophorus maculatus ENSXMAP00000005400	151 (28.01)
Tetraodon nigroviridis ENSTNIP00000013096	159 (29.5)
Takifugu rubripes ENSTRUP00000017331	159 (29.5)
Xiphophorus maculatus ENSXMAP00000002496	160 (29.68)
Oreochromis niloticus ENSONIP00000015985	160 (29.68)
Aplysia californica 524870249	161 (29.87)
Gasterosteus aculeatus ENSGACP00000010853	162 (30.06)
Latimeria chalumnae ENSLACP00000008684	165 (30.61)
Lottia gigantea 556112584	168 (31.17)
Caenorhabditis elegans Y110A7A.10	181 (33.58)
Erinaceus europaeus ENSEEUP00000002251	186 (34.51)
Danio rerio ENSDARP00000106632	258 (47.87)
Nematostella vectensis 156227304	301 (55.84)
Astyanax mexicanus ENSAMXP00000020188	308 (57.14)
Astyanax mexicanus ENSAMXP00000003118	368 (68.27)
Lottia gigantea 556112583	385 (71.43)
Ciona intestinalis ENSCINP00000024453	390 (72.36)
Tetraodon nigroviridis ENSTNIP00000022692	391 (72.54)
Capsaspora owczarzaki 470298756	396 (73.47)
Amphimedon queenslandica 340373849	397 (73.65)
Hydra vulgaris 449671265	399 (74.03)
Trichoplax adhaerens 196000346	401 (74.4)
Monosiga brevicollis 167520620	450 (83.49)

TABLE *B.3* – ***Nombre de gaps dans les séquences du jeu de données des homologues des protéines régulatrice de classe IA, après sélection des sites conservés.*** Les séquences sont triées par pourcentage croissant de gaps.

Organism name and sequence ID	Number of gaps (percentage)
Pteropus vampyrus ENSPVAP00000009486	0 (0.0)
Monodelphis domestica ENSMODP00000010607	0 (0.0)
Canis lupus ENSCAFP00000025622	0 (0.0)
Oreochromis niloticus ENSONIP00000022874	0 (0.0)
Takifugu rubripes ENSTRUP00000009990	0 (0.0)
Danio rerio ENSDARP00000110608	0 (0.0)
Astyanax mexicanus ENSAMXP00000009979	0 (0.0)
Gasterosteus aculeatus ENSGACP00000009090	0 (0.0)
Lepisosteus oculatus ENSLOCP00000016346	0 (0.0)
Loxodonta africana ENSLAFP00000014379	0 (0.0)
Mus musculus ENSMUSP00000021283	0 (0.0)
Chrysemys picta 530646583	0 (0.0)
Otolemur garnettii ENSOGAP00000006947	0 (0.0)
Homo sapiens ENSP00000392812	0 (0.0)
Ficedula albicollis ENSFALP00000009800	1 (0.17)
Gallus gallus NP001025868	1 (0.17)
Anas platyrhynchos ENSAPLP00000015962	1 (0.17)
Xiphophorus maculatus ENSXMAP00000006631	1 (0.17)
Gasterosteus aculeatus ENSGACP00000025604	1 (0.17)
Xiphophorus maculatus ENSXMAP00000014544	1 (0.17)
Bos taurus ENSBTAP00000055710	1 (0.17)
Oryzias latipes ENSORLP00000019168	1 (0.17)
Anolis carolinensis ENSACAP00000017428	2 (0.33)
Pelodiscus sinensis ENSPSIP00000009891	3 (0.5)
Oryzias latipes ENSORLP00000003891	4 (0.67)
Callorhinchus milii 632942859	4 (0.67)
Meleagris gallopavo ENSMGAP00000002675	5 (0.83)
Xenopus tropicalis ENSXETP00000042245	5 (0.83)
Gadus morhua ENSGMOP00000019203	5 (0.83)
Oreochromis niloticus ENSONIP00000010381	12 (2.0)
Petromyzon marinus ENSPMAP00000000327	16 (2.67)
Danio rerio ENSDARP00000111599	44 (7.35)
Otolemur garnettii ENSOGAP00000006938	44 (7.35)
Xenopus tropicalis ENSXETP00000042238	45 (7.51)

