UNIVERSITÉ DE MONTRÉAL

UNE APPROCHE PHYLOGÉNOMIQUE POUR INFÉRER L'ÉVOLUTION DES EUCARYOTES

PAR

NAIARA RODRÍGUEZ-EZPELETA

DÉPARTEMENT DE BIOCHIMIE

FACULTÉ DE MÉDECINE

THÈSE PRÉSENTÉE À LA FACULTÉ DES ÉTUDES SUPÉRIEURES
EN VUE DE L'OBTENTION DU GRADE DE PHILOSOPHIÆ DOCTOR (PH.D.)
EN BIOCHIMIE

FÉVRIER, 2007

UNIVERSITÉ DE MONTRÉAL

FACULTÉ DES ÉTUDES SUPÉRIEURES

CETTE THÈSE INTITULÉE :

« UNE APPROCHE PHYLOGÉNOMIQUE POUR INFÉRER L'ÉVOLUTION DES EUCARYOTES »

PRÉSENTÉE PAR :

NAIARA RODRÍGUEZ-EZPELETA

A ÉTÉ ÉVALUÉE PAR UN JURY COMPOSÉ DES PERSONNES SUIVANTES :

MICHEL BOUVIER – Président-rapporteur

B. FRANZ LANG – Directeur de recherche

ANDREW J. ROGER – Membre du jury

DAVID MOREIRA – Examinateur externe

NORMAND BRISSON – Représentant du doyen de la FES

RÉSUMÉ

En dépit de l'immense quantité de phylogénies moléculaires publiées ces deux dernières décennies, la reconstruction de l'arbre des eucaryotes est loin d'être achevée. Ceci est principalement dû au faible signal phylogénétique contenu dans les jeux de données basés sur un petit nombre de gènes. Dans le but d'adresser des questions encore irrésolues, mais clés dans l'histoire évolutive des eucaryotes, nous avons utilisé la phylogénomique, une nouvelle approche qui cherche à combiner le signal contenu dans une multitude de gènes pour augmenter la résolution des phylogénies.

Nos études ont permis d'élucider trois sujets longuement débattus. Premièrement, les analyses basées sur 50 et 143 protéines codées respectivement dans le plaste et dans le noyau confirment la monophylie des plastes ainsi que celle des eucaryotes photosynthétiques primaires. Ensemble, ces deux résultats supportent une origine unique des plastes par l'endosymbiose entre une cyanobactérie et un eucaryote. Deuxièmement, les phylogénies basées sur trois jeux de données comprenant respectivement 125, 50, et 33 protéines codées dans le noyau, dans le plaste et dans la mitochondrie placent l'algue unicellulaire *Mesostigma* dans les streptophytes. Pour la première fois, les génomes des trois compartiments cellulaires donnent des résultats congruents concernant la position de cette espèce. Finalement, l'analyse de la concaténation de 170 protéines codées dans le noyau a permis de situer pour la première fois deux groupes d'eucaryotes unicellulaires considérés « primitifs », les jakobides et les malawimonadines, dans l'arbre des eucaryotes.

Ces études nous ont permis d'explorer le potentiel de la phylogénomique pour inférer l'histoire évolutive des eucaryotes. Nous concluons que, pour résoudre les branches profondes de l'arbre des eucaryotes, l'utilisation d'un grand nombre de gènes d'une multitude d'espèces s'impose. D'autre part, les artefacts d'inférence phylogénétique sont aggravés avec l'augmentation de la taille des jeux de données et des résultats parfaitement résolus mais incorrects sont aussi possibles. En attendant le développement de meilleurs modèles d'évolution, l'élimination sélective d'espèces et/ou de positions est une méthode efficace pour augmenter le signal phylogénétique et d'éviter ainsi des conclusions erronées.

MOTS-CLÉS : Phylogénomique, erreur systématique, chloroplaste, mitochondrie, Plantae, excavés, *Mesostigma*, jakobides, malawimonadines, glaucophytes.

ABSTRACT

Despite the large number of molecular phylogenies published during the last twenty years, the reconstruction of the eukaryotic tree is far from being completed. This is mainly due to the weak phylogenetic signal present in single or few-gene based datasets. In order to address still unresolved but key events in the evolutionary history of eukaryotes, we have used phylogenomics –a new approach that seeks to combine the signal existing in multiple genes to increase phylogenetic resolution.

Three long debated issues have been deciphered by our studies. First, the analyses based on 50 plastid and 143 nuclear encoded proteins confirm, respectively, the monophyly of plastids and of primary photosynthetic eukaryotes. Together, these results support a single origin of plastids by the endosymbiosis between a cyanobacterium and a eukaryote. Second, phylogenies based on three separate datasets of 125 nuclear, 50 plastid and 33 mitochondrial encoded proteins, respectively place the unicellular alga *Mesostigma* within the streptophytes. For the first time, the genomes of the three cellular compartments give congruent results concerning the position of this organism. Finally, the analysis of the concatenation of 170 nuclear encoded proteins has positioned, for the first time, two groups of so-called "primitive eukaryotes" (the jakobids and the malawimonads) in the eukaryotic tree.

These studies are useful to explore the potential of phylogenomics to infer the evolutionary history of eukaryotes. We conclude that the resolution of the deepest branches in the eukaryotic tree requires the use of a large number of genes from a multitude of species. On the other hand, phylogenetic reconstruction artifacts are exacerbated in large datasets, and perfectly resolved but misleading results are also expected. While waiting for the development of better models of evolution, the selective removal of species and/or sites is an efficient approach to increase phylogenetic signal and thus, to avoid erroneous conclusions.

KEYWORDS : Phylogenomics, systematic error, chloroplast, mitochondrion, Plantae, excavates, *Mesostigma*, jakobids, malawimonads, glaucophytes.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABREVIATIONS

+Γ	Modèle qui prend en compte l'hétérogénéité du taux de substitution entre les sites
+F	Modèle qui inclut les fréquences de chaque état estimées à partir du jeu de données
+I	Modèle qui inclut une fraction de sites invariables estimée à partir du jeu de données
A	Adénine (nucléotide) ou Alanine (acide aminé)
ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
AIC	*Akaike Information Criterion*
ARN	Acide ribonucléique
ARNr	ARN ribosomal
ATP	Adénine triphosphate
atp	Gène codant pour l'ATP synthase
AU	Test *Approximately Unbiased*
BIC	*Bayesian Information Criterion*
BLAST	*Basic Local Alignment Search Tool*
BRH	*Best Reciprocal Hit*
C	Cytosine (nucléotide) ou Cystéine (acide aminé)
Cpn60	Chaperonne 60
cpREV	Matrice de substitution pour protéines chloroplastiques
cob	Gène codant pour le cytochrome b
cox	Gène codant pour la cytochrome oxydase

D	Acide aspartique
DHFR	Dihydrofolate réductase
E	Acide glutamique
EF	Facteur d'élongation
EST	*Expressed sequence tag* (Marqueur de séquence exprimée)
F	Phénylalanine
F81	Modèle Felsenstein 81
FBA	Fructose-bisphosphate aldolase
GAPDH	Glycéraldéhyde-3-phosphate déshydrogénase
gap	Gène codant pour la GAPDH
G	Guanine (nucléotide) ou Glycine (acide aminé)
GTR	Modèle *General Time Reversible*
H	Histidine
HKY	Modèle Hasegawa, Kishino et Yano
HSP	*Heat-Shock Protein*
I	Isoleucine
JC	Modèle Jukes et Cantor
JTT	Matrice de substitution de Jones, Taylor and Thornton
K	Lysine
K2P	Modèle Kimura 2 paramètres
KH	Test de Kishino et Hasegawa
L	Leucine
LogDet	Logarithme du déterminant
LRT	*Likelihood Ratio Test*
M	Méthionine

MCMC	*Markov Chain Monte Carlo*
MCMCMC	*Metropolis Coupling Markov Chain Monte Carlo*
Mpb	Méga paires de base
mtREV	Matrice de substitution pour protéines mitochondriales
N	Asparagine
nad	Géne codant pour la NADH-ubiquinone oxydoréductase
NNI	*Nearest Neighbor Interchange*
P	Proline
PCR	*Polymerase Chain Reaction* (réaction en chaîne de la polymérase)
psbB	Gène codant pour la protéine B du photosystème II
psbH	Gène codant pour la protéine H du photosystème II
psbN	Gène codant pour la protéine N du photosystème II
PSU ARNr	Petite sous unité de l'ARN ribosomal
Q	Glutamine
R	Arginine
rbcL	Gène codant pour la RUBISCO
RELL	*Resampling Esitmated Log-Likelihood*
RNase P	Ribonucléase P
RPB1	ARN polymérase II nucléaire
RPL22	Protéine 22 de la grande sous-unité ribosomale
RPL24A	Protéine 24A de la grande sous-unité ribosomale
rpo	Gène codant pour la ARN polymérase
rps2	Gène codant pour la protéine 2 de la PSU ribosomale
RUBISCO	ribulose-1,5-bisphosphate carboxylase/oxygénase

S	Sérine
SF	*Slow/Fast*
SH	Test de Shimodaira et Hasegawa
SPR	*Subtree Pruning and Regrafting*
T	Tyrosine (nucléotide) ou Thréonine (acide aminé)
TBR	*Tree Bisection and Reconnection*
TIC110	Protéine 110 du translocon de la membrane interne du plaste
TOC34	Protéine 34 du translocon de la membrane externe du plaste
TS	Thymidylate synthase
tufA	Gène codant pour le facteur d'élongation Tu
UT	Unité taxonomique
V	Valine
VB	Valeur de bootstrap
W	Tryptophane
WAG	Matrice de substitution de Whelan et Goldman
Y	Tyrosine

Dena galduta zegoela pentsatu ondoren,
traba guztiak gainditzeko eta bizitza berri
bat hasteko gai izan denari,

eta laguntza handiena eman dionari ere,
noski,

REMERCIEMENTS

Je tiens à exprimer en premier lieu ma reconnaissance envers mon directeur de recherche, B. FRANZ LANG, qui m'a donné l'opportunité de me joindre à son équipe malgré ~~ma~~ manque d'expérience en recherche et mes connaissances limitées en biochimie, en évolution et en informatique. Franz, la liberté que tu m'as donnée pour choisir et pour mener mon projet dès mon arrivée m'a un peu apeuré et quelques fois (oui, je l'avoue) un peu fâché. Par contre, j'ai vite réalisé que tout ça constituait une partie très importante de ma formation et que tu as toujours été là pour me recadrer quand je m'éloignais trop du chemin. J'apprécie énormément tous les moyens (financiers et humains) que tu as mis à ma disposition, la confiance que tu m'as accordée, tes précieux conseils, nos discussions animées et les séances de rédaction dans ton bureau. J'espère sincèrement que nous aurons la chance de travailler ensemble à nouveau.

Officiellement, je n'ai eu qu'un directeur de recherche pendant mon doctorat, mais pratiquement, on peut dire qu'HERVÉ PHILIPPE a bien eu un rôle de co-directeur. Hervé, je serais encore noyée dans les données si tu ne m'avais pas appris à nager dans ce vaste océan qui est la phylogénomique. Ces discussions de vendredi après-midi où on essayait de trouver des chemins dans les forêts d'arbres phylogénétiques vont énormément me manquer. Je te remercie de m'avoir prêté Obelix et son Petit Village Gaulois (qui on fait une grande partie du travail présenté dans cette thèse), de m'avoir écouté dans mes moments de panique, d'avoir compté sur moi pour des nombreux projets et, finalement, de ta façon de ~~donner~~ la pression sans vraiment le faire (et je cite ici ta réponse favorite : « Pour hier! »).

Je tiens à remercier les membres de mon comité de thèse et du jury de mon examen prédoctoral (PASCAL CHARTRAND, LUC DESGROSEILLERS, FRANÇOIS-JOSEPH LAPOINTE et DAVID MORSE), qui ont contribué à l'avancement de mon projet avec des questions, des critiques et des idées. En particulier, les conseils de FRANÇOIS-JOSEPH LAPOINTE ont été d'une grande aide. J'aimerais aussi remercier mes coauteurs GERTRAUD BURGER (Département de Biochimie), MICHAEL W. GRAY (Halifax, Canada), MICHAEL MELKONIAN (Cologne, Allemagne), BUKHARD BECKER (Cologne, Allemagne), WOLFGANG LÖFFELHARDT (Vienne, Autriche), SUZANNE C. BUREY (Vienne, Autriche) et HANS BOHNERT (Urbana-Champaign, USA).

Je remercie le GOUVERNEMENT BASQUE (*Eusko Jaurlaritza*) pour m'avoir octroyé une bourse d'études pour venir faire mon doctorat à Montréal, et l'UNIVERSITÉ DE MONTRÉAL, la FONDATION SIMON-PIERRE NOËL, le CENTRE ROBERT CEDERGREN et GÉNOME QUÉBEC pour du soutien financier additionnel. Je remercie également le Dr. MICHAEL CUMMINGS de m'avoir donné l'opportunité de participer, d'abord comme étudiante et ensuite deux fois comme assistante d'enseignement, au *Workshop on Molecular Evolution*, Woods Hole, USA. Je suis aussi très reconnaissante en vers les étudiants et les professeurs que j'ai rencontrés pendant mes séjours à Woods Hole, avec qui j'ai partagé des discussions très formatrices. Merci aussi aux Drs. DAVID PENNY et PETER LOCKHART pour m'avoir accueilli dans leur groupe de recherche pendant le fabuleux été néozelandais.

J'ai côtoyé un grand nombre de personnes au cours de mes cinq années de doctorat. Parmi celles avec qui j'ai travaillé au *wet-lab*, j'ai une mention très spéciale pour LISE FORGET. Laïs, je ne peux pas imaginer une autre façon d'endurer des expériences qui ne fonctionnent pas qu'en ta compagnie. Ton expertise technique, tes nombreux conseils et ta patience à répondre à toutes mes questions ont été très précieux et souvent suffisants pour me faire recommencer avec le même enthousiasme que la première fois. Ma deuxième mention spéciale va, bien sur, pour ELIAS SEIF, qui m'a appris de « tas de choses », entre autres, comment ne pas dégrader mes ARNs (!). Elias, appart tes cotés animateur, blagueur et commère si nécessaires pour la vie dans un laboratoire, j'ai énormément apprécié ton amitié, nos conversations dans le café d'après-midi, ton écoute et tes pieds à terre (qui nous font descendre quand on monte trop haut). Ah! Et j'espère que tu n'oublieras pas que les vieux riches du sud de l'Espagne seront toujours là comme alternative (*aaaah! No hay que llorar, que la vida es un carnaval...*).

J'ai aussi une pensée très spéciale pour le reste des personnes qui m'on accompagnée pendant mes années au labo: JESSICA, pour tes « vraisemblables » conseils (on se rejoins en Europe!); YANNICK, pour ce contagieux intérêt scientifique; JULIEN, pour ton expérience « radioactive »; DELPHINE, pour ton rire; GUERLINE, pour tes couleurs; DENNIS, *for your questions*; WILLIAM, pour nos discussions, JOANNIE, pour ta passion « microscopique », mais aussi « électronique » (ohmmmmm); NICOLAS, pour tes appétissants dîners; RACHID, pour être si agréablement tannant; SHONA (Teixeriña), *por tu ayuda con los bichitos y por dar tanto ambiente al laboratorio (con tus motes y monigotes todo es mucho más divertido)*; JEAN-SÉBASTIEN, pour être l'esclave le plus optimiste (on a des bandes!); et,

finalement, JUNG HWA, YUN, ZHANG et, très spécialement, JEAN-FRANÇOIS, pour votre rigoureux travail: s'il n'y a pas de séquences, il n'y a pas de phylogénie!

Pendant ma période au *dry-lab*, j'ai eu la chance de travailler avec HENNER BRINKMANN. Henner, ton cours intensif de MUST était, en effet, un *must* pour la continuité de mon travail. Définitivement, ta façon d'argumenter, tes changements instantanés de sujet et tes *cocktails* seront des souvenirs très agréables de mon passage à Montréal. Il y a, par contre, quelque chose de toi que je voudrais vite oublier: toutes les pages sur la GAPDH et les « KKKinases » que tu m'as fait lire! ;-). Merci BÉATRICE, pour déboguer SCaFoS à des heures intempestives de la nuit; OLIVIER, pour ton expertise « scriptienne » et pour être un excellent compagnon de pauses; FRED, pour tes relectures et bonnes suggestions; DAVID, *for your unvaluable help installing programs*; FABRICE, pour être le meilleur « BBQ-man », LIISA, pour être une excellente décontaminatrice; NIC, pour être si calme et attentif; NICO, pour tes idées enchaînées; YAN, *for trying so hard to understand the behavior of my datasets*; CLAUDIA, *por entenderme (y no sólo por que hablamos el mismo idioma)*; DENIS, pour ton hétérogénéité compositionnelle; DANIEL, pour être mon dernier voisin, et NACHO, *por tu risa contagiosa.*

Nombreuses personnes ont participé moins directement, mais pas pour ça de façon moins importante, à l'avancement de cette thèse. À tous, je vous dis merci, *gracias, eskerrik asko*! À ROCÍO, *por estar siempre ahí, dispuesta a hablar y escuchar*; à VINCENT qui, étant arrivé presque « à la fin », aura laisse sa marque (*Ay hijo! A veces eres un poco perrrrro, eh?*); aux compagnons de soirées, camping et autres folies, MARC, STÉPHANE, MANU, HENRY...; *a mis amigas de siempre*, LUISA, CRIS, ESTI, SARRI, INMA, SARA, TAMARA, PILI, BIDATZ, MARTA, AMAIA, NURIA...; *a* ama, aita *y* URTZI; *a* JOSU ZUDAIRE, *por ser el responsable de todo esto... (sí, sí, haz memoria)*; et aux DOYON, chez qui je me suis sentie comme dans ma propre famille (on est-tu ben!). Finalement, le remerciement le plus spécial est pour la personne qui m'a accompagnée pendant mon doctorat et qui a vécu mes plus hauts et plus bas moments de très proche; il va sans dire qu'il a toujours été à la hauteur.

INTRODUCTION

Le but de cette section est de compléter les introductions inclues dans les chapitres I à V et ne cherche, dans aucun cas, à les remplacer. Notamment, elle se focalise sur les notions de base en phylogénie moléculaire et en évolution des eucaryotes qui ont été omises ou sous-entendues dans les sections suivantes.