Organism name and sequence ID	Number of gaps (percentage)
Anas platyrhynchos ENSAPLP00000015850	45 (7.51)
Lepisosteus oculatus ENSLOCP00000016321	47 (7.85)
Canis lupus ENSCAFP00000025610	47 (7.85)
Erinaceus europaeus ENSEEUP00000002256	47 (7.85)
Homo sapiens ENSP00000475670	47 (7.85)
Meleagris gallopavo ENSMGAP00000002634	47 (7.85)
Anolis carolinensis ENSACAP00000017414	48 (8.01)
Pelodiscus sinensis ENSPSIP00000012541	48 (8.01)
Pteropus vampyrus ENSPVAP00000009482	49 (8.18)
Gallus gallus ENSGALP00000001859	49 (8.18)
Danio rerio ENSDARP00000109865	50 (8.35)
Mus musculus ENSMUSP00000052522	50 (8.35)
Xiphophorus maculatus ENSXMAP00000006614	51 (8.51)
Callorhinchus milii 632942857	51 (8.51)
Oreochromis niloticus ENSONIP00000022879	52 (8.68)
Astyanax mexicanus ENSAMXP00000009984	52 (8.68)
Oryzias latipes ENSORLP00000003864	53 (8.85)
Takifugu rubripes ENSTRUP00000009667	55 (9.18)
Latimeria chalumnae ENSLACP00000004268	56 (9.35)
Bos taurus ENSBTAP00000021739	60 (10.02)
Oreochromis niloticus ENSONIP00000010373	62 (10.35)
Chrysemys picta 530646579	71 (11.85)
Ficedula albicollis ENSFALP00000009808	72 (12.02)
Erinaceus europaeus ENSEEUP00000007413	72 (12.02)
Tupaia belangeri ENSTBEP00000013428	81 (13.52)
Tetraodon nigroviridis ENSTNIP00000017101	131 (21.87)
Tupaia belangeri ENSTBEP00000013236	148 (24.71)
Loxodonta africana ENSLAFP00000005536	249 (41.57)
Latimeria chalumnae ENSLACP00000000071	266 (44.41)
Gasterosteus aculeatus ENSGACP00000025600	366 (61.1)
Gadus morhua ENSGMOP00000019166	370 (61.77)
Gasterosteus aculeatus ENSGACP00000025601	432 (72.12)
Gadus morhua ENSGMOP00000019174	438 (73.12)

TABLE *B.4* – ***Nombre de gaps dans les séquences du jeu de données des homologues des protéines régulatrice de classe IB, après sélection des sites conservés.*** Les séquences sont triées par pourcentage croissant de gaps.

Organism name and sequence ID	Number of gaps (percentage)
Chrysemys picta 530644153	0 (0.0)
Danio rerio ENSDARP00000120821	0 (0.0)
Latimeria chalumnae ENSLACP00000019372	0 (0.0)
Anas platyrhynchos ENSAPLP00000010610	0 (0.0)
Meleagris gallopavo ENSMGAP00000006951	0 (0.0)
Gallus gallus ENSGALP00000009863	0 (0.0)
Taeniopygia guttata ENSTGUP00000008892	0 (0.0)
Ficedula albicollis ENSFALP00000000804	0 (0.0)
Pelodiscus sinensis ENSPSIP00000013783	0 (0.0)
Anolis carolinensis ENSACAP00000000044	0 (0.0)
Mus musculus ENSMUSP00000126092	0 (0.0)
Pteropus vampyrus ENSPVAP00000012080	0 (0.0)
Homo sapiens ENSP00000265970	0 (0.0)
Canis lupus ENSCAFP00000036143	0 (0.0)
Loxodonta africana ENSLAFP00000003094	0 (0.0)
Monodelphis domestica ENSMODP00000034801	0 (0.0)
Ornithorhynchus anatinus ENSOANP00000012246	0 (0.0)
Callorhinchus milii 632941134	1 (0.09)
Chrysemys picta 530639296	1 (0.09)
Lepisosteus oculatus ENSLOCP00000003495	1 (0.09)
Xenopus tropicalis ENSXETP00000034942	1 (0.09)
Pteropus vampyrus ENSPVAP00000006751	1 (0.09)
Homo sapiens ENSP00000356155	1 (0.09)
Otolemur garnettii ENSOGAP00000010659	1 (0.09)
Mus musculus ENSMUSP00000076911	1 (0.09)
Anolis carolinensis ENSACAP00000007419	1 (0.09)
Meleagris gallopavo ENSMGAP00000002505	1 (0.09)
Gallus gallus ENSGALP00000000882	1 (0.09)
Monodelphis domestica ENSMODP00000001932	1 (0.09)
Xiphophorus maculatus ENSXMAP00000012426	2 (0.18)
Oreochromis niloticus ENSONIP00000019580	2 (0.18)
Bos taurus ENSBTAP00000017388	2 (0.18)
Loxodonta africana ENSLAFP00000014283	2 (0.18)
Canis lupus ENSCAFP00000014201	2 (0.18)
Astyanax mexicanus ENSAMXP00000025863	3 (0.27)
Gadus morhua ENSGMOP00000011250	3 (0.27)
Takifugu rubripes ENSTRUP00000038296	4 (0.36)
Tetraodon nigroviridis ENSTNIP00000005672	4 (0.36)
Gasterosteus aculeatus ENSGACP00000008939	4 (0.36)
Latimeria chalumnae ENSLACP00000019798	4 (0.36)
Oryzias latipes ENSORLP00000013515	7 (0.63)
Erinaceus europaeus ENSEEUP00000006595	7 (0.63)
Xiphophorus maculatus ENSXMAP00000012596	9 (0.81)
Apis mellifera 571502742	10 (0.9)
Lepisosteus oculatus ENSLOCP00000014902	10 (0.9)
Danio rerio ENSDARP00000110942	10 (0.9)
Astyanax mexicanus ENSAMXP00000010380	12 (1.08)
Oreochromis niloticus ENSONIP00000000365	14 (1.26)
Drosophila melanogaster FBpp0075818	14 (1.26)
Ficedula albicollis ENSFALP00000000443	15 (1.35)
Tupaia belangeri ENSTBEP00000006933	19 (1.71)
Callorhinchus milii 632963288	27 (2.43)
Gallus gallus ENSGALP00000021380	37 (3.32)
Mus musculus AAI50814	37 (3.32)