Dans la première partie, les principes de l'inférence phylogénétique sont résumés. Nous exposons les différentes méthodes de reconstruction d'arbres avec une emphase particulière sur les méthodes probabilistes et sur les modèles d'évolution de séquences. Nous décrivons également les tests de comparaison de modèles et les moyens pour calculer la robustesse des arbres. Finalement, nous parlons des deux types d'erreurs, stochastique et systématique, qui affectent les analyses phylogénétiques. Ces concepts sont tous importants pour comprendre les analyses phylogénétiques présentées dans les chapitres suivants.

Dans la deuxième partie, nous nous focalisons sur l'évolution des eucaryotes. Premièrement, une vue historique sur la classification des eucaryotes est présentée. Après, nous décrivons chacun des groupes reconnus selon la vision actuelle de l'arbre des eucaryotes, se focalisant sur les Plantae et sur les excavés, les deux ensembles étudiés dans ce projet. On se concentrera sur les apports de la phylogénie pour résoudre certaines des branches de l'arbre des eucaryotes et sur les questions qui restent encore à éclaircir.

1. L'inférence phylogénétique

1.1. Les données

1.1.1. Homologie et analogie

L'inférence des liens de parenté entre les espèces requiert la définition d'un ensemble de caractères comparables à travers tous les organismes étudiés. Ces caractères doivent être homologues, c'est-à-dire hérités d'un même ancêtre (Owen, 1843). Il est important de ne pas confondre homologie et similarité. En effet, il existe des caractères très différents mais homologues et des caractères très similaires qui ne partagent pas d'ancêtre commun (Fitch, 2000). Ces derniers sont dits analogues. Un exemple d'homologie sans similarité est illustré par les membres supérieurs des grenouilles, des oiseaux, des lapins et des lézards qui, malgré leurs différences significatives, partagent le même ancêtre (Figure 1A). Un exemple du cas opposé sont les cactus et les plantes succulentes, tous les deux caractérisés par de grosses branches remplies d'eau et couvertes d'épines (Figure 1B) bien qu'ils ne soient pas groupes frères –les cactus sont plus proches du gazon commun et les plantes succulentes, de l'œillet. La distinction entre homologie et analogie est donc cruciale dans toute étude évolutive.

Les caractères morphologiques ont été largement utilisés pour inférer des arbres phylogénétiques représentant l'évolution des espèces. Maintenant, avec l'accès grandissant aux séquences génomiques, il est possible d'étudier les liens de parenté entre les organismes en comparant leurs séquences protéiques ou nucléiques. Cette approche a de nombreux avantages par rapport à la comparaison de caractères morphologiques : (i) le nombre d'états de caractère est fixe (quatre nucléotides et vingt acides aminés) et ceux-ci peuvent être comparés à travers tous les organismes vivants, que ce soient des animaux, des plantes, des bactéries ou des virus, ce qui est impossible avec les caractères morphologiques; (ii) l'évolution des protéines et des séquences nucléiques suit un patron plus ou moins régulier qui permet l'utilisation de modèles mathématiques pour formaliser leurs changements; et (iii) les génomes de tous les organismes sont composés de très longues séquences nucléiques fournissant une immense quantité d'information, surpassant celle fournie par les caractères morphologiques (Nei et Kumar, 2000).

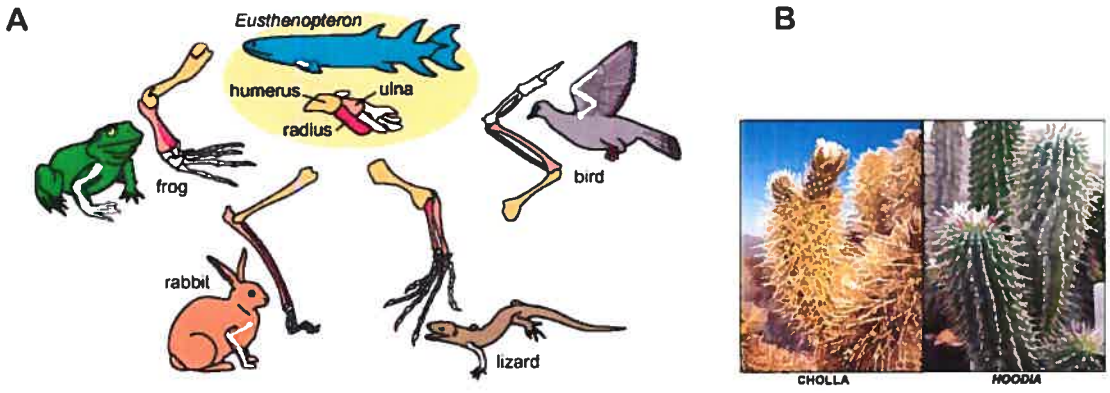

Figure 1 : Homologie et analogie

A : Les membres supérieurs des grenouilles, des oiseaux, des lapins et des lézards sont homologues, car ils dérivent des membres supérieurs de leur ancêtre, l'*Eusthenopteron*. **B** : Les grosses branches remplies d'eau et couvertes d'épines des cactus (p. ex., *Cholla*) et des plantes succulentes (p. ex., *Hoodia*) sont analogues, car elles n'étaient pas présentes chez leur ancêtre commun, mais ont été acquises indépendamment.

Source : http://evolution.berkeley.edu/evolibrary/search/topicbrowse2.php?topic_id=55

Hérité de la biologie évolutive classique, le concept d'homologie s'applique également aux séquences nucléiques et protéiques. Une étape fondamentale et préliminaire à toute analyse phylogénétique utilisant des données moléculaires est l'alignement, qui consiste à définir les positions homologues entre toutes les séquences. De nombreuses méthodes d'alignement de séquences ont été développées (Wallace, Blackshields et Higgins, 2005), mais, malheureusement, l'assignement d'homologie reste parfois impossible. Par exemple, dans les régions très divergentes qui ont subi des insertions et/ou des délétions, il est parfois impossible de déterminer quels sont les nucléotides ou les acides aminés homologues (Figure 2). Dans de tels cas, les régions d'homologie ambiguë doivent être éliminées avant de procéder à l'analyse phylogénétique (Castresana, 2000).

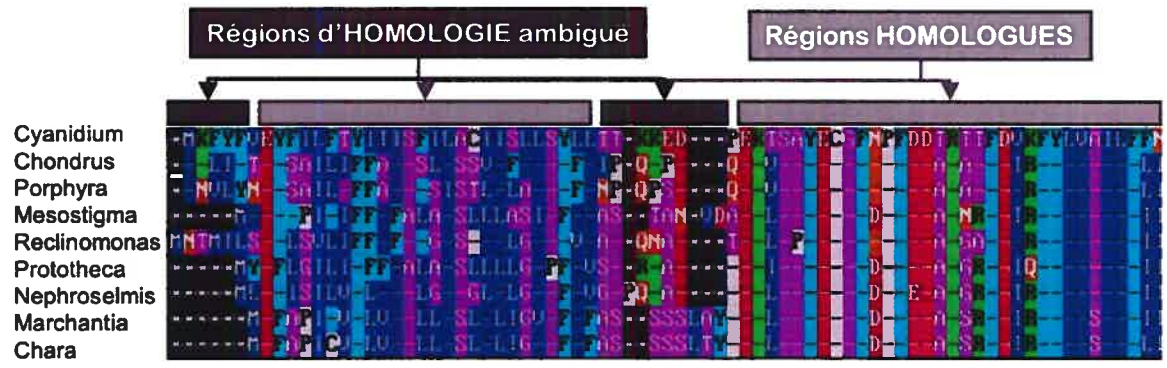

Figure 2 : Homologie dans les alignements

Alignement de la sous-unité 3 de la NADH déshydrogénase mitochondriale illustrant la distinction entre les régions d'homologie certaine et ambiguë. Gris clair: Régions où l'assignation d'homologie est possible pour chaque position; gris foncé : régions où l'assignation d'homologie est pratiquement impossible; on parle ici de régions alignées de façon ambiguë.

1.1.2. Orthologie, paralogie et xénologie

L'orthologie, la paralogie et la xénologie sont trois cas particuliers d'homologie. Il est important de savoir que, quand on s'intéresse à l'évolution des espèces, les gènes (ou protéines[1]) qui sont utilisé(e)s pour l'inférence phylogénétique doivent être orthologues. Par définition, les gènes orthologues et paralogues sont respectivement issus d'événements de spéciation et de duplication (Fitch, 1970), ce qui explique pourquoi seulement les orthologues représentent l'évolution des espèces. Par exemple, considérons un gène qui a subi une ou plusieurs duplications dans l'ancêtre d'un groupe d'espèces de tel manière qu'il existe autant d'arbres phylogénétiques que de copies de ce gène, chaque arbre reflétant l'arbre des espèces (Figure 3A). Supposons maintenant qu'une seule copie est retenue pour chaque individu; alors, la majorité des arbres obtenus seront différents de l'arbre représentant l'évolution des espèces (Figure 3B). Par conséquent, l'identification des gènes orthologues et paralogues est cruciale dans l'interprétation des résultats des analyses phylogénétiques.

[1] Les termes gène et protéine seront souvent utilisés sans distinction à travers le texte.

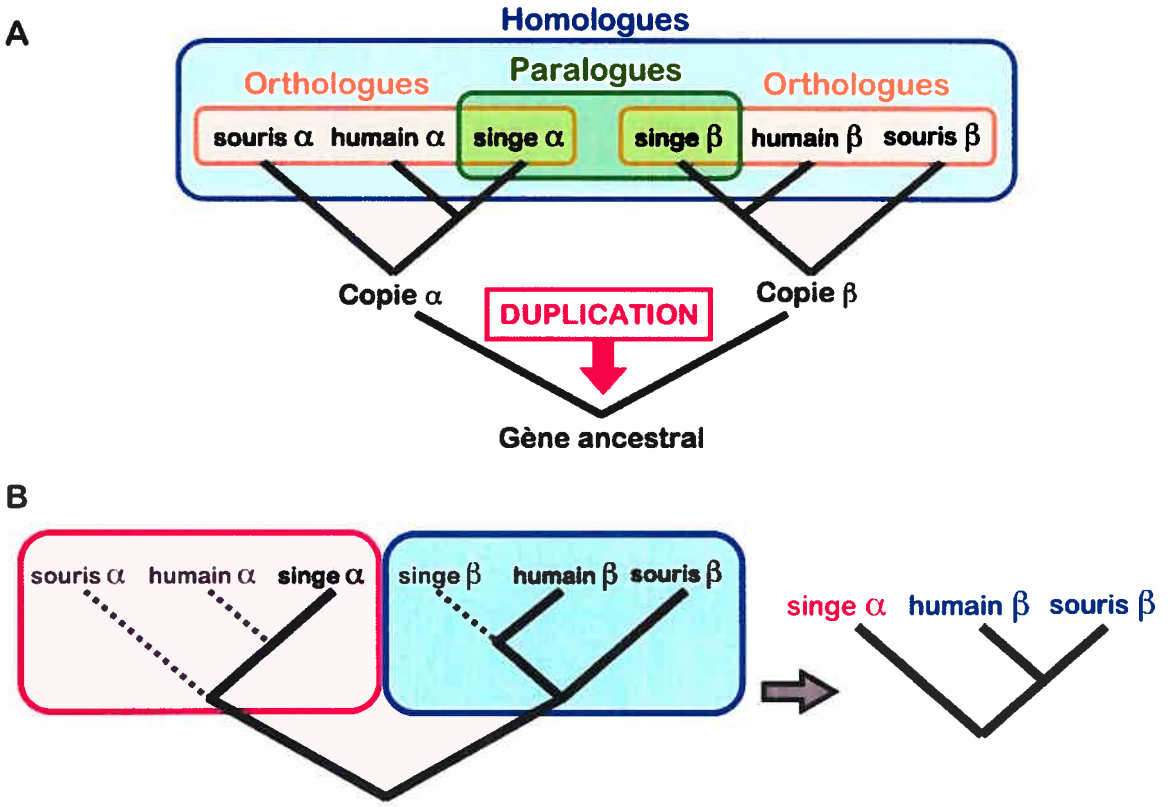

Figure 3 : Orthologie et paralogie

A : Une duplication d'un gène ancestral a donné lieu à deux copies, α et β, et celles-ci ont par la suite subi des spéciations. Tous les gènes α sont orthologues entre eux, alors qu'ils sont paralogues par rapport aux gènes β, et *vice-versa*. Tous les gènes (orthologues et paralogues) représentés dans cet arbre sont homologues entre eux car ils partagent le même ancêtre (nommé ici « gène ancestral »). **B** : Lorsqu'on infère un arbre qui n'inclut que des orthologues (α ou β), la vraie phylogénie des espèces est obtenue; par contre, si par exemple les copies α de la souris et de l'humain et la copie β du singe sont absentes (soit qu'elles ont été perdues ou pas encore séquencées), la phylogénie inférée ne représente pas l'évolution des espèces.

En plus des de gènes paralogues, les de gènes xénologues, c'est-à-dire ceux issus d'un transfert horizontal de matériel génétique entre deux espèces (Gray et Fitch, 1983), peuvent aussi fausser l'inférence de la phylogénie des espèces. D'ailleurs, la xénologie non détectée est un des phénomènes ayant le plus d'impact négatif dans la reconstruction phylogénétique. Tel qu'illustré dans la Figure 4, l'inclusion de gènes acquis par transfert horizontal lors de la reconstruction phylogénétique a comme conséquence le regroupement

des espèces impliquées dans le transfert, alors que celles-ci ne sont pas forcément de proches parents.

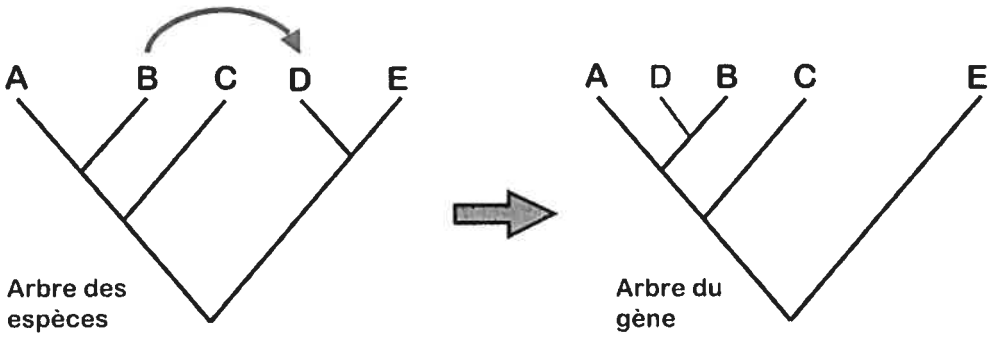

Figure 4 : Xénologie

Considérons un arbre phylogénétique représentant les vrais liens de parenté entre les espèces A, B, C, D et E (à gauche). Si un gène de l'espèce B est transféré à l'espèce D et que ce même gène est utilisé pour l'inférence phylogénétique, l'arbre obtenu placera D comme espèce sœur de B (à droite), alors que D est en vérité plus proche de E que de B.

Plusieurs méthodes existent pour identifier des gènes paralogues et xénologues dans un ensemble de gènes homologues. Les plus classiques impliquent des analyses phylogénétiques, particulièrement des procédures de réconciliation (Page et Charleston, 1997), tandis que d'autres se basent sur la comparaison des séquences (Tatusov et al., 2000). Dans les premières, l'arbre obtenu pour un gène donné est comparé à l'arbre connu des espèces et les deux sont réconciliés selon un nombre minimal de duplications et de pertes de gènes. Ces méthodes ont le désavantage de requérir une connaissance *a priori* de la phylogénie des espèces. Dans les deuxièmes, les séquences de plusieurs génomes sont comparées par BLAST (Altschul et al., 1990) et regroupées selon le principe du BRH (pour *Best Reciprocal Hit*) : une séquence A d'un premier génome a pour meilleur score BLAST une séquence B d'un deuxième génome et *vice-versa*. Ces méthodes supposent que les génomes des espèces étudiées sont complets, qu'il n'y a pas eu de perte de gène et que les séquences les plus proches selon le score BLAST le sont aussi selon la phylogénie, ce qui n'est pas toujours le cas (Koski et Golding, 2001). En résumé, comme il n'existe pas, pour le moment, de méthode fiable pour détecter les orthologues, l'approche recommandée est

d'éviter les gènes à fort taux de duplication et de transfert et de réaliser des analyses phylogénétiques préliminaires pour vérifier la consistance du jeu de données (Philippe, Lartillot et Brinkmann, 2005).

1.2. Les arbres phylogénétiques

1.2.1. Propriétés des arbres phylogénétiques

Le but de toute analyse phylogénétique est d'obtenir, à partir d'un ensemble de données, un arbre qui illustre les liens de parenté entre les unités taxonomiques (UT) d'intérêt, qu'elles soient des espèces, des individus, des gènes, etc. Un arbre phylogénétique est composé de nœuds et de branches, représentant respectivement les unités taxonomiques (UT) et leurs liens de parenté. Généralement, la longueur des branches représente le nombre de changements de caractère entre deux nœuds et un arbre phylogénétique sans longueurs de branches est appelé topologie.

Dans les arbres phylogénétiques, on distingue des groupes monophylétiques, paraphylétiques et polyphylétiques (Figure 5). Un groupe monophylétique (ou clade) est composé de toutes les UTs dérivées d'un ancêtre commun et l'ancêtre commun lui-même; un groupe paraphylétique est formé d'une partie des UTs dérivées d'un ancêtre commun et l'ancêtre commun lui-même; finalement, un groupe polyphylétique est composé de plusieurs UTs excluant leur ancêtre commun.

Un arbre phylogénétique est soit enraciné, ou soit non-enraciné. Un arbre enraciné possède une UT particulière nommée nœud racine, dont descendent toutes les autres UTs, la racine étant leur ancêtre commun. En revanche, un arbre non-enraciné spécifie seulement les liens de parenté entre les UTs sans indiquer dans quelle branche se trouve l'UT ancestrale.