Organism name and sequence ID	Number of gaps (percentage)
Loxodonta africana ENSLAFP00000011952	37 (3.32)
Canis lupus ENSCAFP00000018530	37 (3.32)
Bos taurus ENSBTAP00000040697	38 (3.41)
Caenorhabditis elegans F39B1.1	39 (3.5)
Amphimedon queenslandica 340378016	44 (3.95)
Echinococcus granulosus 556520602	49 (4.4)
Erinaceus europaeus ENSEEUP00000013771	56 (5.03)
Pelodiscus sinensis XP006133177	59 (5.3)
Nematostella vectensis 156225157	60 (5.39)
Otolemur garnettii ENSOGAP00000001905	65 (5.84)
Oreochromis niloticus ENSONIP00000010680	68 (6.11)
Monosiga brevicollis 167533638	70 (6.29)
Clonorchis sinensis 358342191	73 (6.56)
Homo sapiens ENSP00000266497	75 (6.74)
Otolemur garnettii ENSOGAP00000002350	77 (6.92)
Schistosoma mansoni 353229343	78 (7.01)
Erinaceus europaeus ENSEEUP00000009401	78 (7.01)
Aplysia californica 524911800	78 (7.01)
Branchiostoma floridae 260790325	79 (7.1)
Pteropus vampyrus ENSPVAP00000000598	79 (7.1)
Latimeria chalumnae ENSLACP00000004519	82 (7.37)
Salpingoeca rosetta 326432665	85 (7.64)
Tupaia belangeri ENSTBEP00000003810	86 (7.73)
Tupaia belangeri ENSTBEP00000011511	86 (7.73)
Xiphophorus maculatus ENSXMAP00000007238	89 (8.0)
Chrysemys picta 530644775	91 (8.18)
Anas platyrhynchos ENSAPLP00000009675	91 (8.18)
Takifugu rubripes ENSTRUP00000043606	96 (8.63)
Capsaspora owczarzaki 470291293	103 (9.25)
Tetraodon nigroviridis ENSTNIP00000008764	109 (9.79)
Ornithorhynchus anatinus XP001506878	116 (10.42)
Taeniopygia guttata ENSTGUP00000012743	129 (11.59)
Meleagris gallopavo ENSMGAP00000014352	141 (12.67)
Ciona intestinalis ENSCINP00000032884	169 (15.18)
Pelodiscus sinensis ENSPSIP00000002245	253 (22.73)
Crassostrea gigas 405975165	268 (24.08)
Petromyzon marinus ENSPMAT00000006926	303 (27.22)
Capitella teleta 443696965	379 (34.05)
Lottia gigantea 556098278	452 (40.61)
Hydra vulgaris 449666338	471 (42.32)
Saccoglossus kowalevskii 585646586	496 (44.56)
Crassostrea gigas 405976593	556 (49.96)
Petromyzon marinus ENSPMAT00000010071	695 (62.44)
Saccoglossus kowalevskii 585718267	716 (64.33)
Bos taurus ENSBTAP00000043907	752 (67.57)
Tetraodon nigroviridis ENSTNIP00000019124	827 (74.3)
Helobdella robusta 555694902	829 (74.48)
Gasterosteus aculeatus ENSGACP00000004463	905 (81.31)
Petromyzon marinus ENSPMAT00000007051	907 (81.49)
Gadus morhua ENSGMOP00000012312	914 (82.12)
Gadus morhua ENSGMOP00000005503	948 (85.18)
Gasterosteus aculeatus ENSGACP00000013210	948 (85.18)
Gasterosteus aculeatus ENSGACP00000004467	969 (87.06)
Gadus morhua ENSGMOP00000012319	969 (87.06)

TABLE *B.5* – ***Nombre de gaps dans les séquences du jeu de données des homologues MIC des protéines catalytiques de classe II, après sélection des sites conservés.*** Les séquences sont triées par pourcentage croissant de gaps.