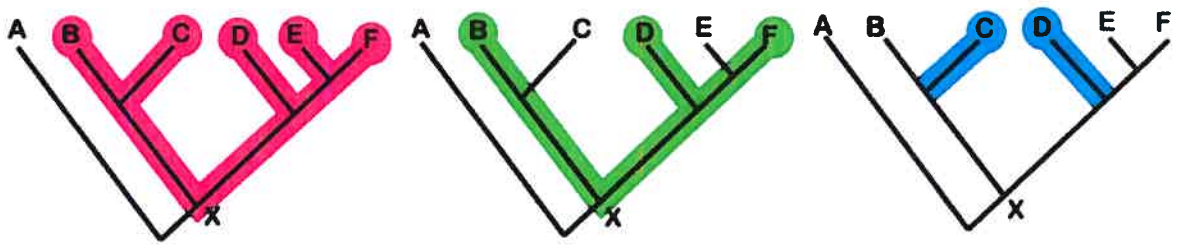

Figure 5 : Monophylie, paraphylie et polyphylie

Prenons un arbre phylogénétique montrant les liens de parenté entre les espèces A-F et choisissons des ensembles d'espèces pour illustrer la monophylie, la paraphylie et la polyphylie. Le groupe composé des espèces B, C, D, E et F (rouge) est monophylétique car il contient toutes les UTs dérivées de l'ancêtre X, ainsi qu'X lui-même. Le groupe composé des espèces B, D et F (vert) est paraphylétique car il inclut leur ancêtre X, mais seulement quelques-uns de ses descendants. Finalement, le groupe composé des espèces C et D (bleu) est polyphylétique car il n'inclut pas leur ancêtre commun.

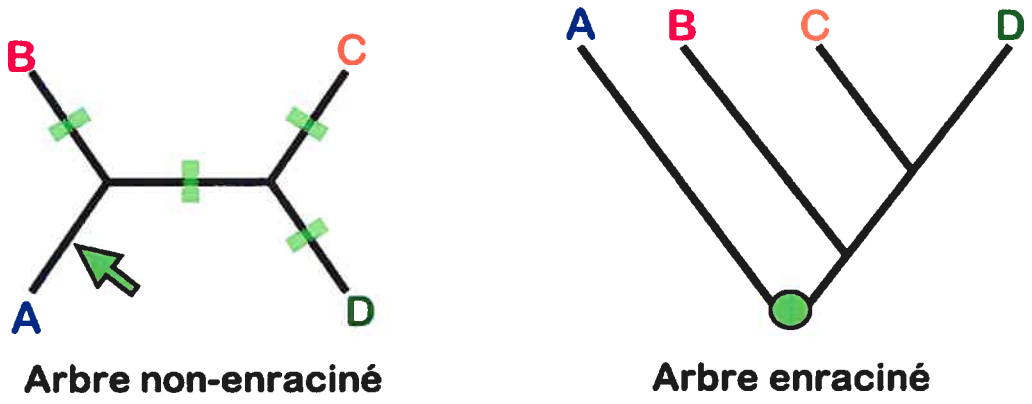

Figure 6 : Arbre non-enraciné et arbre enraciné

Il est possible d'obtenir un arbre enraciné à partir d'un arbre non-enraciné en choisissant de façon arbitraire une branche (flèche verte) et en y insérant le nœud racine (cercle vert). Ici, la branche choisie pour ajouter la racine est celle qui relie A avec le reste de l'arbre, mais il faut souligner qu'il existe quatre autres façons d'enraciner l'arbre (indiquées par les rectangles verts).

Pour trois UTs, il existe un seul arbre non-enraciné et trois arbres enracinés[2]. En généralisant, un arbre non-enraciné de $n \geq 3$ UTs a $2n$-3 branches, chacune d'elles pouvant « recevoir » la racine et, par conséquent, le même nombre d'arbres enracinés est possible. Le nombre d'arbres non-enracinés pour n UTs est $N_N = (2n-5)!/2^{n-3}(n-3)!$ et d'arbres enracinés, $N_R = (2n-3)!/2^{n-2}(n-2)!$ (Felsenstein, 1978b). Pour n UTs, ces nombres augmentent très rapidement en fonction de n (voir Tableau I). Il faut aussi noter que ces formules mathématiques s'appliquent aux arbres sans multifurcations, c'est-à-dire, que chaque nœud interne a exactement deux descendants. Si on permet les multifurcations, le nombre d'arbres possibles est encore plus grand.

Tableau I : Nombre d'arbres non-enracinés et enracinés possibles pour *n* UTs

Nombre de UTs (*n*)[3]	Nombre d'arbres non-enracinés	Nombre d'arbres enracinés
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425
11	34 459 425	654 729 075
12	654 729 075	13 749 310 575
13	13 749 310 575	316 234 143 225
14	316 234 143 225	7 905 853 580 625
15	7 905 853 580 625	213 458 046 676 875
20	2×10^{20}	8×10^{21}
30	7×10^{36}	5×10^{38}
40	1×10^{55}	1×10^{57}
50	3×10^{74}	3×10^{76}

[2] Ceci n'est pas tout à fait exact, car un arbre est aussi défini par ses longueurs de branches. Comme le nombre de combinaisons de longueurs de branches qu'un arbre peut adopter est infini, il y a un nombre infini d'arbres pour chaque ensemble d'espèces. Le terme correct dans cette section est donc topologie, mais pour être consistant avec la littérature, on utilise arbre, terme plus couramment utilisé.

[3] Les chiffres à partir de 20 UTs sont approximés.

1.2.2. Recherche du meilleur arbre phylogénétique

La plupart des méthodes de reconstruction phylogénétique consistent à identifier les meilleurs arbres selon un critère d'optimalité (par exemple, la vraisemblance ou le score de parcimonie, qui seront expliqués dans la section 1.3) parmi un ensemble de solutions possibles. Malheureusement, le nombre d'arbres est généralement trop grand (voir Tableau I) pour qu'une exploration complète et exhaustive de tout l'ensemble soit réalisable. Des recherches dites exactes ou heuristiques explorant partiellement l'ensemble d'arbres ont donc été proposées. Seules les premières assurent l'identification du meilleur arbre.

1.2.2.1. Méthode exacte : l'algorithme de « branch and bound »

L'algorithme de séparation et d'évaluation progressive (ou *branch and bound*[4]) (Land et Doig, 1960) est une méthode exacte pour l'identification d'un arbre optimal sans une exploration complète de l'ensemble des solutions possibles (Hendy et Penny, 1982). Tel que ce sera expliqué dans la section 1.3.1, le score de parcimonie d'un arbre ne peut pas diminuer avec l'addition d'UTs, une condition nécessaire à l'application du *branch and bound*. L'algorithme débute avec l'unique arbre de trois UTs et consiste à explorer implicitement l'ensemble des solutions selon un parcours en profondeur et l'ajout d'une espèce à l'arbre courant. Dans une recherche exhaustive, toutes les solutions sont énumérées explicitement, tandis que l'exploration implicite de tous les arbres se fait en comparant la valeur de la meilleure solution jusqu'à maintenant avec celle de l'arbre courant (même s'il est incomplet). Dans l'exemple de la Figure 7, on commence par visiter l'une des voies d'exploration jusqu'à la fin (arbres 1-10) et le score du meilleur arbre devient le score de référence (241). Ensuite, l'exploration d'une autre voie continue tant que le score est inférieur au score de référence, car un score ne peut qu'augmenter ou rester égal avec le rajout d'UTs. Si un arbre complet a un score inférieur au score de référence, il devient le meilleur arbre et le score de référence est mis à jour. On procède de cette façon jusqu'à ce que toutes les voies partant de l'arbre initial aient été implicitement explorées. Grâce à cet algorithme, le nombre d'arbres évalués afin de trouver le meilleur est

[4] Les noms anglais des algorithmes seront souvent utilisés car ils sont plus connus que ceux en français.

souvent drastiquement réduit, mais il reste encore immense pour de grands jeux de données. L'algorithme de *branch and bound* peut généralement être utilisé pour des jeux de données de moins de 15 espèces avec un temps de calcul raisonnable.

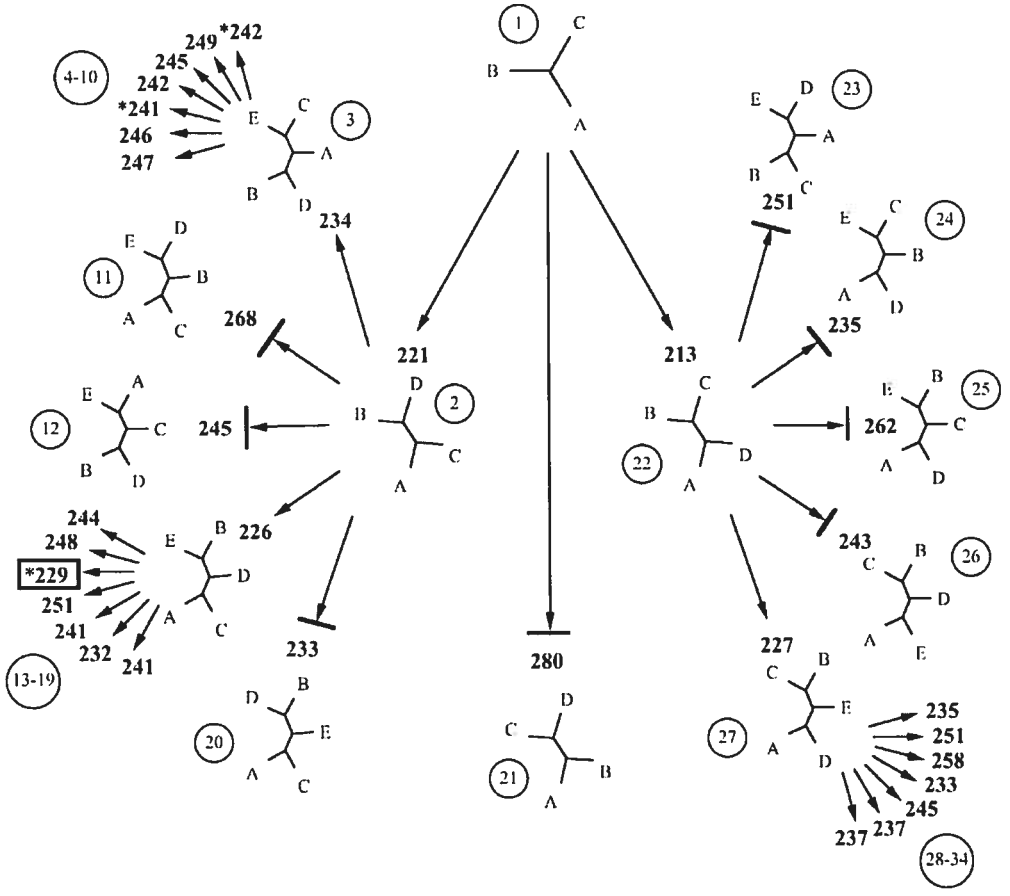

Figure 7 : Exemple d'application de l'algorithme de *branch and bound*

Le but est de trouver un arbre optimal selon la parcimonie parmi les 945 possibilités pour six UTs sans devoir toutes les explorer. Dans ce schéma, les numéros encerclés représentent l'ordre selon lequel l'algorithme a visité les solutions et ceux en gras, le score correspondant; les astérisques indiquent les arbres dont la valeur est meilleure que celle actuellement connue et le rectangle indique le score de la solution optimale. Après l'exploration implicite de toutes les solutions, le meilleur arbre est celui qui a un score de 229. Des 945 arbres possibles avec 6 UTs, l'exploration de 21 d'entre eux (plus 13 autres dans les étapes intermédiaires) a été suffisante pour trouver l'arbre optimal.

Source : http://workshop.molecularevolution.org/people/faculty/swofford_david.php

1.2.2.2. Les méthodes heuristiques

La plupart des recherches heuristiques se basent sur le même principe (Maddison et Maddison, 1992; Swofford et Begle, 1993): des réarrangements sont appliqués sur un arbre initial et, parmi les arbres obtenus, celui avec le meilleur score est retenu; ce processus est répété jusqu'à ce que (i) aucun des réarrangements n'améliore le score ou (ii) le nombre maximal de réarrangements imposé par l'utilisateur est atteint. L'arbre obtenu à la fin du processus est dit optimum local car il n'y a aucune garantie qu'il soit l'optimum global. Comme illustré dans la Figure 8, l'arbre de départ d'une recherche heuristique affecte ses chances de trouver l'optimum global. Ces types de stratégies sont dites gloutonnes car elles ne considèrent que le voisinage immédiat de l'arbre courant pour faire un choix

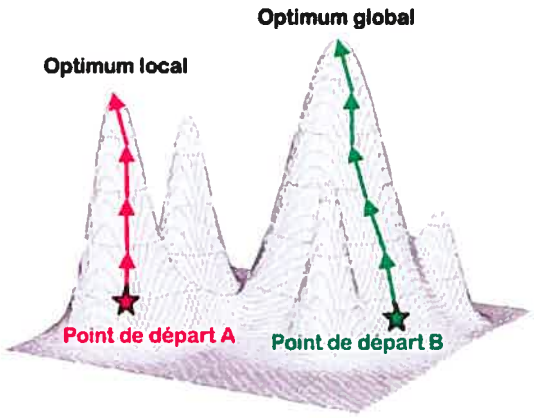

Figure 8 : Recherche heuristique du meilleur arbre

Dans ce diagramme, les arbres avec les meilleurs scores sont représentés par les points les plus hauts, les étoiles indiquent deux points de départ de deux recherches indépendantes et les flèches, des réarrangements avec succès, c'est-à-dire, qui ont donné lieu à de meilleurs arbres. Une telle recherche débute avec un arbre donné (étoile rouge ou verte) et l'exploration de ses voisins; parmi eux, celui qui a le meilleur score est choisi et le processus est répété jusqu'à ce qu'on ne puisse plus améliorer le score de l'arbre courant. Ici, la recherche commencée à l'étoile verte permet de trouver l'optimum global, alors que celle débutée par l'étoile rouge reste coincée dans un optimum local.

Il existe plusieurs méthodes de réarrangement d'arbres, les plus connues étant l'échange entre les plus proches voisins (*nearest-neighbor interchange*; NNI), l'élagage et le greffage de sous arbres (*subtree pruning and regrafting*; SPR) et le coupage et la reconnexion d'arbres (*tree bisection and reconnection*; TBR). Une représentation schématique du fonctionnement de chacune de ces méthodes est montrée dans la Figure 9 et une brève description est donnée ci-dessous.

Le **NNI** consiste à échanger deux branches adjacentes dans l'arbre. Plus précisément, une branche interne et les deux branches connectées à chacune de ses extrémités sont effacées (soient cinq branches effacées) et les quatre sous arbres résultant sont reconnectés des deux autres façons possibles. Dans un arbre non-enraciné et sans multifurcation avec n UTs, il y a n-3 branches internes, chacune ayant 2 arbres voisins à examiner, et donc $2(n-3)$ réarrangements NNI possibles. À titre d'exemple, un arbre avec 20 UTs a 34 voisins selon cette méthode.

Le **SPR** (Swofford *et al.*, 1996) consiste à effacer une branche (interne ou externe) de l'arbre et à coller le sous-arbre résultant dans toutes les autres branches. Tel que démontré par Allen et Steel (2001), le nombre d'arbres différents obtenus par des réarrangements SPR sur un arbre de n UTs est $2(n-3)(2n-7)$. Avec cette méthode, il y a 1 122 arbres voisins d'un arbre avec 20 UTs. Parce que le NNI est un cas particulier du SPR, les 1 122 arbres incluent les 34 arbres voisins obtenus par le NNI.

Le **TBR** (Swofford *et al.*, 1996) consiste à couper une branche interne et à connecter les deux arbres résultants de toutes les façons possibles. Ici, le nombre d'arbres voisins dépends de la topologie originale, mais il y a au plus $(2n-3)(n-3)^2$ réarrangements possibles peu importe l'arbre (Allen et Steel, 2001). Par exemple, pour un arbre avec 20 UTs, il y a un maximum de 10 693 arbres voisins. Les réarrangements NNI et SPR sont des cas particuliers du TBR et sont considérés dans ce dénombrement.

Parmi les trois, la méthode NNI est celle qui explore le plus petit nombre d'arbres et qui a le moins de chance de trouver l'arbre optimal. À l'opposé, la méthode TBR explore le plus grand nombre de voisins et elle a donc plus de chance de trouver le meilleur arbre. Évidemment, le choix de la méthode à utiliser dépend de la taille du jeu de données, de la puissance de calcul disponible et du temps alloué pour les analyses.

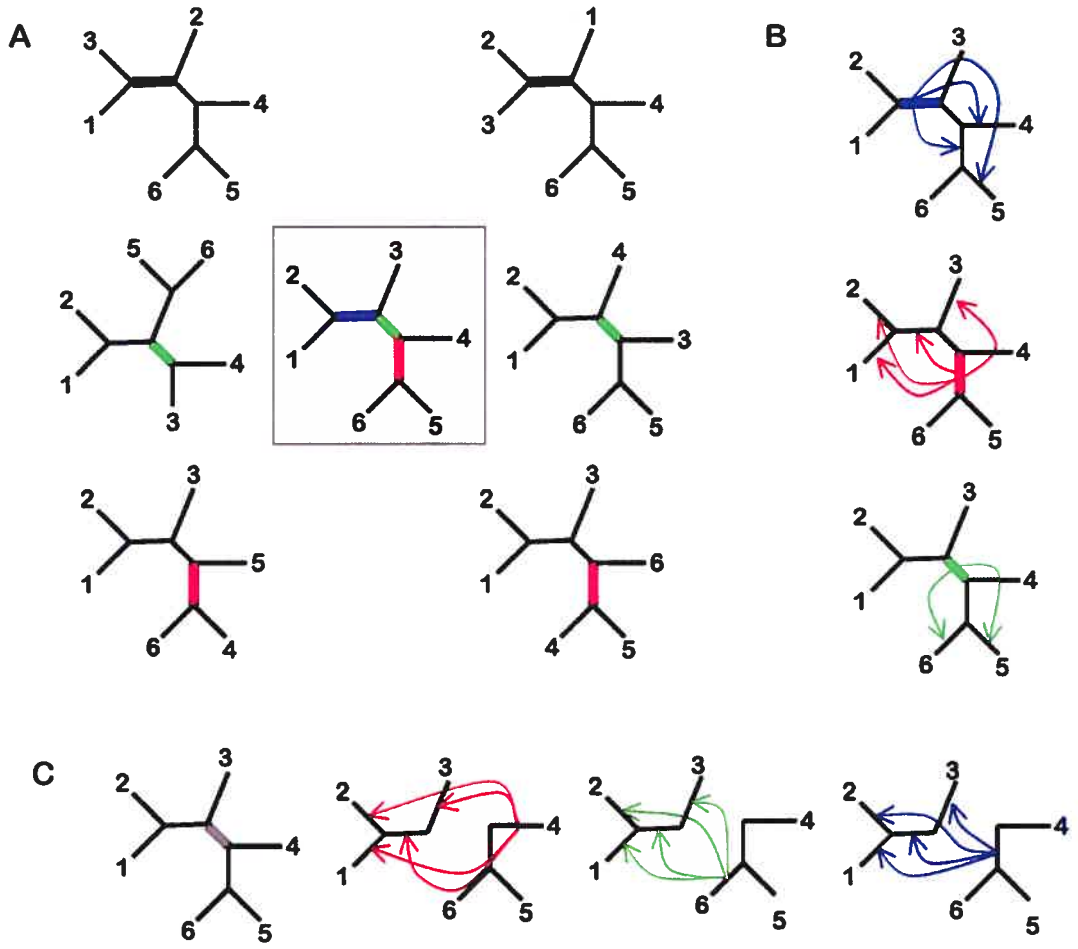

Figure 9 : Exemples de méthodes de réarrangement d'arbres

A : Le NNI (*nearest-neighbor interchange*) consiste à effacer une branche interne et à reconnecter les quatre sous-arbres résultants selon les deux autres façons possibles; dans la figure, tous les NNIs possibles pour un arbre avec six UTs (encadré) sont montrés. **B** : Le SPR (*subtree pruning and regrafting*) consiste à effacer une branche et à connecter le sous-arbre résultant dans toutes les autres branches possibles. La figure montre des exemples de réarrangements possibles en coupant les trois branches internes d'un arbre avec six UTs, mais il faut noter que les SPRs incluent aussi les réarrangements impliquant les branches externes. **C** : Le TBR (*tree bisection and reconnection*) consiste à couper l'arbre en deux et à reconnecter les deux sous-arbres résultant de toutes les façons possibles, c'est-à-dire, toutes les branches du premier sous arbre avec toutes les branches du deuxième; dans la figure, on montre des exemples de réarrangements TBR possibles dans un arbre de six espèces.

Les réarrangements NNI, SPR et TBR sont appliqués successivement sur un arbre initial généré aléatoirement ou par l'application soit d'un algorithme glouton d'addition par étapes (*greedy stepwise addition*), soit par la méthode de décomposition d'un arbre en étoile (*star decomposition*) (Figure 10). Le premier algorithme est similaire à la méthode de *branch and bound*, à l'exception qu'à chaque fois qu'une UT est ajoutée, seulement la solution avec le meilleur score est retenue pour la prochaine étape (Figure 10A). Quant au deuxième algorithme, il consiste à résoudre graduellement un arbre en groupant deux UTs à la fois et en évaluant le score des arbres obtenus à chaque étape (Figure 10B). Cette méthode fut l'une des premières à être utilisée, alors qu'il est maintenant plus commun d'utiliser l'algorithme d'addition par étapes pour obtenir un arbre de départ sur lequel les réarrangements NNI, SPR ou TBR seront appliqués.

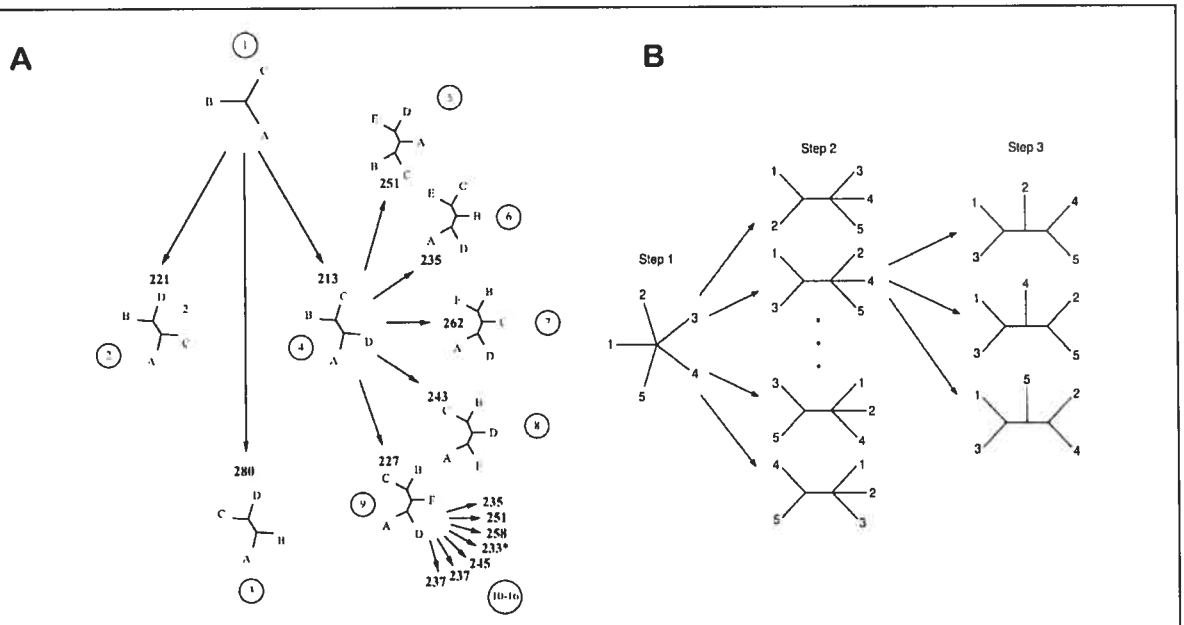

Figure 10 : Méthodes heuristiques non basées sur des réarrangements

A : L'algorithme glouton d'addition par étapes ajoute les UTs une à une dans toutes les branches possibles, calcule le score de chacun des arbres obtenus et seulement celui avec le meilleur score est retenu pour l'étape suivante. Il faut noter que cet exemple est le même que celui utilisé pour expliquer le *branch and bound* et qu'ici le meilleur arbre n'a pas été trouvé. **B :** L'algorithme de décomposition d'un arbre en étoile résout graduellement un arbre en groupant deux UTs à la fois, et en retenant l'arbre avec le meilleur score à chaque étape.

Source : http://workshop.molecularevolution.org/people/faculty/swofford_david.php

1.3. Les méthodes d'inférence phylogénétique

Il existe trois groupes de méthodes d'inférence phylogénétique : les méthodes de parcimonie, les méthodes probabilistes (vraisemblance maximale et inférence bayésienne) et les méthodes de distance. Les deux premiers évaluent un ensemble d'arbres (obtenu par une recherche exhaustive, exacte ou heuristique) et identifient celui qui explique le mieux les données observées. En revanche, les méthodes de distance construisent un arbre en se basant sur une matrice de distances générée à partir de séquences nucléiques ou protéiques. Dans les prochaines sections, les méthodes de parcimonie et probabilistes ainsi que leurs principales différences seront expliquées. Les méthodes de distance ne seront pas décrites ici car elles utilisent les mêmes modèles que les méthodes probabilistes, mais sans prendre en considération toute l'information provenant des séquences (voir Felsenstein, 2004 pour une description approfondie sur ce type de méthodes).

1.3.1. La méthode de parcimonie maximale

La parcimonie fut premièrement développée pour l'analyse de caractères morphologiques (Henning, 1966). Eck et Dayhoff (1966) furent les premiers à l'utiliser pour construire des arbres à partir des séquences protéiques et, plus tard, Fitch (1971b) et Hartigan (1973) développèrent des algorithmes plus rigoureux pour l'analyse de séquences nucléiques. La parcimonie implique l'identification d'un arbre requérant le nombre minimal de changements évolutifs pour expliquer les différences observées entre les UTs étudiées. Dans un contexte de données moléculaires, cette méthode infère, selon une topologie donnée, les états ancestraux de chacune des positions en minimisant le nombre de substitutions requises pour expliquer le processus évolutif de l'alignement. L'idée générale de cette méthode fut donnée par Edwards et Cavalli-Sforza (1963) dans leur déclaration selon laquelle le meilleur arbre phylogénétique est celui qui implique « la plus petite quantité d'évolution » et se base sur l'affirmation du moine franciscain et philosophe Guillaume d'Ockham (XIVème siècle), qui disait que la meilleure hypothèse pour expliquer un processus est celle ayant besoin du plus petit nombre de présomptions[5].

[5] La citation originale en latin, *Pluralitas non est ponenda sine necessitate*, se traduit littéralement en français par « La pluralité ne doit pas être proposée sans nécessité ».

1.3.1.1. Estimation du plus petit nombre de changements

Nous illustrons ci-dessous le principe de parcimonie à l'aide d'un exemple. Considérons un alignement de séquences nucléiques avec quatre espèces (Tableau II), qui peuvent être regroupées selon trois arbres non-enracinés possibles (Figure 11) et calculons le nombre de changements nécessaire dans chacun de ces arbres afin de choisir le plus parcimonieux, soit celui qui requiert le plus petit nombre de changements pour expliquer les données observées.

Tableau II : Alignement de quatre séquences hypothétiques

	Position						
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**
Espèce 1	A	A	G	A	G	T	A
Espèce 2	A	G	C	C	G	T	T
Espèce 3	A	G	A	T	A	C	A
Espèce 4	A	G	A	G	A	C	T

Le calcul doit être fait pour chaque position. Commençons par la position III et inférons le scénario d'évolution impliquant le nombre minimal de substitutions selon la topologie X (Figure 12). Il y a trois façons également parcimonieuses pour expliquer l'évolution de la position III selon cette topologie. Comme l'état ancestral de ce caractère est inconnu, il est impossible de savoir lequel des trois scénarios a eu lieu. Cependant, pour choisir l'arbre le plus parcimonieux, seul le nombre minimum de changements requis pour expliquer l'évolution de la position III dans la topologie X est important.

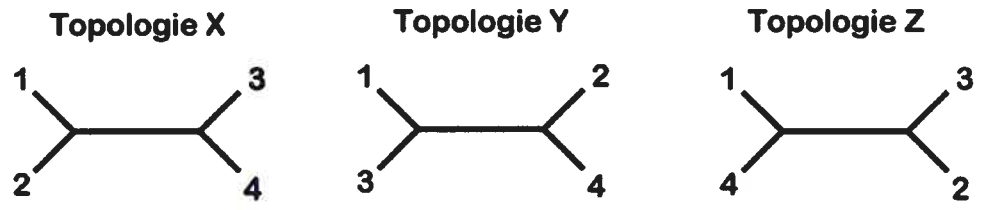

Figure 11 : Trois topologies possibles pour quatre UTs
Il existe trois arbres non-enracinés pour quatre UTs.

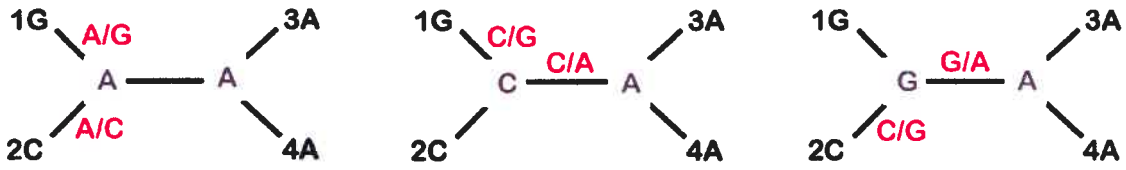

Figure 12 : Reconstruction de l'évolution de la position III dans la topologie X
Il existe trois façons également parcimonieuses de cartographier l'évolution de la position III dans la topologie X de la Figure 11. Les états observés pour chacune des espèces et ceux inférés pour les nœuds ancestraux sont représentés en noir et en gris, et les changements inférés dans chacun des scénarios possibles sont donnés en rouge.

Le nombre de changements requis pour chacune des positions selon chacune des trois topologies est calculé de la même façon. La Figure 13 illustre le nombre minimal de changements inférés pour chacune des positions de l'alignement du Tableau II dans chacune des topologies possibles (X, Y et Z). Tel qu'illustré pour la position III (Figure 12), il existe plusieurs façons de placer les changements dans les arbres sans que leur nombre change, mais seulement un exemple pour chaque position est donné.

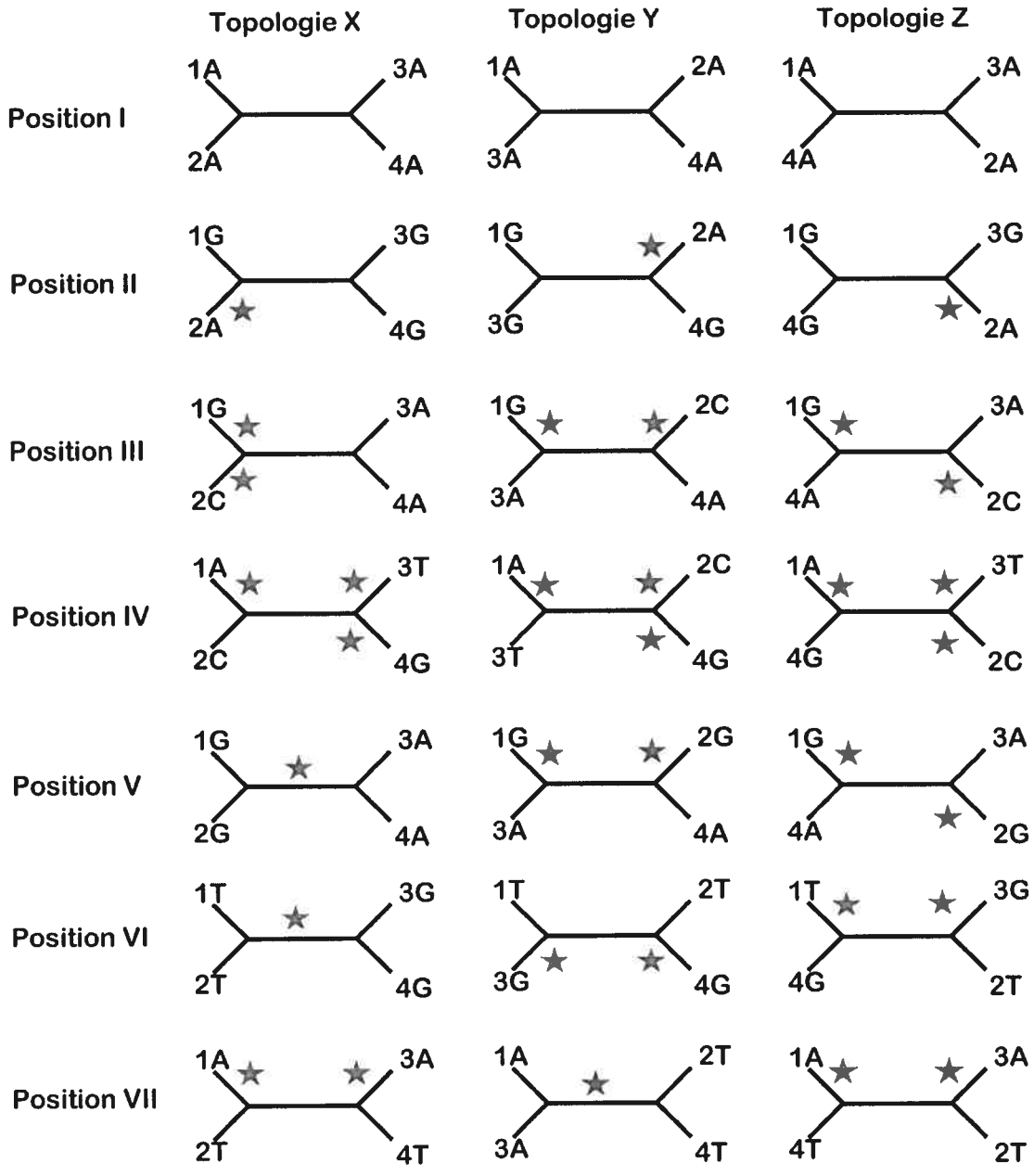

Figure 13 : Reconstruction de l'évolution des positions I-VII

Les étoiles indiquent les endroits où les substitutions ont pu avoir lieu. Il faut noter que les changements peuvent avoir eu lieu dans des branches différentes de celles où l'étoile a été placée.

Une fois que le nombre minimal de changements pour chacune des positions selon chaque topologie est connu, le nombre total de changements requis selon chaque topologie est calculé (Tableau III). L'arbre le plus parcimonieux est celui qui a la plus petite longueur, c'est-à-dire, celui qui requiert le plus petit nombre de changements pour expliquer l'alignement (X dans notre exemple). Notons qu'il est possible que plusieurs arbres aient la même longueur; dans un tel cas, on ne peut pas déterminer lequel est le meilleur et ils sont considérés arbres également parcimonieux.

Tableau III : Nombre de changements par position et total pour les topologies X, Y et Z

	Position							
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**TOTAL**
Topologie X	0	1	2	3	1	1	2	10
Topologie Y	0	1	2	3	2	2	1	11
Topologie Z	0	1	2	3	2	2	2	12

Le nombre de changements nécessaires par arbre a été très facile à calculer pour un alignement de quatre espèces et de sept positions. Par contre, quand le nombre d'UTs et/ou de positions est plus grand, il est nécessaire d'utiliser des algorithmes développés spécialement pour ces types de calculs (p. ex., (Fitch, 1971b) et (Sankoff, 1975)). À ce jour, il existe de nombreux programmes qui implémentent la méthode de parcimonie pour la reconstruction phylogénétique, les plus connus étant PAUP* (Swofford, 2002) et PHYLIP (Felsenstein, 2001).

1.3.1.2. Positions informatives et non informatives

Les positions invariables (p. ex., la position I) ne nous aident pas à discriminer entre deux topologies et seulement les positions variables sont considérées en parcimonie. Par contre, certaines positions variables peuvent ne pas être utiles pour trouver l'arbre le plus parcimonieux. En effet, une position dans laquelle aucun nucléotide n'apparaît plus d'une fois (p. ex., la position IV) est non informative et peut toujours être expliquée par le même

nombre de substitutions pour n'importe lequel des arbres possibles. Pour qu'une position soit utile pour trouver l'arbre le plus parcimonieux, elle doit avoir au moins deux caractères avec le même état (p. ex., positions V, VI et VII). Ce type de positions sont appelées informatives (Fitch, 1977). Il faut noter que les positions non informatives pour la parcimonie (incluant les positions invariables) peuvent l'être pour les autres méthodes de reconstruction phylogénétique. C'est pourquoi, à la place de la terminologie proposée par Fitch, « positions non informatives », il est préférable d'utiliser « positions non informatives pour la parcimonie ». Pour déterminer l'arbre le plus parcimonieux, il suffit de considérer seulement les positions informatives pour la parcimonie; par contre, les positions variables mais non-informatives peuvent être utilisées pour calculer la longueur des branches, c'est-à-dire, le nombre de substitutions inférées pour une branche donnée.