Organism name and sequence ID	Number of gaps (percentage)
Nematostella vectensis 156221153	0 (0.0)
Capitella teleta 443734586	1 (0.12)
Callorhinchus milii 632934759	2 (0.24)
Lepisosteus oculatus ENSLOCP00000005286	2 (0.24)
Xiphophorus maculatus ENSXMAP00000006655	2 (0.24)
Oreochromis niloticus ENSONIP00000009986	2 (0.24)
Takifugu rubripes ENSTRUP00000017655	2 (0.24)
Homo sapiens ENSP00000418143	3 (0.36)
Loxodonta africana ENSLAFP00000013155	3 (0.36)
Mus musculus ENSMUSP00000035037	3 (0.36)
Erinaceus europaeus ENSEEUP00000007636	3 (0.36)
Pteropus vampyrus ENSPVAP00000005473	3 (0.36)
Bos taurus ENSBTAP00000009083	3 (0.36)
Canis lupus ENSCAFP00000011245	3 (0.36)
Anolis carolinensis ENSACAP00000003172	3 (0.36)
Pelodiscus sinensis ENSPSIP00000002511	3 (0.36)
Ficedula albicollis ENSFALP00000004251	3 (0.36)
Gallus gallus ENSGALP00000008740	3 (0.36)
Meleagris gallopavo ENSMGAP00000004152	3 (0.36)
Xenopus tropicalis ENSXETP00000056285	3 (0.36)
Latimeria chalumnae ENSLACP00000014968	3 (0.36)
Gadus morhua ENSGMOP00000014980	3 (0.36)
Oryzias latipes ENSORLP00000019450	3 (0.36)
Drosophila melanogaster FBpp0083348	3 (0.36)
Branchiostoma floridae 260815076	4 (0.48)
Taeniopygia guttata ENSTGUP00000005489	4 (0.48)
Danio rerio ENSDARP00000102724	4 (0.48)
Tetraodon nigroviridis ENSTNIP00000017338	4 (0.48)
Lepisosteus oculatus ENSLOCP00000001301	5 (0.6)
Gasterosteus aculeatus ENSGACP00000021206	5 (0.6)
Takifugu rubripes ENSTRUP00000016050	5 (0.6)
Tetraodon nigroviridis ENSTNIP00000010637	5 (0.6)
Oryzias latipes ENSORLP00000011212	5 (0.6)
Oreochromis niloticus ENSONIP00000013444	5 (0.6)
Xiphophorus maculatus ENSXMAP00000004675	5 (0.6)
Gadus morhua ENSGMOP00000003184	5 (0.6)
Anolis carolinensis ENSACAP00000004867	5 (0.6)
Gallus gallus ENSGALP00000014504	5 (0.6)
Meleagris gallopavo ENSMGAP00000010212	5 (0.6)
Anas platyrhynchos ENSAPLP00000012778	5 (0.6)
Ficedula albicollis ENSFALP00000008750	5 (0.6)
Taeniopygia guttata ENSTGUP00000011093	5 (0.6)
Monodelphis domestica ENSMODP00000026633	5 (0.6)
Canis lupus ENSCAFP00000016537	5 (0.6)
Bos taurus ENSBTAP00000012168	5 (0.6)
Otolemur garnettii ENSOGAP00000005899	5 (0.6)
Loxodonta africana ENSLAFP00000002561	5 (0.6)
Homo sapiens ENSP00000263967	5 (0.6)
Danio rerio ENSDARP00000124781	5 (0.6)
Astyanax mexicanus ENSAMXP00000020452	5 (0.6)
Gadus morhua ENSGMOP00000011408	5 (0.6)
Oreochromis niloticus ENSONIP00000022585	5 (0.6)
Xiphophorus maculatus ENSXMAP00000008625	5 (0.6)
Gasterosteus aculeatus ENSGACP00000001555	5 (0.6)
Callorhinchus milii 632968054	5 (0.6)
Callorhinchus milii 632934984	5 (0.6)
Chrysemys picta 530584406	5 (0.6)
Monodelphis domestica ENSMODP00000023181	5 (0.6)
Otolemur garnettii ENSOGAP00000013693	5 (0.6)
Canis lupus ENSCAFP00000029210	5 (0.6)
Mus musculus ENSMUSP00000036434	5 (0.6)
Anas platyrhynchos ENSAPLP00000007963	5 (0.6)
Ficedula albicollis ENSFALP00000007638	5 (0.6)
Gallus gallus ENSGALP00000004061	5 (0.6)
Meleagris gallopavo ENSMGAP00000003869	5 (0.6)
Anolis carolinensis ENSACAP00000013119	5 (0.6)
Pelodiscus sinensis ENSPSIP00000004971	5 (0.6)
Gadus morhua ENSGMOP00000000741	5 (0.6)
Xiphophorus maculatus ENSXMAP00000011635	5 (0.6)
Takifugu rubripes ENSTRUP00000025672	5 (0.6)
Oryzias latipes ENSORLP00000006834	5 (0.6)
Oreochromis niloticus ENSONIP00000002983	5 (0.6)
Mus musculus ENSMUSP00000103878	5 (0.6)
Latimeria chalumnae ENSLACP00000017606	5 (0.6)
Gasterosteus aculeatus ENSGACP00000008159	5 (0.6)
Oreochromis niloticus ENSONIP00000010535	5 (0.6)
Tetraodon nigroviridis ENSTNIP00000009567	5 (0.6)
Takifugu rubripes ENSTRUP00000009530	5 (0.6)
Oryzias latipes ENSORLP00000014606	5 (0.6)
Xiphophorus maculatus ENSXMAP00000000601	5 (0.6)
Lottia gigantea 556102838	6 (0.72)
Crassostrea gigas 405975190	6 (0.72)
Tetraodon nigroviridis ENSTNIP00000018220	6 (0.72)
Gasterosteus aculeatus ENSGACP00000009701	6 (0.72)
Xenopus tropicalis ENSXETP00000015756	7 (0.85)
Anas platyrhynchos ENSAPLP00000003138	7 (0.85)
Gasterosteus aculeatus ENSGACP00000007464	7 (0.85)
Danio rerio ENSDARP00000017757	7 (0.85)
Lepisosteus oculatus ENSLOCP00000008374	7 (0.85)
Capitella teleta 443701283	8 (0.97)
Tupaia belangeri ENSTBEP00000010880	9 (1.09)
Tetraodon nigroviridis ENSTNIP00000005428	9 (1.09)
Takifugu rubripes ENSTRUP00000013971	10 (1.21)