1.3.1.3. Inconsistance et parcimonie

La méthode de parcimonie ne fait aucune supposition explicite, excluant celle qui dicte que le meilleur arbre est celui qui requiert le plus petit nombre de substitutions. Notamment, un arbre qui minimise le nombre de substitutions, minimise aussi le nombre de réversions ou de convergences (homoplasie). Quand le degré de divergence entre les séquences comparées est petit, c'est-à-dire que l'homoplasie est rare, la parcimonie fonctionne correctement. Par contre, si les séquences ont divergé au point que l'homoplasie est commune, la parcimonie est dite inconsistante, c'est-à-dire qu'elle convergera vers une solution erronée lorsque la quantité de données tend vers l'infini.

L'homoplasie est plus probable dans les séquences à fort taux d'évolution, ce qui a pour conséquence le regroupement artificiel des espèces correspondantes (Figure 14), un phénomène appelé attraction des longues branches (Felsenstein, 1978a; Hendy et Penny, 1989). En 1978, Felsenstein présenta les conditions sous lesquelles la parcimonie est inconsistante pour quatre espèces et pour des caractères à deux états. Si $p^2 < q(1 - q)$, p et q étant les longueurs des branches tel qu'illustré dans la Figure 15, la parcimonie trouvera le bon arbre si une quantité suffisante de caractères est utilisée. Dans le cas contraire, l'artefact d'attraction des longues branches dominera et un arbre incorrect sera inféré. La région où les valeurs de p et de q causent l'inconsistance de la parcimonie est appelée « zone de Felsenstein » (Huelsenbeck et Hillis, 1993).

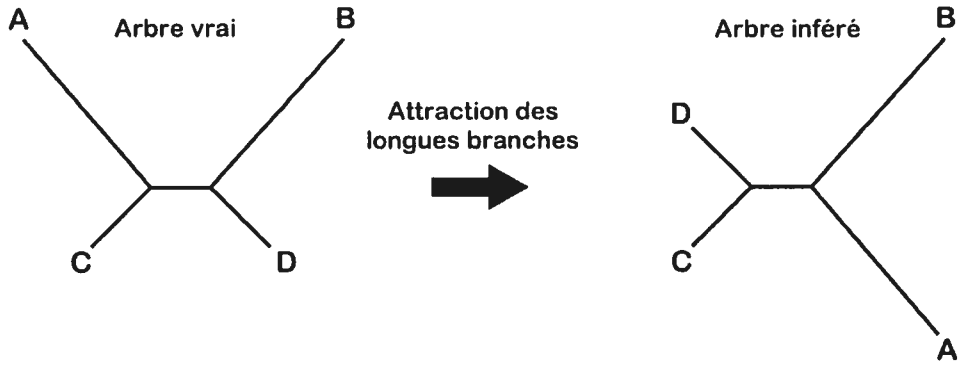

Figure 14 : Attraction des longues branches

Selon le phénomène d'attraction des longues branches, deux longues branches qui ne sont pas proches parentes (arbre vrai) peuvent se trouver ensemble de façon artificielle sous certaines conditions (arbre inféré).

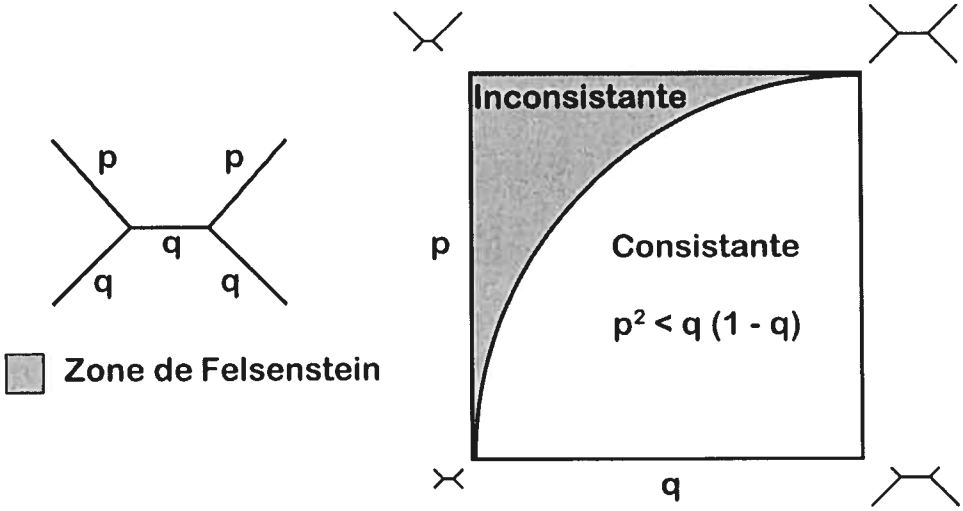

Figure 15 : Conditions dans lesquelles la parcimonie est consistante ou inconsistante

p et *q* indiquent la probabilité de changement le long d'une branche. Les valeurs de *p* et *q* se situent entre 0 et 0,5 car même une branche de longueur infinie a une probabilité de 0,5 que l'état au début de la branche soit différent de celui à la fin. La zone de Felsenstein indique les combinaisons de valeurs de *p* et de *q* pour lesquelles la parcimonie est inconsistante.

Modifié d'après Felsenstein (2004)

1.3.2. Les méthodes probabilistes

Contrairement à la parcimonie, qui présume implicitement que les différences observées entre les UTs doivent s'expliquer par un nombre minimal de changements, les méthodes probabilistes n'ont pas de suppositions implicites. En effet, leur robustesse réside dans l'incorporation explicite d'un modèle d'évolution de séquences dont les paramètres peuvent être estimés au cours de l'analyse phylogénétique. D'autre part, ces méthodes permettent l'application de tests statistiques pour évaluer différentes hypothèses évolutives (Huelsenbeck et Crandall, 1997). Leurs propriétés statistiques et la possibilité de prendre en compte l'histoire évolutive de manière explicite sont à l'origine de la popularité de ces méthodes. Deux méthodes probabilistes appliquées à la reconstruction phylogénétique ont été développées : la méthode de vraisemblance maximale et l'inférence bayésienne. Il existe plusieurs similarités entre les deux, entre autres, elles sont basées sur la vraisemblance et elles utilisent les mêmes modèles d'évolution. Par contre, tel que ce sera expliqué ci-dessous, elles utilisent les probabilités de façon différente.

1.3.2.1. La méthode de vraisemblance maximale

La méthode de vraisemblance maximale fut inventée par le généticien statisticien Sir R. A. Fisher entre 1912 et 1922 (Fisher, 1912; Fisher, 1921; Fisher, 1922). En 1964, Edwards et Cavalli-Sforza (1964) appliquèrent cette méthode à la phylogénie, mais la première application aux séquences moléculaires ne fut réalisée qu'au début des années 70 par le fameux statisticien Jerzy Neyman (1971). En 1981, Felsenstein décrivit l'algorithme de *prunning*, une méthode efficace pour le calcul de la vraisemblance à partir de séquences nucléiques d'un nombre modéré d'UTs. À ce jour, la méthode de vraisemblance maximale est implémentée dans plusieurs programmes d'inférence phylogénétique, les plus connus étant PROTML (Adachi et Hasegawa, 1996b), PAML (Yang, 1997), PHYLIP (Felsenstein, 2001), PAUP (Swofford, 2002), TREE-PUZZLE (Schmidt et al., 2002), PhyML (Guindon et Gascuel, 2003) et TreeFinder (Jobb, von Haeseler et Strimmer, 2004).

1.3.2.1.1. La fonction de vraisemblance

La vraisemblance en phylogénie moléculaire (L) est la probabilité d'observer les données D, soient des séquences nucléiques ou d'acides aminés, étant donné un modèle d'évolution (M) de paramètres $\theta = (\theta_1, \theta_2, ..., \theta_k)$ et une topologie τ de longueurs de branches $\upsilon = (\upsilon_1, \upsilon_2, ..., \upsilon_s)$:

$$L = \Pr(D \mid M, \theta, \tau, \upsilon).$$

L'objectif de la méthode de vraisemblance maximale est de trouver l'arbre avec la plus grande valeur de L. Soit un arbre phylogénétique avec des longueurs de branches données et un alignement de 5 séquences nucléiques avec n positions

$$D = \left[D_{ij}\right] = \begin{bmatrix} AAAGCACTCTAA.....N \\ CAGCTAGTCTAA.....N \\ CAAGTAGGTTTG.....N \\ CAGATGAGCTGA.....N \\ GGTTTGAAGTAA.....N \end{bmatrix},$$

supposant que l'évolution de chaque position est indépendante de l'évolution du reste des positions[6], la probabilité $L = \Pr(D \mid M, \theta, \tau, \upsilon)$ est égale au produit des probabilités d'observer chaque position $D^{[i]}$ étant donné les mêmes hypothèses, soient le modèle M, la topologie τ et l'ensemble de longueurs de branches υ :

$$L = \Pr(D \mid M, \theta, \tau, \upsilon) = \prod_{i=1}^{n} \Pr(D^{[i]} \mid M, \theta, \tau, \upsilon).$$

Pour éviter de travailler avec des valeurs très petites (les probabilités prennent des valeurs entre 0 et 1), on utilise plutôt le logarithme de la vraisemblance, qui donne des valeurs plus facilement manipulables d'un point de vue informatique. Le logarithme de la vraisemblance de l'alignement de séquences D devient alors la somme des logarithmes de vraisemblance de chaque position $D^{[i]}$:

[6] Voir section 1.4.4 pour une discussion à ce sujet

$$\ln L = \ln(\Pr(D \mid M,\theta,\tau,\upsilon)) = \sum_{i=1}^{n} \Pr(D^{[i]} \mid M,\theta,\tau,\upsilon)$$

1.3.2.1.2. Calcul de la vraisemblance d'une position

Pour calculer la vraisemblance totale d'un alignement, il suffit de considérer celle de chaque position, qui est égale à la somme des vraisemblances de chacun des scénarios d'évolution possibles. Prenant comme exemple l'arbre de la Figure 16, la vraisemblance de cette position pour l'arbre donné est :

$$L^{[i]} = \sum_{x}\sum_{y}\sum_{z}\sum_{w} \Pr(x)\Pr(y \mid x,\upsilon_6,\theta)\Pr(A \mid y,\upsilon_1,\theta)\Pr(C \mid y,\upsilon_2,\theta)\Pr(z \mid x,\upsilon_8,\theta)$$
$$\Pr(C \mid z,\upsilon_3,\theta)\Pr(w \mid z,\upsilon_7,\theta)\Pr(C \mid w,\upsilon_4,\theta)\Pr(G \mid w,\upsilon_5,\theta),$$

où chaque sommation doit se faire pour chacun des quatre nucléotides, A, T, G et C. Le calcul de ces sommations avec un grand nombre de séquences est difficile, mais peut être réalisé grâce à l'algorithme de *prunning* (Felsenstein, 1981).

La probabilité d'observer un état de caractère dans une position donnée dépend donc de la topologie de l'arbre τ, de ses longueurs de branches $\upsilon = (\upsilon_1, \upsilon_2, \ldots, \upsilon_s)$ et d'un modèle M. Ce dernier décrit l'évolution des séquences et définit la probabilité de changement d'un état i à un état j le long d'une branche de longueur υ selon les paramètres $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Généralement, les longueurs de branche υ et les paramètres du modèle θ ne sont pas connus *a priori* et sont optimisés pour maximiser la fonction de vraisemblance.

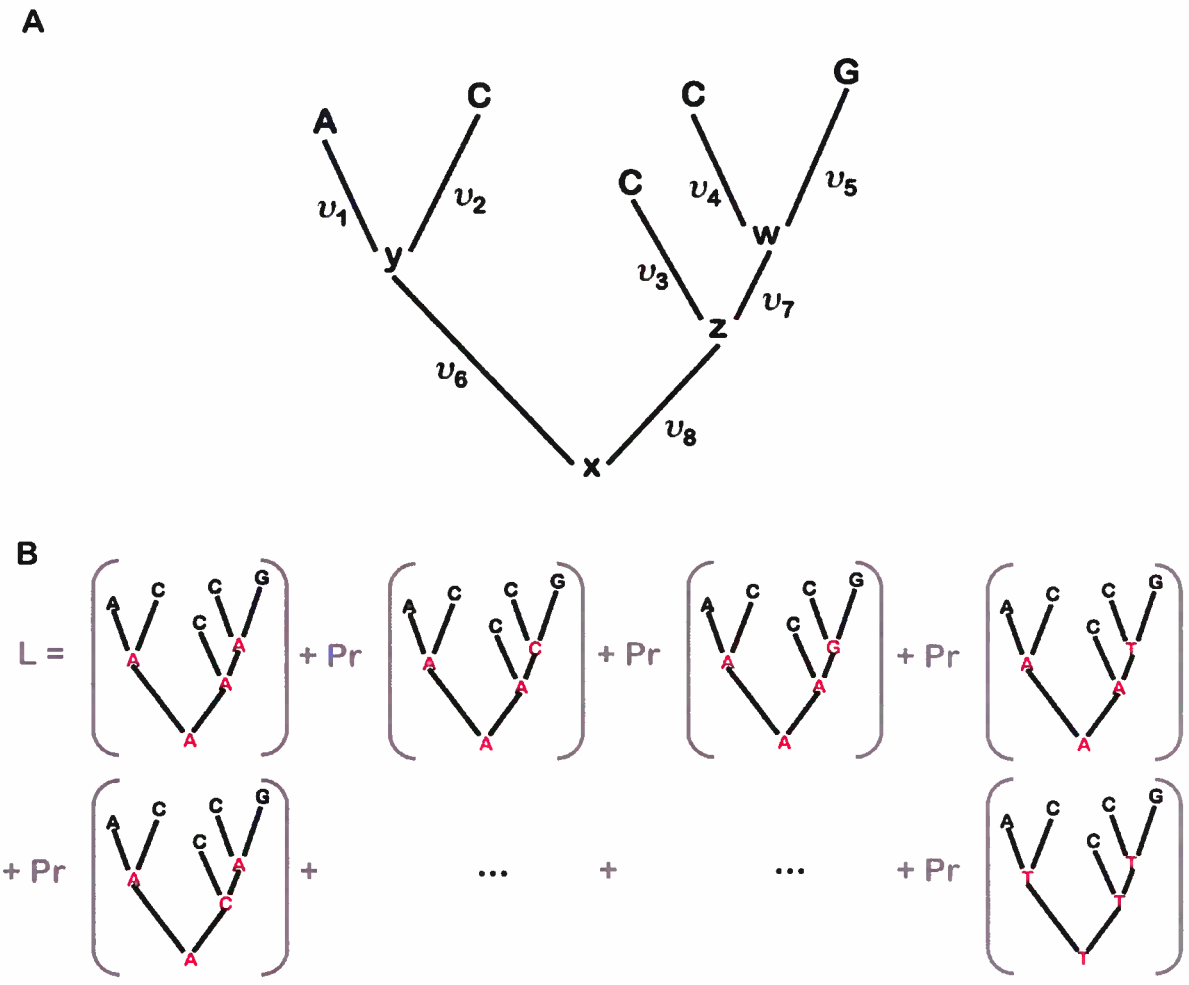

Figure 16 : Calcul de la vraisemblance pour une position donnée

A : Arbre utilisé comme exemple dans le texte pour décrire le calcul de la vraisemblance de la position 1 de l'alignement *D*. w, x, y et z représentent les états des nœuds internes et peuvent être A, T, G ou C. v_1, v_2, v_3, v_4, v_5, v_6, v_7 *et* v_8 représentent les longueurs des branches. **B :** La vraisemblance de la position 1 se calcule en sommant les probabilités de toutes les reconstructions possibles des états ancestraux étant donné des longueurs de branches v_1, v_2, v_3, v_4, v_5, v_6, v_7 *et* v_8 et un modèle d'évolution donné. Les états en rouge représentent des états inconnus. Chaque terme de la sommation représente une des reconstructions possibles des états ancestraux.

D'après Felsenstein (2004)

1.3.2.2. L'inférence bayésienne

Contrairement à la méthode de vraisemblance maximale, qui sélectionne l'hypothèse avec la meilleure vraisemblance, l'inférence bayésienne se base sur une distribution de probabilités *a priori* pour calculer la distribution des probabilités postérieures des hypothèses à évaluer. Étant l'une des méthodes d'inférence statistique les plus anciennes, l'application de l'approche bayésienne à l'inférence phylogénétique est relativement récente (Rannala et Yang, 1996). De plus, étant donné que le calcul des probabilités postérieures des arbres est analytiquement impossible, il aura fallu attendre l'implémentation de méthodes numériques telles que les chaînes de Markov avec la technique de Monte Carlo (ou MCMC pour *Markov Chain Monte Carlo*) pour rendre la méthode bayésienne applicable à la phylogénie (Larget et Simon, 1999; Mau, Newton et Larget, 1999; Yang et Rannala, 1997). Évidemment, l'utilisation de l'approche bayésienne en phylogénie moléculaire a été popularisée suite à son implémentation dans des programmes tels que BAMBE (Simon et Larget, 1998) et MrBayes (Huelsenbeck et Ronquist, 2001; Ronquist et Huelsenbeck, 2003).

1.3.2.2.1. Le théorème de Bayes

La méthode bayésienne se base sur la probabilité postérieure (Pr) d'une hypothèse (H) étant donné les données, $\Pr(H|D)$, qui est calculée par le théorème de Bayes

$$\Pr(H \mid D) = \frac{\Pr(D \mid H)\Pr(H)}{\Pr(D)},$$

où $\Pr(D|H)$ est la fonction de vraisemblance, $\Pr(H)$, la probabilité *a priori* de l'hypothèse et $\Pr(D)$, la probabilité des données. Cette probabilité d'apparence simple implique des sommations à travers tous les arbres possibles et, pour chaque arbre, l'intégration sur toutes les combinaisons possibles des longueurs de branches et des paramètres du modèle. Évidemment, ceci est analytiquement impossible. Pour pallier cette difficulté, le MCMC (Gilks, Richardson et J., 1996) est utilisé. Cette technique permet l'approximation de la probabilité postérieure d'un arbre et a révolutionné l'application de l'inférence bayésienne à la phylogénie (Huelsenbeck *et al.*, 2001; Rannala et Yang, 1996).

1.3.2.2.2. Le MCMC

L'idée sous-jacente au MCMC est la construction d'une chaîne qui, prenant la forme d'une marche guidée à travers l'espace multidimensionnel des paramètres, peut être utilisée pour estimer une distribution de probabilité en échantillonnant les valeurs des paramètres de façon périodique. L'approximation de la distribution sera d'autant plus exacte que le nombre de pas effectués par la chaîne est élevé. Dans le contexte phylogénétique, chaque pas correspond à une modification aléatoire de la topologie, des longueurs de branches ou d'un paramètre du modèle de substitution (Lewis, 2001). Un pas est accepté si la probabilité de la modification proposée est plus élevée que celle de l'état actuel; dans le cas contraire, le pas est accepté avec une probabilité qui dépend de la magnitude de diminution dans la probabilité selon la méthode de Metropolis-Hastings (Hastings, 1970; Metropolis et al., 1953). Dans un MCMC proprement construit, la proportion du nombre de visites d'un arbre pendant la marche guidée sur le nombre total d'arbres est une bonne estimation de la probabilité postérieure de cet arbre (Tierney, 1994). Autrement dit, si le nombre d'arbres visités tend vers l'infini, ces proportions vont tendre vers les probabilités postérieures correspondantes.

1.3.2.2.3. La méthode de Metropolis-Hastings

Pour illustrer le principe du MCMC avec la méthode de Metropolis-Hastings imaginons un espace d'arbres dans lequel nous nous promenons de manière aléatoire jusqu'à trouver une distribution d'arbres de probabilité postérieure supérieure au reste de l'espace. Un tel espace est représenté dans la Figure 17, où les cercles représentent les arbres avec les plus hautes probabilités postérieures. La méthode implique plusieurs étapes (Felsenstein, 2004):

1. L'arbre de départ est un arbre aléatoire T_i.
2. Une modification (p. ex., un réarrangement NNI) est appliquée à cet arbre

afin de créer une nouvelle proposition d'arbre T_j[7].

3. Le ratio entre les probabilités (fonctions de densité) des deux arbres T_i et T_j est calculé.

$$R = \frac{f(T_j)}{f(T_i)}$$

4. Si R ≥ 1, T_j est accepté et l'algorithme **recommence à l'étape 2 avec T_j** comme arbre de départ.

5. Si R < 1, un nombre aléatoire entre 0 et 1 est choisi.

    a. Si ce nombre est inférieur à R, l'algorithme **recommence à l'étape 2 avec T_j** comme arbre de départ.

    b. Sinon, T_j est rejeté et l'algorithme **recommence à l'étape 2 avec T_i** comme arbre de départ.

En suivant ces règles élémentaires, les points de l'espace seront explorés de manière proportionnelle à leur altitude, les points les plus élevés étant les plus fréquemment visités. Tel que décrit précédemment, cet algorithme ne termine jamais car il recommence toujours à l'étape 2, peu importe la valeur de R. De plus, comme c'est un processus aléatoire de type chaîne de Markov, l'acceptation d'un changement dépend seulement de l'état actuel et pas des états précédents. Il est donc nécessaire de déterminer le moment où le processus doit être arrêté, ce qui se mesure par la convergence de la chaîne.

[7] Pour simplifier l'exemple, nous imaginons que les paramètres du modèle restent constants pendant la recherche. Par contre, il faut noter que normalement les modifications dans paramètres font aussi partie des propositions.

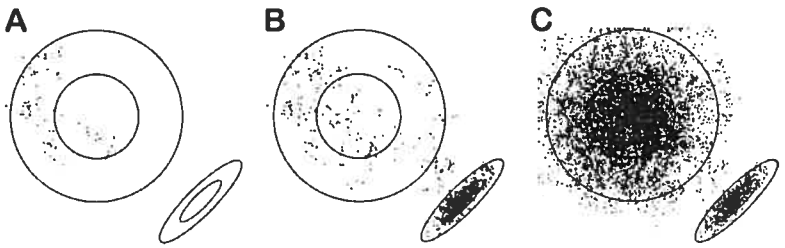

Figure 17 : Marche aléatoire illustrant le MCMC

Trajectoire après 100 (**A**), 1 000 (**B**) et 10 000 (**C**) pas sur une surface avec des collines et des vallées. Les cercles représentent les points les plus hauts de l'espace.

D'après Lewis (2001)

1.3.2.2.4. Convergence des MCMC

Une des limitations des méthodes bayésiennes pour l'inférence phylogénétique est l'ambiguïté dans la définition de la convergence d'une chaîne et l'incertitude concernant son arrêt. Le problème est illustré dans la Figure 17. Avec seulement 100 itérations (ou pas) de l'algorithme, une seule colline a été explorée. La rapidité avec laquelle une chaîne trouve un optimum global dépend du point de départ de la chaîne. Une approche courante pour valider les analyses bayésiennes est d'en réaliser plusieurs, chacune ayant son propre point de départ. Une autre solution est l'utilisation d'une variante de l'algorithme de Metropolis-Hastings, soit le MCMCMC (pour *Metropolis Coupling Markov Chain Monte Carlo*). Cet algorithme permet l'utilisation simultanée de plusieurs MCMC « chauffés » de façon graduelle. Les chaînes chaudes font de plus grands pas permettant une exploration plus extensive de l'espace et sont utilisées pour guider la chaîne froide à partir de la quelle les inférences sont faites (Huelsenbeck et Ronquist, 2001).

Une façon de vérifier la convergence des chaînes est l'étude de l'évolution de la fonction de vraisemblance pendant le parcours de la chaîne froide. Après une période pendant laquelle la vraisemblance augmente très rapidement, appelée « allumage » ou *burn-in* et exclue de l'analyse, la vraisemblance de la chaîne froide se stabilise. La stationnarité atteinte, un bon aperçu de la distribution de la probabilité postérieure peut en être inféré. Cette méthode est très rapide, mais peut donner des résultats erronés quant à la convergence de la chaîne. Il est donc recommandé d'utiliser d'autres analyses de convergence comme la déviation standard des fréquences des partitions (Huelsenbeck et

Ronquist, 2001). Il existe deux programmes conçus spécialement pour analyser la convergence des analyses bayésiennes d'inférence phylogénétique : TRACER (Rambaut et Drummond, 2003) et AWTY (Wilgenbusch, Warren et Swofford, 2004).

1.4. Les modèles d'évolution

La performance des méthodes probabilistes, que ce soit la méthode de vraisemblance ou l'inférence bayésienne, dépend de la fidélité du modèle évolutif à décrire le processus évolutif qui a généré les données observées. Ce modèle décrit la façon selon laquelle les séquences ont évolué par des remplacements de nucléotides ou d'acides aminés le long des branches de l'arbre. Ces remplacements sont les produits de substitutions dont les occurrences à chaque position peuvent être modélisées selon un processus de Markov, où la probabilité de passer de l'état *i* à l'état *j* dépend uniquement de l'état *i* et non pas des états précédents. D'autre part, il est en général supposé que ce processus de Markov est homogène, stationnaire et réversible. L'homogénéité implique que les probabilités de substitutions ne changent pas le long des branches, la stationnarité, que les fréquences de chaque état sont constantes le long de l'arbre, et la réversibilité, que la probabilité de passer de l'état *i* à l'état *j* est égale à la probabilité de passer de *j* à *i* (Lio et Goldman, 1998; Whelan, Lio et Goldman, 2001).

Les modèles d'évolution peuvent être construits de façon empirique en utilisant des propriétés calculées à partir de la comparaison d'un grand nombre de séquences observées, ou de façon paramétrique en se basant sur les propriétés chimiques ou biologiques des séquences étudiées. Les modèles empiriques sont composés de paramètres dont les valeurs fixes sont calculées une seule fois et supposées valides pour tous les jeux de données. En revanche, dans les modèles paramétriques, les valeurs des paramètres sont estimées à partir du jeu de données au cours de chaque analyse. Bien évidemment, les modèles paramétriques ont une meilleure performance que les modèles empiriques, mais ils demandent aussi plus de puissance de calcul (Whelan, Lio et Goldman, 2001).

1.4.1. Les matrices de substitution

1.4.1.1. Séquences nucléiques

La modélisation du taux de substitution entre chaque paire de nucléotides est généralement faite selon des modèles paramétriques. Les composantes principales des matrices de substitution de nucléotides sont les paramètres de fréquence des bases à l'équilibre (p_j) et les paramètres de taux d'échange (r_{ij}) d'un nucléotide *i* en un nucléotide *j*. On peut donc représenter un modèle de substitution de nucléotides par une matrice de taux instantanés, chaque taux caractérisant le remplacement d'un nucléotide par un autre,

$$Q = \begin{bmatrix} -r_{AC}\pi_C - r_{AG}\pi_G - r_{AT}\pi_T & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{CA}\pi_A & -r_{CA}\pi_A - r_{CG}\pi_G - r_{CT}\pi_T & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{GA}\pi_A & r_{GC}\pi_C & -r_{GA}\pi_A - r_{GC}\pi_C - r_{GT}\pi_T & r_{GT}\pi_T \\ r_{TA}\pi_A & r_{TC}\pi_C & r_{TG}\pi_G & -r_{TA}\pi_A - r_{GT}\pi_C - r_{TG}\pi_G \end{bmatrix},$$

où les éléments de la diagonale sont choisis de tel sorte que la somme des éléments de chaque lignes soit égale à 0.

Il existe plusieurs variantes de cette matrice selon le nombre des fréquences en nucléotides à l'équilibre et des taux d'échange qui sont considérés. Les variantes plus connues sont montrées dans le Tableau IV. Le modèle le plus simple, nommé Jukes et Cantor ou JC (Jukes et Cantor, 1969), assume que les fréquences en nucléotides à l'équilibre sont égales et que tous les échanges entre nucléotides ont la même probabilité ($\pi_A = \pi_C = \pi_G = \pi_T$ et $r_{AC} = r_{AT} = r_{CG} = r_{GT} = r_{AG} = r_{CT}$). Puisque dans les séquences d'ADN, la probabilité des transitions (échanges entre purines ou pyrimidines) est différente de celle des transversions (échanges entre une purine et une pyrimidine), Kimura introduisit, 10 ans plus tard, son modèle à deux paramètres, K2P (Kimura, 1980), qui prend en compte cette différence. Un an plus tard, Felsenstein introduisit son modèle F81 (Felsenstein, 1981), qui se base sur le JC, mais qui relaxe la supposition que les fréquences des quatre nucléotides sont les mêmes. Combinant les modèles K2P et F81, le modèle de Hasegawa, Kishino et Yano (HKY) tient compte de la différence entre transitions et transversions et entre les fréquences des nucléotides en même temps (Hasegawa, Kishino et Yano, 1985). Finalement, le modèle le plus complexe est le GTR (*General Time Reversible*), aussi appelé

REV (pour réversible), qui considère que les fréquences des nucléotides et que tous les taux d'échange peuvent être différents (Yang, 1994a).

Tableau IV : Exemples de modèles de substitution pour séquences nucléiques

Modèle[1]	Fréquences des bases	Taux d'échange	Paramètres libres[2]
JC	$\pi_A = \pi_C = \pi_G = \pi_T$	$r_{AC} = r_{AT} = r_{CG} = r_{GT} = r_{AG} = r_{CT}$	0
K2P	$\pi_A = \pi_C = \pi_G = \pi_T$	$r_{AC} = r_{AT} = r_{CG} = r_{GT} \neq r_{AG} = r_{CT}$	1
F81	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$r_{AC} = r_{AT} = r_{CG} = r_{GT} = r_{AG} = r_{CT}$	3
HKY	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$r_{AC} = r_{AT} = r_{CG} = r_{GT} \neq r_{AG} = r_{CT}$	4
GTR	$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$	$r_{AC} \neq r_{AT} \neq r_{CG} \neq r_{GT} \neq r_{AG} \neq r_{CT}$	8

[1]JC, Jukes et Cantor (Jukes et Cantor, 1969); K2P, Kimura deux paramètres (Kimura, 1980); F81, Felsenstein 81 (Felsenstein, 1981); HKY, Hasegawa, Kimura et Yano (Hasegawa, Kishino et Yano, 1985); GTR, *General time reversible* (Yang, 1994a).

[2] Nombre de paramètres qui doivent être estimés pendant le calcul de la vraisemblance.

Étant donné qu'on parle ici de modèles paramétriques, une matrice plus complexe implique plus de paramètres à estimer lors du calcul de la vraisemblance (Tableau IV). Par exemple, dans le modèle JC, il n'y a aucun paramètre à estimer : les fréquences des nucléotides sont égales, soit ¼ chacune, ainsi que les taux de substitution, qui se mesurent de façon relative les uns par rapport aux autres. Seulement un paramètre à estimer est ajouté dans le modèle K2P, celui du taux de transition versus transversion (N), on dit alors que les transitions sont N fois plus probables que les transversions. Le modèle le plus gourmand quant au nombre de paramètres est le GTR, avec 8 paramètres à estimer : 3 fréquences de nucléotide (et non pas quatre, car l'addition de fréquences doit être égale à 1), et 5 taux de substitution (et non pas six, car ces 5 taux se calculent par rapport au sixième).

1.4.1.2. Séquences protéiques

Contrairement aux matrices de substitution de nucléotides, qui sont paramétriques, les matrices de substitution d'acides aminés sont généralement empiriques, car le nombre de paramètres à estimer est plus grand (matrices 20x20 *versus* 4x4). En 1978, Dayhoff *et al.* ont proposé un modèle d'évolution de protéines qui a résulté dans le développement d'un ensemble de matrices de substitution d'acides aminés très largement utilisées. Leur modèle (Figure 18) fut simplement dérivé du comptage des substitutions d'acides aminés observées dans de grandes bases de données de protéines globulaires. Ils utilisèrent uniquement des séquences très proches pour réduire la fréquence auxquelles des substitutions observées (p. ex. A→S) étaient le résultat d'une suite de remplacements non observés (p. ex. A→x→y→S). Plus récemment, Jones et collaborateurs (Jones, Taylor et Thornton, 1992) ont utilisé la même méthodologie que Dayhoff *et al.*, mais avec des bases de données plus complètes et en n'incluant pas que des protéines globulaires. Le modèle résultant (JTT; Figure 18) est actuellement l'un des plus utilisés en phylogénie. Parmi les matrices les plus utilisées, il y a aussi WAG (Whelan et Goldman, 2001), mtREV (Adachi et Hasegawa, 1996a) et cpREV (Adachi et al., 2000), les deux dernières étant respectivement conçues pour les protéines mitochondriales et chloroplastiques.

Normalement, l'analyse phylogénétique de séquences protéiques utilise des paramètres de fréquence estimés à partir du jeu de données (π_i) et des paramètres d'échange dérivés des matrices mentionnées ci-dessus. Ceci force les fréquences en acides aminés de ce modèle hybride à correspondre à celles des données observées, mais en incorporant de l'information sur les paramètres d'échange dérivée de la base de données utilisée pour créer la matrice. Ces applications sont normalement désignées avec un suffixe +F.

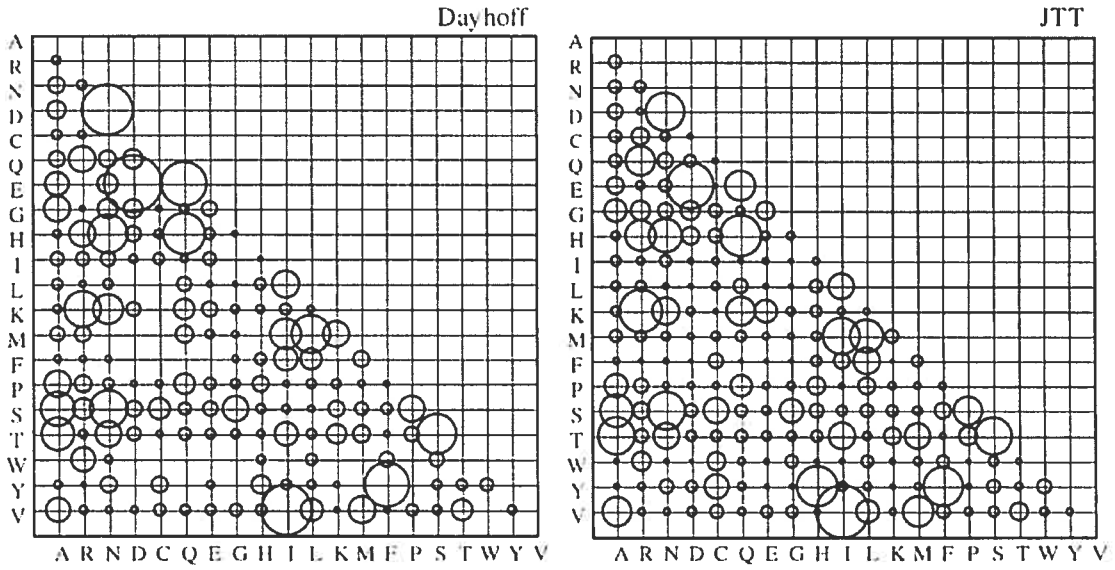

Figure 18 : Paramètres d'échange pour deux modèles de substitution en acides aminés
La valeur de chaque paramètre d'échange entre chaque pair d'acides aminés est indiquée par la surface des cercles. Par exemple, l'échange de deux acides aminés ayant des propriétés similaires (ex. D/E, V/I, FY) est très probable, tandis que celui de deux acides aminés peu similaires a une probabilité très petite.

D'après Whelan et Goldman (2001)

1.4.2. L'hétérogénéité du taux de substitution entre les positions

Les modèles présentés précédemment supposent que toutes les positions d'une protéine ou d'une séquence nucléique évoluent à la même vitesse. Cependant, les différentes positions d'une séquence sont soumises à des contraintes évolutives différentes dont dépend le taux de fixation d'une mutation. Il existe donc des modèles qui estiment les taux de substitution propres à chaque position (Figure 19).

L'estimation d'un taux d'évolution propre à chaque position est analytiquement difficile due au grand nombre de paramètres impliqués. Des alternatives à cette approche ont donc été développées. Celles-ci consistent à grouper les positions en catégories (p. ex., très lente, lente, moyenne, rapide et très rapide). La méthode la plus simple pour une telle classification est d'estimer la fraction des positions invariable, le reste représentant les positions variables (Adachi et Hasegawa, 1995; Fitch et Margoliash, 1967). Ce modèle est désigné avec le suffixe +I.