Organism name and sequence ID	Number of gaps (percentage)
Ciona savignyi ENSCSAVP00000004359	11 (1.33)
Caenorhabditis elegans B0334.8	12 (1.45)
Schistosoma mansoni 353228720	13 (1.57)
Trichoplax adhaerens 196007024	13 (1.57)
Clonorchis sinensis 358334260	13 (1.57)
Xenopus tropicalis ENSXETP00000013019	16 (1.93)
Oryzias latipes ENSORLP00000011853	18 (2.17)
Xiphophorus maculatus ENSXMAP00000008835	18 (2.17)
Latimeria chalumnae ENSLACP00000007893	18 (2.17)
Gallus gallus ENSGALP00000013101	18 (2.17)
Meleagris gallopavo ENSMGAP00000014459	18 (2.17)
Taeniopygia guttata ENSTGUP00000003158	18 (2.17)
Anas platyrhynchos ENSAPLP00000006296	18 (2.17)
Pelodiscus sinensis ENSPSIP00000006579	18 (2.17)
Branchiostoma floridae 260831840	18 (2.17)
Lepisosteus oculatus ENSLOCP00000019486	19 (2.29)
Oreochromis niloticus ENSONIP00000014923	19 (2.29)
Helobdella robusta 555689542	19 (2.29)
Xiphophorus maculatus ENSXMAP00000016592	19 (2.29)
Apis mellifera 571563007	19 (2.29)
Ciona intestinalis ENSCINP00000010684	19 (2.29)
Gasterosteus aculeatus ENSGACP00000025368	20 (2.42)
Tetraodon nigroviridis ENSTNIP00000009686	20 (2.42)
Loxodonta africana ENSLAFP00000003840	20 (2.42)
Otolemur garnettii ENSOGAP00000011733	20 (2.42)
Homo sapiens ENSP00000392258	20 (2.42)
Canis lupus ENSCAFP00000005911	20 (2.42)
Mus musculus ENSMUSP00000082596	20 (2.42)
Anolis carolinensis ENSACAP00000014629	20 (2.42)
Callorhinchus milii 632943520	20 (2.42)
Amphimedon queenslandica 340370959	20 (2.42)
Tupaia belangeri ENSTBEP00000008487	21 (2.54)
Callorhinchus milii 632946473	21 (2.54)
Capsaspora owczarzaki 470309276	21 (2.54)
Astyanax mexicanus ENSAMXP00000004300	22 (2.66)
Danio rerio ENSDARP00000002818	22 (2.66)
Monodelphis domestica ENSMODP00000020477	22 (2.66)
Oreochromis niloticus ENSONIP00000002013	22 (2.66)
Helobdella robusta 555699397	22 (2.66)
Ciona intestinalis ENSCINP00000014231	23 (2.78)
Latimeria chalumnae ENSLACP00000012857	23 (2.78)
Homo sapiens ENSP00000446444	23 (2.78)
Bos taurus ENSBTAP00000025274	23 (2.78)
Pteropus vampyrus ENSPVAP00000004422	24 (2.9)
Amphimedon queenslandica 340370156	25 (3.02)
Ornithorhynchus anatinus ENSOANP00000012983	25 (3.02)
Gadus morhua ENSGMOP00000009961	26 (3.14)
Lepisosteus oculatus ENSLOCP00000016225	26 (3.14)
Capsaspora owczarzaki 470303443	26 (3.14)
Monosiga brevicollis 167521039	26 (3.14)
Petromyzon marinus ENSPMAT00000002453	31 (3.74)
Echinococcus granulosus 576697681	31 (3.74)
Lottia gigantea 556100131	33 (3.99)
Danio rerio ENSDARP00000054650	34 (4.11)
Oryzias latipes ENSORLP00000002186	34 (4.11)
Ciona savignyi ENSCSAVP00000000553	36 (4.35)
Pteropus vampyrus ENSPVAP00000013295	37 (4.47)
Latimeria chalumnae ENSLACP00000010097	37 (4.47)
Ornithorhynchus anatinus ENSOANP00000023160	41 (4.95)
Pelodiscus sinensis ENSPSIP00000016189	43 (5.19)
Loxodonta africana XP003413227	46 (5.56)
Gasterosteus aculeatus ENSGACP00000005515	53 (6.4)
Takifugu rubripes ENSTRUP00000008194	54 (6.52)
Tetraodon nigroviridis ENSTNIP00000003526	55 (6.64)
Amphimedon queenslandica 340369052	56 (6.76)
Gadus morhua ENSGMOP00000000052	58 (7.0)
Aplysia californica 524908046	65 (7.85)
Ciona savignyi ENSCSAVP00000007755	68 (8.21)
Erinaceus europaeus ENSEEUP00000013678	71 (8.57)
Takifugu rubripes XP003973030	73 (8.82)
Tupaia belangeri ENSTBEP00000005289	78 (9.42)
Salpingoeca rosetta 326430078	79 (9.54)
Aplysia californica 524899805	92 (11.11)
Tupaia belangeri ENSTBEP00000002551	93 (11.23)
Pteropus vampyrus ENSPVAP00000002313	98 (11.84)
Erinaceus europaeus ENSEEUP00000006927	104 (12.56)
Crassostrea gigas 405975626	109 (13.16)
Ficedula albicollis ENSFALP00000010709	129 (15.58)
Tetraodon nigroviridis ENSTNIP00000006420	148 (17.87)
Saccoglossus kowalevskii 585654121	150 (18.12)
Otolemur garnettii ENSOGAP00000003608	152 (18.36)
Salpingoeca rosetta 326435786	159 (19.2)
Monodelphis domestica ENSMODP00000003651	171 (20.65)
Bos taurus ENSBTAP00000027780	219 (26.45)
Hydra vulgaris 449686903	421 (50.85)
Hydra vulgaris 449686315	437 (52.78)
Petromyzon marinus ENSPMAT00000002863	438 (52.9)
Saccoglossus kowalevskii 585661864	459 (55.43)
Amphimedon queenslandica 340383255	486 (58.7)
Petromyzon marinus ENSPMAT00000000829	573 (69.2)
Monosiga brevicollis 167520402	599 (72.34)
Saccoglossus kowalevskii 585706768	599 (72.34)

TABLE *B.6* – ***Nombre de gaps dans les séquences du jeu de données des homologues MIC des protéines catalytiques de classe I, après sélection des sites conservés.*** Les séquences sont triées par pourcentage croissant de gaps.

Annexe C

Précisions pour EPINe

C.1 Références des PPI de la voie AKT/mTOR

N	Références
1	PI3K / Akt Signaling Pathway [416]
2	S.-Y. Lin *et al.* [417] *Protein phosphorylation-acetylation cascade connects growth factor deprivation to autophagy.*
3	B. Ravikumar *et al.* [333] *Regulation of mammalian autophagy in physiology and pathophysiology.*
4	P. Sini *et al.* [418] *Simultaneous inhibition of mTORC1 and mTORC2 by mTOR kinase inhibitor AZD8055 induces autophagy and cell death in cancer cells.*
5	N. Oshiro *et al.* [419] *Amino acids activate mammalian target of rapamycin (mTOR) complex 1 without changing Rag GTPase guanyl nucleotide charging.*
6	J. A. Martina *et al.* [420] *MTORC1 functions as a transcriptional regulator of autophagy by preventing nuclear transport of TFEB.*
7	T. B. Huber *et al.* [421]. *mTOR and rapamycin in the kidney : signaling and therapeutic implications beyond immunosuppression.*
8	H. Roca *et al.* [422] *CCL2, survivin and autophagy : new links with implications in human cancer.*
9	H. Roca *et al.* [423] *CCL2 protects prostate cancer PC3 cells from autophagic death via phosphatidylinositol 3-kinase/AKT-dependent survivin up-regulation.*
10	M C. Maiuri *et al.* [424] *Self-eating and self- killing : crosstalk between autophagy and apoptosis.*
11	E. Rozengurt. [425] *Mechanistic target of rapamycin (mTOR) : a point of convergence in the action of insulin/IGF-1 and G protein-coupled receptor agonists in pancreatic cancer cells.*
12	C. Muñoz-Pinedo, N. *et al.* [426]*Cancer metabolism : current perspectives and future directions.*
13	V. Deretic *et al.* [] *Autophagy in infection, inflammation and immunity.*
14	J. A. McCubrey *et al.* [427] *Multifaceted roles of GSK-3 and Wnt/β-catenin in hematopoiesis and leukemogenesis : opportunities for therapeutic intervention.*
15	S. Kongara et V. Karantza [334] *The interplay between autophagy and ROS in tumorigenesis.*
16	D. Tang *et al.* [428] *Endogenous HMGB1 regulates autophagy.*
17	N. Chen et J. Debnath [429] *Autophagy and tumorigenesis.*
18	H. Cheong *et al.* [430] *Therapeutic targets in cancer cell metabolism and autophagy.*
19	S. Alers *et al.* [431] *Role of AMPK-mTOR- Ulk1/2 in the regulation of autophagy : cross talk, shortcuts, and feedbacks..*
20	Y.-M. Kim et D.-H. Kim [432] *dRAGging amino acid-mTORC1 signaling by SH3BP4.*
21	Y.-M. Kim *et al.* [433] *SH3BP4 is a negative regulator of amino acid-Rag GTPase-mTORC1 signaling.*
22	F. Nazio et F. Cecconi [434] *mTOR, AMBRA1, and autophagy : an intricate relationship.*
23	C. A. Mercer *et al.* [435] *A novel, human Atg13 binding protein, Atg101, interacts with ULK1 and is essential for macroautophagy.*

TABLE *C.1* – ***Références des articles utilisés pour reconstruire la voie de signalisation de la Figure 5.3.***