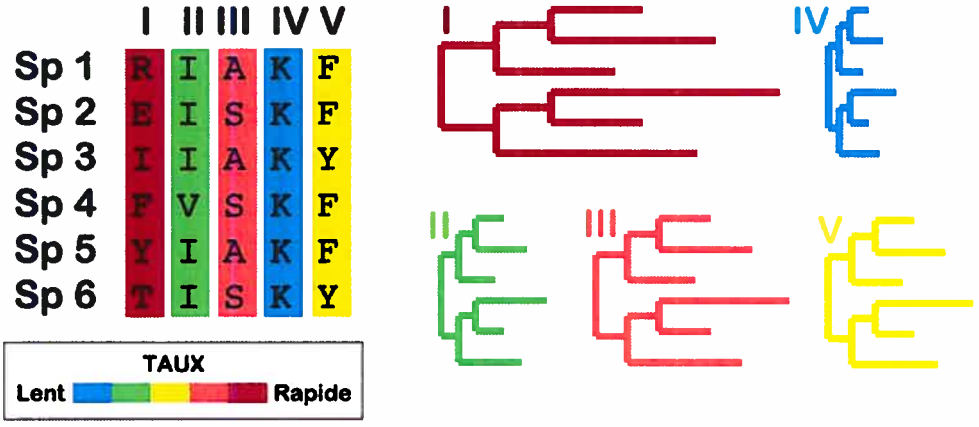

Figure 19 : Hétérogénéité du taux de substitution entre positions

Selon un modèle où l'hétérogénéité du taux de substitution des positions est considérée, l'évolution de chaque position peut être plus lente ou plus rapide que le taux d'évolution moyen de l'alignement. Noter que les longueurs de branches relatives sont les mêmes pour toutes les positions. Sp, espèce.

Pour une modélisation plus réaliste de l'hétérogénéité du taux de substitution entre les positions, plusieurs auteurs ont proposé des modèles basés sur des distributions Gamma continues (Nei et Gojobori, 1986; Yang, 1993). Afin de rendre de tels modèles plus pratiques d'un point de vue analytique, Yang (1994b) proposa une solution efficace consistant en une distribution Gamma discrète, qui comporte au moins quatre catégories pour approcher la distribution continue. Avec ce modèle, le taux de substitution de chaque position (r) est échantillonné d'une distribution Gamma de paramètre de forme α, où $\beta=1/\alpha$ pour que la distribution aie une moyenne de 1 (Felsenstein, 2004):

$$f(r)=\frac{\alpha^{\alpha}}{\Gamma(\alpha)}r^{\alpha-1}e^{-\alpha r}.$$

Si la valeur de α est inférieur à 1, alors la forme de la distribution est telle que la majorité des positions évoluent très lentement et seulement quelques-unes évoluent très rapidement. Pour des valeurs de α supérieures à 1, la distribution change de forme : la majorité des positions ont un taux très similaire (Figure 20). Les formes de distribution disponibles avec toutes les valeurs de α ($0 < \alpha < \infty$) sont considérées suffisantes pour décrire toute la variation retrouvée dans les séquences. Les modèles incluant une

distribution Gamma pour modéliser le taux d'hétérogénéité entre les positions sont désignés par +Γ.

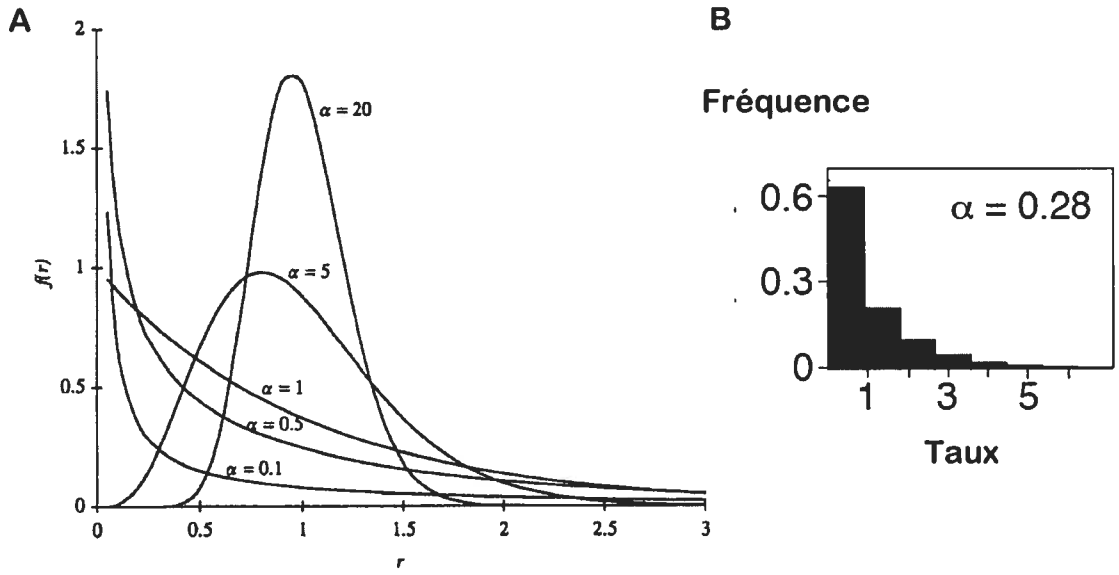

Figure 20 : Représentation graphique de la distribution gamma

A : Fonction de densité *f(r)* de la distribution Gamma des taux de substitution par position (*r*) pour différentes valeurs du paramètre α. **B :** Exemple de distribution Gamma (pour α=0.28) discrétisée.

D'après Yang et Kumar (1996) et Whelan, Lio et Goldman (2001)

1.4.3. L'hétérogénéité du processus de substitution

Les modèles décrits précédemment supposent que le processus évolutif est stationnaire et homogène, c'est-à-dire que les fréquences de nucléotides ou d'acides aminés sont constantes le long de l'arbre (stationnarité) et que le taux de substitution ne varie pas entre les positions à travers le temps (homotachie). Cependant, ces deux hypothèses sont souvent violées par le processus évolutif, et plusieurs modèles alternatifs ont été développés afin de les prendre en considération.

1.4.3.1. L'hétérogénéité compositionnelle

Les paramètres de fréquence (π_j) des modèles de substitution définissent la distribution des fréquences des états à l'équilibre. Dans les modèles stationnaires, ces fréquences ne varient pas le long de l'arbre, ce qui implique que la composition en nucléotides ou en acides aminés est supposée la même chez toutes les espèces considérées, incluant leurs ancêtres. Or, ceci n'est pas toujours le cas (Mooers et Holmes, 2000). Par exemple, les génomes des bactéries *Micrococcus luteus* et *Mycoplasma capricolum* ont respectivement une composition en G+C de 74% et 25%. Il a été démontré que l'hétérogénéité dans la composition en nucléotides ou en acides aminés (hétérogénéité compositionnelle) peut avoir un impact drastique sur l'inférence phylogénétique, provocant le regroupement d'espèces avec des compositions similaires indépendamment de leurs vraies relations phylogénétiques (Foster et Hickey, 1999; Galtier et Gouy, 1995; Lockhart *et al.*, 1992).

L'utilisation de modèles non homogènes qui considèrent l'hétérogénéité compositionnelle permet d'améliorer l'inférence phylogénétique lors de l'analyse de jeux de données présentant de fortes déviations par rapport à l'hypothèse de stationnarité (Foster, 2004; Galtier et Gouy, 1998). Toutefois, ces types de modèles empêchent d'utiliser l'hypothèse de réversibilité des modèles évolutifs courants, ce qui rend difficile leur implémentation et ils sont donc peu utilisés. La méthode LogDet (Logarithme du déterminant) fut développée comme alternative plus simple aux modèles non homogènes (Lake, 1994; Lockhart et al., 1994). Cette méthode calcule les distances entre chaque paire d'espèces en corrigeant pour la variation de la composition en nucléotides ou acides aminés. Par contre, étant basée sur des distances, la méthode LogDet ne tient pas compte des autres facteurs du processus évolutif et s'avère inefficace dans certaines situations (Conant et Lewis, 2001).

1.4.3.2. L'hétérotachie

Les modèles homogènes ne tiennent pas compte de la variation du taux de substitution à travers les positions au cours du temps, un phénomène connu sous le nom d'hétérotachie (Philippe et Lopez, 2001). Pourtant, cette variabilité existe et est due au fait

que les positions critiques pour le fonctionnement d'une protéine ne sont pas les mêmes à travers le temps. Dans les années 70, Fitch et Markowitz (1970) montrèrent que, à un instant donné, seulement une petite fraction des positions sont susceptibles de varier, permettant la fixation de mutations. Quand une ou plusieurs mutations ont été fixées, les contraintes agissant sur la protéine (p. ex., pour permettre son repliement ou pour maintenir sa fonction) ont pu changer et les positions qui sont libres de varier ne sont peut-être plus les mêmes (Figure 21).

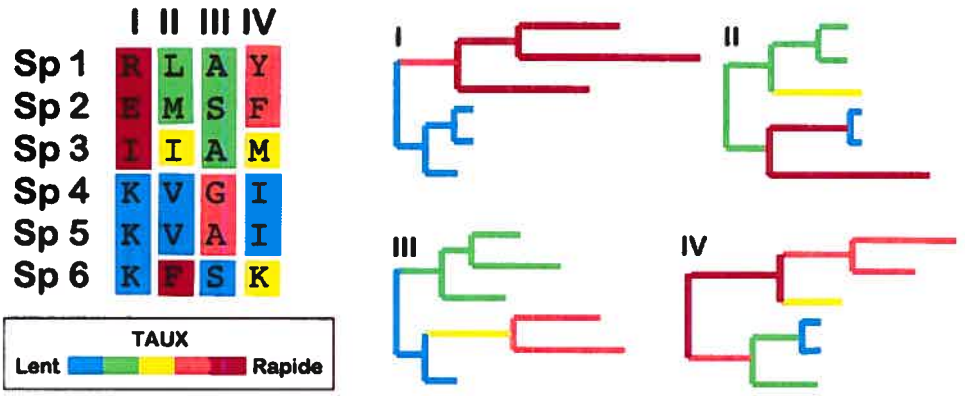

Figure 21 : Représentation schématique de l'hétérotachie
La variation du taux de substitution au cours du temps peut être différente d'une position à une autre. Noter qu'ici, contrairement à la Figure 19, les longueurs de branches relatives changent entre les positions. Sp, Espèce.

Plusieurs études ont confirmé de façon convaincante l'impact de l'hétérotachie sur les reconstructions phylogénétiques (Kolaczkowski et Thornton, 2004; Lockhart *et al.*, 2006; Lockhart *et al.*, 1996; Philippe *et al.*, 2005b; Spencer, Susko et Roger, 2005) et des tests ont été développés pour la détecter (Baele *et al.*, 2006; Lockhart *et al.*, 1998; Lopez, Forterre et Philippe, 1999). Pour surmonter les effets de l'hétérotachie dans la reconstruction phylogénétique, Fitch développa le modèle covarion (pour *concomitantly variable codons*), qui postule qu'au cours du temps les positions peuvent passer d'un état libre de varier (*on*) à un état fixé (*off*) (Fitch, 1971a). Suite à la formalisation mathématique

de ce modèle (Tuffley et Steel, 1998), plusieurs adaptations ont été implémentées (Galtier, 2001; Huelsenbeck, 2002; Wang et al., 2006).

1.4.4. Le modèle parfait

Les méthodes probabilistes permettent de rendre explicite l'ensemble des suppositions sous-jacentes à la reconstruction phylogénétique sous la forme d'un modèle d'évolution de séquences. De plus, ces méthodes sont dites consistantes, c'est-à-dire qu'elles convergent vers la solution correcte au fur et à mesure qu'on augmente le nombre de caractères utilisés. Toutefois, la consistance est garantie seulement lorsque le modèle utilisé décrit correctement l'évolution des séquences réelles (Huelsenbeck et Hillis, 1993). Or, les modèles actuellement existants sont loin de capturer toute la complexité du comportement des séquences moléculaires au cours de l'évolution.

Plusieurs hypothèses des modèles les plus couramment utilisés en phylogénie moléculaire sont violées par l'évolution réelle des séquences (Philippe *et al.*, 2005a). Tel que décrit ci-dessus, l'homogénéité et la stationnarité du processus évolutif figurent parmi ces suppositions, mais elles ne sont pas les seules. Par exemple, la majorité des modèles supposent que les positions d'une séquence nucléique ou protéique évoluent de façon indépendante les unes des autres, alors que les contraintes structurales imposent un certain degré de co-évolution entre elles. De même, la majorité des modèles supposent que chaque position peut accepter tous les états de caractère possibles (évolution qualitativement homogène), alors que, dû à des contraintes structurales ou fonctionnelles, certaines positions n'acceptent qu'un nombre limité d'états.

Depuis plusieurs années, des modèles sont développés afin de prendre en compte de plus en plus de facteurs affectant le processus évolutif, comme ceux qui considèrent l'hétérogénéité compositionnelle (Foster, 2004; Galtier et Gouy, 1995), l'hétérotachie (Galtier, 2001; Huelsenbeck, 2002), la non indépendance entre les positions (Robinson *et al.*, 2003; Rodrigue *et al.*, 2005) ou l'hétérogénéité qualitative du processus évolutif (Lartillot et Philippe, 2004). Cependant, même un modèle qui tient compte de toutes ces caractéristiques ne sera pas parfait, car le processus évolutif est complexe et plusieurs facteurs impliqués dans l'évolution des séquences nous sont encore inconnus. Dans cette

perspective, au lieu de chercher le « modèle parfait », le but doit plutôt être de trouver le modèle le mieux adapté aux données, soit celui ayant un nombre suffisant de paramètres permettant de capturer les caractéristiques clés des données (Steel, 2005).

1.5. Les tests statistiques de comparaison de modèles

La sélection d'un modèle demeure un problème majeur dans la reconstruction phylogénétique (Cunningham, Zhu et Hillis, 1998). En effet, l'utilisation de modèles différents peut amener à des résultats drastiquement distincts (Cunningham, Zhu et Hillis, 1998; Kelsey, Crandall et Voevodin, 1999; Sullivan et Swofford, 1997). Il est donc nécessaire d'identifier le modèle qui s'ajuste le mieux à chaque jeu de données (Huelsenbeck, 1995).

Les modèles les plus complexes (avec plus de paramètres) tendent à mieux s'ajuster aux données (Goldman, 1993; Yang, Goldman et Friday, 1994), estimant des phylogénies plus robustes (Huelsenbeck, 1995), mais ceci n'est pas toujours vrai. De plus, l'utilisation de modèles complexes a plusieurs désavantages : (i) un grand nombre de paramètres doivent être estimés, ce qui rend l'analyse intense et longue; et (ii) plus le nombre de paramètres estimés est grand, plus grande est l'erreur incluse dans leur estimation (Huelsenbeck et Crandall, 1997). Pour identifier le modèle qui s'ajuste le mieux aux données et qui possède le nombre minimum de paramètres, plusieurs tests statistiques ont été développés, les plus connus étant le test du ratio des vraisemblances, le critère d'information d'Akaike et le critère d'information bayésien. Ceux-ci peuvent être calculés de manière automatique par les programmes MODELTEST (Posada et Crandall, 1998) et ProtTest (Abascal, Zardoya et Posada, 2005).

1.5.1. Le test du ratio des vraisemblances

Le test du ratio des vraisemblances (LRT pour *Likelihood Ratio Test*) vérifie si un modèle plus complexe a significativement de meilleures performances qu'un modèle plus simple pour le jeu de données considéré. Ce test n'est valide que pour comparer des modèles imbriqués, c'est-à-dire que le modèle le plus complexe peut être réduit au modèle

le plus simple en modifiant les valeur des paramètres (Huelsenbeck et Crandall, 1997).

Les modèles plus riches en paramètres donnant de meilleures vraisemblances, le LRT permet de voir si cette amélioration est significative étant donné la complexité ajoutée. Cette statistique, qui suit une distribution χ^2 avec un nombre de degrés de liberté égal au nombre de paramètres supplémentaires dans le modèle le plus complexe, est calculée selon la formule suivante :

$$LRT = 2(\log L_1 - \log L_2).$$

1.5.2. Le critère d'information d'Akaike

Le critère d'information d'Akaike (ou AIC pour *Akaike Information Criterion*) (Akaike, 1973) d'un modèle est calculé selon la formule suivante :

$$AIC = -2\log L + 2K,$$

où K est le nombre de paramètres libres du modèle. En phylogénie, l'AIC peut être vu comme la quantité d'information perdue quand on utilise un modèle donné pour faire l'approximation du processus réel d'évolution des séquences. Les AICs de faibles valeurs sont donc recherchés. Quand le nombre de paramètres des modèles à comparer est grand, des modèles plus complexes seront favorisés de façon artificielle (Forster et Sober, 2004). En conséquence, quand la taille de l'alignement (n) est petite par rapport au nombre de paramètres (K), soit $n/K < 40$, l'utilisation de l'AIC de deuxième ordre (Hurvich et Tsai, 1989),

$$AIC_c = -2\log L + 2K + \frac{2K(K+1)}{n-K-1},$$

corrigé pour cette caractéristique, est recommandée. Le AIC a deux avantages principaux par rapport au LRT : il n'est pas limité aux modèles imbriqués et il permet la comparaison d'un grand nombre à la fois, pas seulement deux à deux comme avec le LRT.

1.5.3. Le critère d'information bayésien

Le critère d'information bayésien (BIC, pour *Bayesian Information Criterion*) correspond à une approximation du logarithme de la vraisemblance marginale d'un modèle (Schwartz, 1978). De même que l'AIC_c, il est pondéré par rapport à la taille de l'alignement (n) :

$$BIC = -2\log L + 2K\log n.$$

Les modèles avec les plus petites valeurs BICs sont ceux qui ont la plus grande probabilité postérieure. Le BIC ressemble au AIC et a les mêmes avantages par rapport au LRT. Toutefois, il a été suggéré, mais avec certaines réserves, que dans certaines conditions le BIC est statistiquement consistant lorsque l'AIC ne l'est pas (Forster, 2002).