C.2 Résultats des 62 protéines de la voie AKT/mTOR

Nom de la protéine	ID	Nbe d'homologues		Transcrit Altern. (%)	Longueur d'alignement	Nbe sites conservés		Modèle évolutif sélectionné
		avant tri	après tri			(aa)	(%)	
AKTS1	Q96B36	48	26	45,83	513	254	49,51	JTT+G
INS	P01308	48	31	35,42	232	113	48,71	JTT+G
ATG13	O75143	82	44	46,34	822	500	60,83	JTT+G
BAKOR	Q6ZNE5	52	48	7,69	1039	433	41,67	LG+G
SH3B4	Q9P0V3	78	65	16,67	1271	800	62,94	JTT+G
TSC1	Q92574	87	65	25,29	2410	676	28,05	JTT+G
PI3R5	Q8WYR1	98	67	31,63	1125	578	51,38	JTT+G
PI3R6	Q5UE93	97	67	30,93	1126	583	51,78	JTT+G
MDM2	Q00987	150	74	50,67	1468	358	24,39	JTT+G
PRR5	P85299	113	75	33,63	1057	291	27,53	JTT+G
SIN1	Q9BPZ7	100	75	25,00	1681	371	22,07	LG+G
ATG101	Q9BSB4	104	87	16,35	571	163	28,55	LG+G
RICTR	Q6R327	125	104	16,80	8742	775	8,87	LG+G
IRS1	P35568	136	109	19,85	3500	630	18,00	JTT+G
UVRAG	Q9P2Y5	144	130	9,72	2951	126	4,27	LG+G
P53	P04637	234	131	44,02	1747	269	15,40	JTT+G
BECN1	Q14457	166	143	13,86	3092	282	9,12	LG+G
BECN2	A8MW95	167	144	13,77	2750	273	9,93	LG+G
TFEB	P19484	293	165	43,69	4647	300	6,46	JTT+G
RRAGB	Q5VZM2	221	207	6,33	4282	235	5,49	LG+G
RS6	P62753	241	226	6,22	889	232	26,10	WAG+G
AMRA1	Q9C0C7	291	259	11,00	8565	226	2,64	JTT+G
RRAGA	Q7L523	293	266	9,22	4344	233	5,36	LG+G
RRAGC	Q9HB90	298	270	9,40	4434	234	5,28	LG+G
RRAGD	Q9NQL2	298	270	9,40	4434	234	5,28	LG+G
RUBIC	Q92622	375	279	25,60	4751	133	2,80	LG+G
DPTOR	Q8TB45	408	310	24,02	8199	39	0,48	LG+G
BCL2	P10415	399	328	17,79	1891	55	2,91	LG+G
B2CL1	Q07817	403	330	18,11	2056	59	2,87	LG+G
PI3R4	Q99570	370	355	4,05	10382	209	2,01	LG+G
BIRC5	O15392	498	370	25,70	8935	80	0,90	LG+G
LST8	Q9BVC4	648	372	42,59	8995	194	2,16	LG+G
P55G	Q92569	538	391	27,32	5714	64	1,12	LG+G
KAT5	Q92993	553	425	23,15	8241	191	2,32	LG+G
RPTOR	Q8N122	560	432	22,86	12419	58	0,47	LG+G
RBCC1	Q8TDY2	581	435	25,13	29859	165	0,55	LG+G
PTEN	P60484	679	459	32,40	7365	131	1,78	LG+G
SHLB1	Q9Y371	747	512	31,46	7776	120	1,54	LG+G
TSC2	P49815	811	556	31,44	10961	107	0,98	LG+G
1433G	P61981	819	700	14,53	6111	181	2,96	LG+G
1433Z	P63104	821	702	14,49	5464	179	3,28	LG+G
HMGB1	P09429	933	707	24,22	4723	33	0,70	LG+G
P85B	O00459	1020	727	28,73	8861	79	0,89	LG+G
FOXO3	O43524	1269	810	36,17	3320	68	2,05	LG+G
MTOR	P42345	1069	858	19,74	22722	154	0,68	LG+G
PK3CA	P42336	1092	918	15,93	17116	157	0,92	LG+G
PK3CG	P48736	1108	927	16,34	15163	136	0,90	LG+G
PK3CD	O00329	1107	928	16,17	14628	170	1,16	LG+G
PK3C3	Q8NEB9	1115	940	15,70	15032	170	1,13	LG+G
PK3CB	P42338	1126	948	15,81	17546	122	0,70	LG+G
P85A	P27986	1383	952	31,16	12314	60	0,49	LG+G
KS6B1	P23443	1212	965	20,38	5186	216	4,17	LG+G
AKT1	P31749	1326	1025	22,70	5709	225	3,94	LG+G
PDPK1	O15530	1493	1208	19,09	6955	193	2,77	LG+G
AAPK1	Q13131	1567	1226	21,76	5406	192	3,55	LG+G
AAPK2	P54646	1598	1243	22,22	5281	169	3,20	LG+G
GSK3A	P49840	1628	1293	20,58	2807	150	5,34	LG+G
GSK3B	P49841	1594	1307	18,01	2947	157	5,33	LG+G
RAB5A	P20339	1614	1459	9,60	3188	131	4,11	LG+G
IGF1R	P08069	2013	1503	25,34	10137	118	1,16	LG+G
ULK1	O75385	2169	1779	17,98	6202	113	1,82	LG+G
RHEB	Q15382	2039	1826	10,45	3307	89	2,69	LG+G
moyennes :		687,89	542,79	22,38	6376,79	226,39	11,23	

TABLE *C.2* – ***Résultats obtenus pour les 62 protéines humaines de la voie AKT/mTOR.*** Les jeux de données sont triés par nombre d'homologues croissant.

C.3 Paramètres d'EPINe utilisés pour l'analyse de la voie AKT/mTOR

Les paramètres d'EPINe utilisés pour l'étude de la voie AKT

- E-value seuil $< 1 \times 10^{-15}$, 5 itérations de PSI-BLAST,
- BATfinder avec les options :
  - MAFFT `-auto`,
  - `-short`,
  - `-gap`,
  - sélection du modèle ou de l'option `-SD` selon la table 5.3,
  - si modèle évolutif : 30 réplicats de bootstrap,
- alignement du fichier avec un seul transcrit par locus génomique avec MAFFT en mode automatique (option `-auto`),
- options de BMGE :
  - si l'alignement fait moins de 6000 aa :
    - matrice BLOSUM30,
    - région conservée d'au moins 3 aa (`-b 3`),
    - une proportion de gaps autorisée de 40% (`-g 0.4`),
  - si l'alignement fait plus de 7000 aa :
    - matrice BLOSUM30,
    - région conservée d'au moins 1 aa (`-b 1`),
    - une proportion de gaps autorisée de 60% (`-g 0.6`),
- PhyML avec les options :
  - modèle évolutif sélectionné par ProtTest,
  - 4 catégories de taux d'évolution (`-c 4`),
  - si loi Γ sélectionnée par ProtTest, estimation du paramètre α par maximum de vraisemblance ((`-a e`).

C.4 Arbre de référence des Eucaryotes

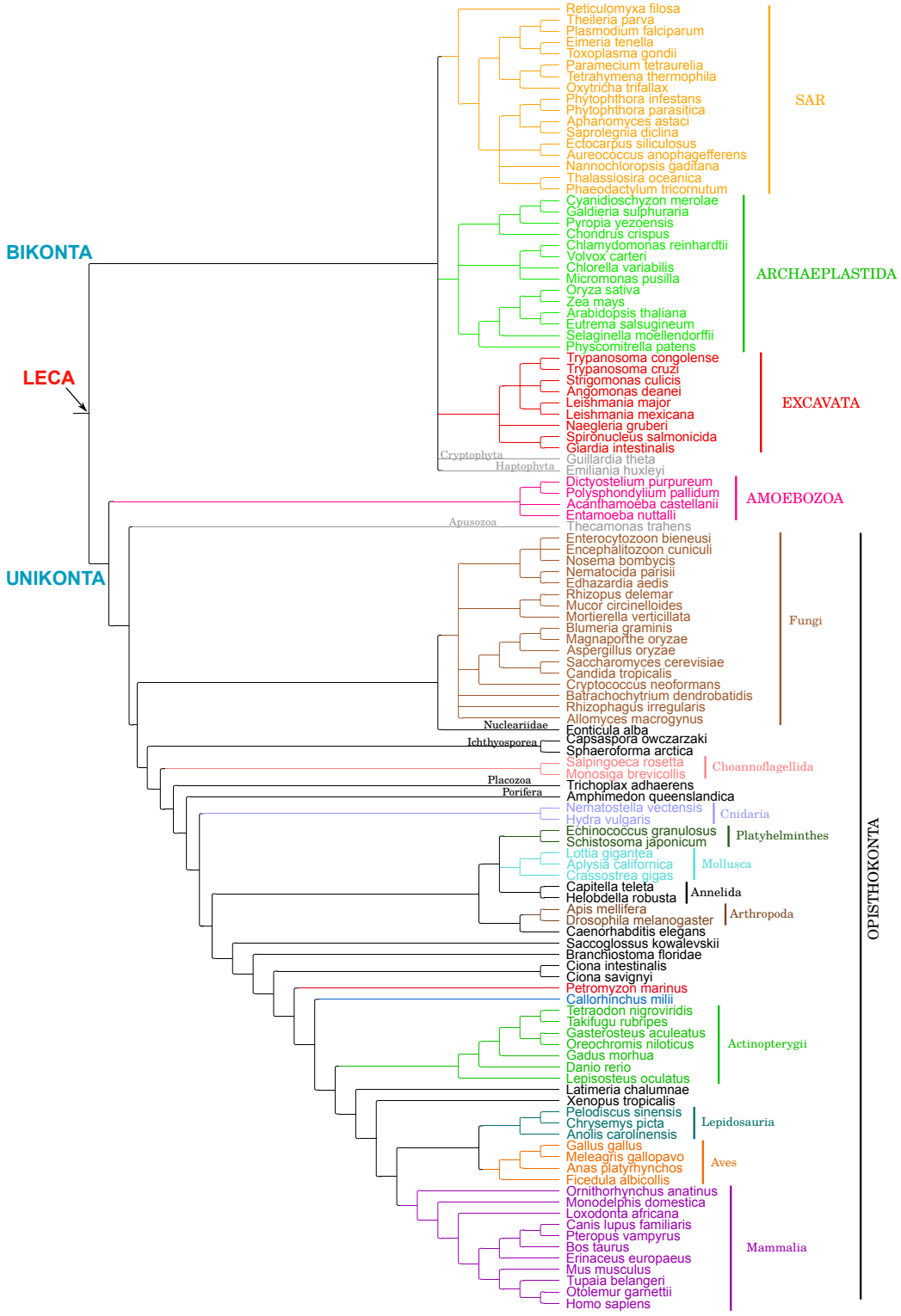

FIGURE *C.1* – ***Arbre phylogénétique eucaryote utilisé comme référence.*** L'arbre est composé des 116 espèces dont les génomes sont dans les deux banques de données interrogées par EPINe.