1.6. Les indices de robustesse des topologies

Indépendamment de la méthode utilisée pour inférer une phylogénie, il est important d'estimer le degré de confiance accordé à chacune de ses branches. Ceci est généralement fait en utilisant les méthodes de bootstrap et de jackknife non-paramétriques, utilisées dans un contexte phylogénétique en premier par Felsenstein (1985), Mueller et Ayala (1982), et Penny et Hendy (1985; 1986). Ces deux méthodes consistent à échantillonner la population d'étude une multitude de fois et à comparer les estimations obtenues à partir des différents échantillons. Plus précisément, le bootstrap génère n matrices pseudo-répliquées de même taille que la matrice initiale par tirage aléatoire avec remise des positions de cette dernière. Ces matrices pseudo-répliquées sont ensuite analysées par la même méthode de reconstruction que la matrice originale, et les fréquences d'apparition des clades dans l'ensemble de ces n matrices sont interprétées comme des pourcentages de confiance pour chacune des branches (Figure 22). Le jackknife est très similaire : la seule différence repose dans le fait que les matrices pseudo-répliquées se génèrent en éliminant K positions de la matrice originale, K étant généralement 25 ou 50.

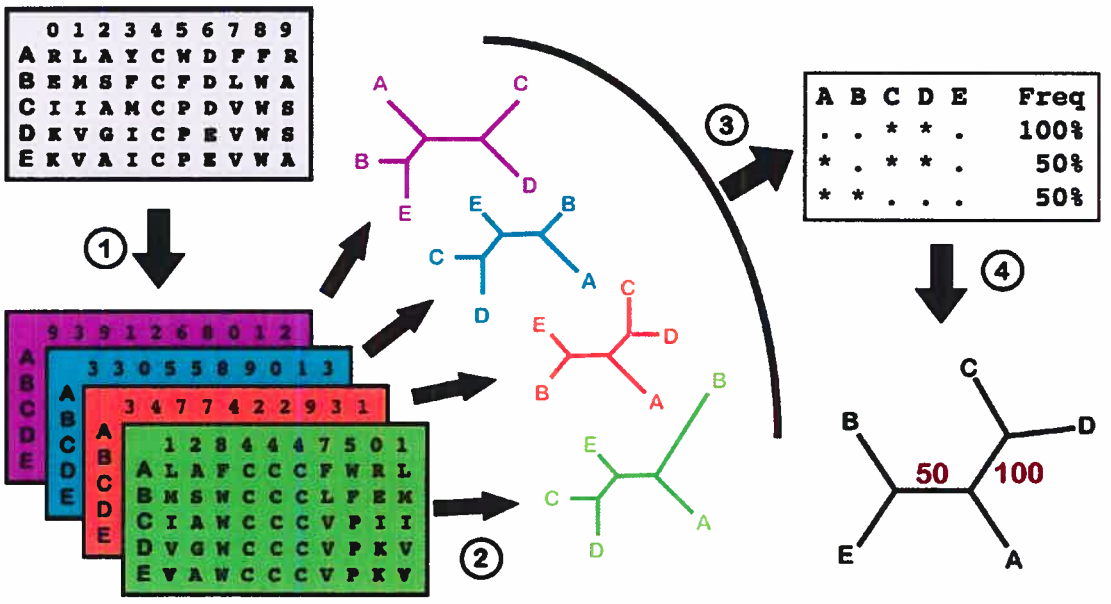

Figure 22 : Le bootstrap

Des matrices pseudo-répliquées sont construites en échantillonnant avec remise des positions de la matrice originale (1) et un arbre phylogénétique est généré pour chacune des matrices (2). Les fréquences d'occurrence de chacun des clades sont calculées (3) et représentées dans un arbre dit consensus (4).

1.7. Les tests statistiques de comparaison de topologies

Il existe plusieurs tests pour comparer des topologies alternatives pour un même jeu de données. Les premiers furent développés par Kishino et Hasegawa (1989) et avaient pour but d'estimer la variance et l'intervalle de confiance de la différence de vraisemblance entre deux topologies proposées pour un même jeu de données. Pour expliquer le principe des tests Kishino et Hasegawa (KH), considérons deux topologies pour lesquelles on veut déterminer si elles sont également supportées par les données (hypothèse nulle). Intuitivement, nous attendons que, par erreur d'échantillonnage, les vraisemblances des deux topologies (L_1 et L_2) ne soient pas identiques même si l'hypothèse nulle est vraie. Par contre, si on était capable d'obtenir plusieurs jeux de données (p. ex., des pseudo-répliqués de la matrice originale), on espérerait qu'« en moyenne » les vraisemblances des deux topologies soient égales si l'hypothèse nulle est vraie. Ceci pourrait être calculé avec la technique du bootstrap, mais étant donné le temps de calcul requis, les tests KH se basent

plutôt sur la méthode RELL (pour *Resampling Esitmated Log-Likelihood*), qui re-échantillonne les logarithmes de vraisemblance déjà estimés pour chaque position (Kishino, Miyata et Hasegawa, 1990).

Shimodaira et Hasegawa (1999) ont proposé une version modifiée du test KH, le SH, permettant la comparaison de multiples topologies simultanément. Le test SH étant trop conservatif, Shimodaira (2002) proposa un nouveau test, le AU (pour *Approximately Unbiased*), qui apparaît plus efficace que le KH, mais moins conservatif que le SH. L'utilisation de ces tests pour discriminer entre topologies est très répandue et plusieurs d'entre eux sont implémentés dans des logiciels tels que PAUP (Swofford, 2002), TREE-PUZZLE (Schmidt *et al.*, 2002) et CONSEL (Shimodaira et Hasegawa, 2001).

1.8. Erreurs dans l'inférence phylogénétique

L'inférence phylogénétique peut être affectée par deux types d'erreur, stochastique et systématique, entraînant la déviation entre un paramètre d'une population et son estimation. La différence entre les deux types d'erreur repose sur le fait que, dans l'erreur stochastique, cette déviation est causée strictement par la taille limitée de l'échantillon, alors que dans l'erreur systématique, elle est due à des présomptions incorrectes de la méthode d'estimation. Par définition, l'erreur stochastique disparaît dans des échantillons de taille infinie, contrairement à l'erreur systématique, qui persiste, et peut même être intensifiée (Swofford *et al.*, 1996).

1.8.1. L'erreur stochastique

Même si l'évolution a eu lieu exactement telle que supposée par le modèle évolutif utilisé pour l'inférence phylogénétique, un arbre incorrect peut être obtenu. Ceci est dû à la taille finie des alignements, qui génère des événements de chance induisant l'erreur stochastique (Swofford *et al.*, 1996). Pour illustrer ceci, imaginons un tirage à pile ou face avec une monnaie non truquée, c'est-à-dire avec une probabilité de 1/2 de tomber sur face. Plus grand est le nombre de lancés, plus petite est la différence entre le nombre de faces

obtenues et 50%. En généralisant, plus la taille de l'échantillon est grande, plus l'estimation des paramètres sera proche de leur valeur réelle.

Le même principe s'applique à la reconstruction phylogénétique, où la taille des alignements est toujours finie. D'ailleurs, plus la taille de l'alignement utilisé est petite, plus l'arbre obtenu est affecté par l'erreur stochastique. Une manière de mesurer l'influence de l'erreur stochastique est de calculer la variance des estimations obtenues à partir de plusieurs échantillons, ce qui peut être fait avec des méthodes telles que le bootstrap et le jackknife expliquées ci-dessus (Swofford *et al.*, 1996).

1.8.2. L'erreur systématique

L'erreur systématique survient lorsque le processus évolutif viole les suppositions du modèle utilisé pour la reconstruction phylogénétique. Cette erreur est d'autant plus grande que le nombre de caractères utilisés est grand (Phillips, Delsuc et Penny, 2004). Pour que l'erreur systématique mène à un arbre erroné, la magnitude du biais induit pas les violations du modèle (signal non-phylogénétique) doit excéder le support légitime pour l'arbre correct (signal phylogénétique). De plus, le biais doit aller dans la direction de l'arbre incorrect. Contrairement à l'erreur stochastique, les analyses de bootstrap ou de jackknife n'indiquent pas la présence d'erreur systématique dans un jeu de données, il faut donc trouver d'autres moyens de le détecter.

Pour minimiser l'erreur systématique et ses effets dans la reconstruction phylogénétique, plusieurs stratégies ont été proposées. Celles-ci consistent à éliminer le signal non-phylogénétique des jeux de données, c'est-à-dire les substitutions mal interprétées par les méthodes de reconstruction et qui supportent une topologie incorrecte. La majorité de ces techniques se basent sur le fait que les artefacts de reconstruction phylogénétique son dus à des substitutions multiples qui ne sont pas correctement identifiées comme des reversions ou conversions par les méthodes d'inférence (Olsen, 1987). Les approches les plus simples consistent soit à éliminer les espèces à taux d'évolution rapide, par définition, celles qui accumulent des substitutions multiples (Aguinaldo *et al.*, 1997; Philippe, Lartillot et Brinkmann, 2005), soit à ajouter des espèces,

préférablement à taux d'évolution lent, pour diviser les longues branches (Hendy et Penny, 1989; Hillis, 1996).

Une autre approche qui s'est avérée efficace dans le combat des artefacts de reconstruction phylogénétique est l'élimination des positions à taux d'évolution rapide. Celles-ci sont saturées par des substitutions multiples et ont donc perdu leur signal phylogénétique. Par exemple, la méthode SF (pour *Slow/Fast*) (Brinkmann et Philippe, 1999) consiste à identifier les positions qui n'ont subi aucune substitution à l'intérieur des groupes prédéfinis (positions les plus lentes), et à rajouter progressivement les positions qui ont subi au plus une, deux, trois, etc., substitutions. Ceci crée un ensemble de matrices imbriquées, qui contiennent de plus en plus de positions rapides. Plusieurs analyses ont montré que les matrices qui contiennent plus de positions rapides ont plus tendance à induire des artefacts de reconstruction phylogénétique (Brinkmann et Philippe, 1999; Brochier et Philippe, 2002; Delsuc, Brinkmann et Philippe, 2005; Philippe *et al.*, 2000).

Pour éliminer le signal non-phylogénetique causé par l'hétérogénéité compositionnelle, plusieurs auteurs ont proposé des techniques pour ajuster l'évolution des données à un modèle stationnaire. Ceci consiste à regrouper les états de caractère en groupes fonctionnels. Par exemple, le codage RY (Woese et al., 1991) consiste à remplacer les nucléotides A et G par R (purine) et C et T par Y (pyrimidine). Ainsi, le fait que dans quelques branches les substitutions A→G et T→C sont plus probables que les substitutions G→A et C→T est compensé. Un système de codage similaire a été proposé pour des séquences d'acides aminés, où ceux-ci sont codés selon leurs caractéristiques physico-chimiques (Hrdy et al., 2004). Ces méthodes ont été utilisées avec succès pour résoudre des questions phylogénétiques difficiles (Delsuc, Phillips et Penny, 2003; Hrdy *et al.*, 2004; Phillips et Penny, 2002).

1.9. La phylogénomique

Les analyses phylogénétiques basées sur un seul gène sont souvent affectées par l'erreur stochastique due au nombre limité de positions informatives contenu dans les petits jeux de données. En conséquence, les analyses basées sur une multitude de gènes sont de plus en plus populaires. En effet, la quantité croissante de données générée par les

projets de séquençage à échelle génomique a donné naissance à un nouveau champ de recherche, la phylogénomique, qui se base sur les principes de la phylogénétique pour expliquer les données génomiques (Eisen et Fraser, 2003). L'utilisation de données à large échelle pour inférer l'histoire évolutive des organismes constitue une des branches de cette nouvelle discipline.

Il existe plusieurs méthodes d'inférence phylogénomique basées soit sur des caractéristiques génomiques, soit sur des séquences primaires (Figure 23). Les premières consistent à comparer des génomes en se basant sur leur contenu en gènes, sur l'ordre de ceux-ci et sur la présence de changements génomiques rares, etc., alors que les deuxièmes se basent sur la comparaison des séquences primaires pour construire des arbres phylogénétiques (Delsuc, Brinkmann et Philippe, 2005). Le point de départ aux analyses phylogénomiques basées sur des séquences primaires sont les alignements de chaque gène. Une fois ceux-ci obtenus, deux approches alternatives peuvent être utilisées (Figure 23). La stratégie la plus populaire est l'analyse de la concaténation des alignements individuels en utilisant les méthodes standard d'inférence phylogénétique (super-matrice). Dans cette approche, les séquences des gènes qui sont absentes pour quelques-unes des espèces sont codées par des points d'interrogation. Des études récentes ont montré que la proportion de données manquantes peut être relativement haute sans que ceci entraîne une perte d'efficacité dans la reconstruction phylogénétique (Driskell *et al.*, 2004; Philippe *et al.*, 2004; Wiens, 2003). La deuxième approche consiste à analyser chaque partition (p. ex., chaque gène) séparément et à combiner les arbres résultants dans un super-arbre. Bien que moins populaire que la méthode de super-matrice, cette approche s'avère très utile pour combiner des arbres obtenus à partir de jeux de données disparates (p. ex., des données moléculaires et des données morphologiques) (Liu *et al.*, 2001).

Depuis quelques années, plusieurs études empiriques ont montré que l'utilisation de jeux de données à l'échelle génomique peut résoudre des questions pour lesquelles les phylogénies basées sur un seul gène avaient échoué (Bapteste *et al.*, 2002; Madsen *et al.*, 2001; Murphy *et al.*, 2001; Qiu *et al.*, 1999; Rokas *et al.*, 2003; Soltis, Soltis et Chase, 1999). La phylogénomique a alors été considérée comme la panacée pour résoudre l'arbre du vivant (Gee, 2003; Rokas *et al.*, 2003). Nonobstant, étant donné que l'erreur systématique est aggravée dans les grands jeux de données, des résultats parfaitement

résolus mais incorrects sont aussi attendus (Delsuc, Brinkmann et Philippe, 2005; Jeffroy *et al.*, 2006; Phillips, Delsuc et Penny, 2004).

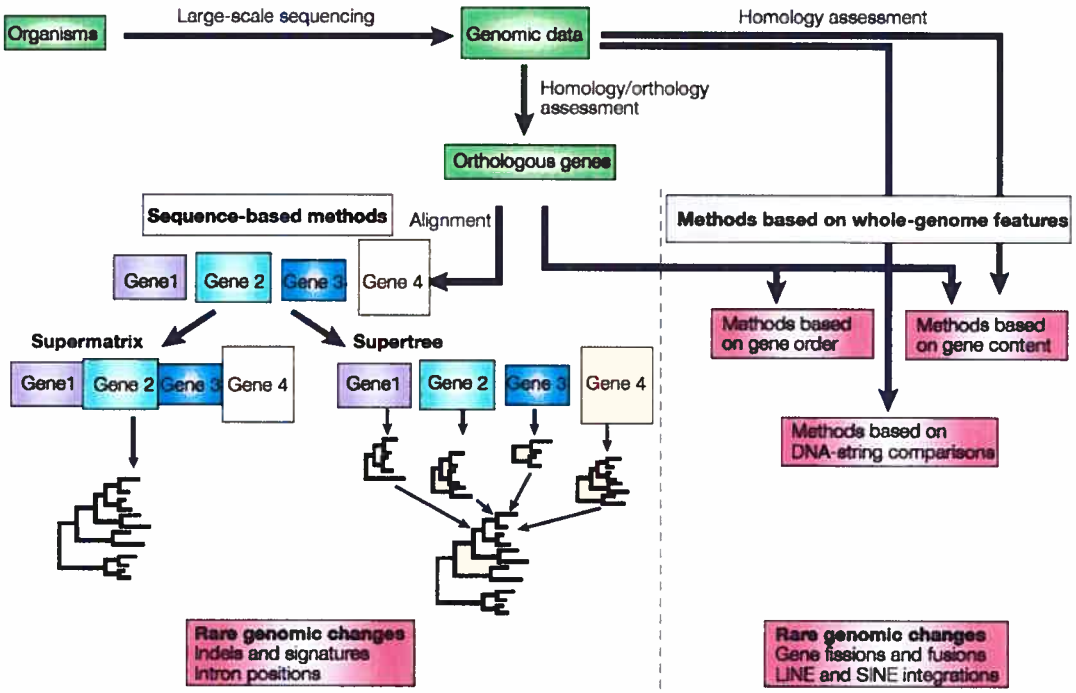

Figure 23 : Méthodes d'inférence phylogénomique

Méthodes basées sur des séquences primaires : la méthode de super-matrice consiste à concaténer les alignements des gènes avant l'analyse phylogénétique, tandis que la méthode de super-arbre consiste à combiner les arbres optimaux obtenus à partir de chaque gène. Méthodes basées sur des caractéristiques génomiques : ces méthodes consistent à comparer des génomes selon le contenu en gènes, l'ordre des gènes, l'occurrence de certaines oligonucléotides et les changements génomiques rares.

D'après Delsuc, Brinkmann et Philippe (2005)

2. Les eucaryotes

On distingue deux types d'organismes dans le monde vivant : les procaryotes (bactéries et archées) et les eucaryotes (Doolittle, 1998). Les différences entre eux sont remarquables. Les cellules eucaryotes ont en général des structures plus complexes et contrastent avec les cellules procaryotes par la présence d'un 'vrai' (*eu*) noyau (*caryon*) entouré d'une double membrane et contenant l'ADN. D'autres structures différencient les eucaryotes des procaryotes : le réticulum endoplasmique, l'appareil de Golgi et le cytosquelette. Un génome de plus grande taille et de structure plus complexe, la division cellulaire par mitose ou par méiose et la capacité de générer des formes multicellulaires différentiées sont d'autres caractéristiques propres aux eucaryotes.

Parmi les organites distinctifs des eucaryotes, sont à noter les mitochondries et les chloroplastes, chacun entouré d'une double membrane. Les mitochondries (ou les organites qui en dérivent) sont universelles à toutes les cellules eucaryotes, tandis que les chloroplastes se retrouvent seulement chez quelques-uns, par exemple, les plantes. La théorie de l'endosymbiose en série, qui postule que ces organites résultent de l'association entre des bactéries endosymbiontes et une cellule hôte (Margulis, 1970; Sagan, 1967; Taylor, 1974), est le modèle le plus accepté pour expliquer leur origine (Cavalier-Smith, 1987b; 1989; Gray, 1992; 1993).

2.1. Les organites d'origine endosymbiotique

2.1.1. La mitochondrie

La mitochondrie est un organite entouré de deux membranes qui contient son propre matériel génétique. Dans la majorité des cellules eucaryotes, elle est utilisée pour générer de l'ATP grâce à la phosphorylation oxydative (Saraste, 1999), mais, mis à part son rôle dans la production d'énergie, elle participe aussi à d'autres fonctions comme le métabolisme intermédiaire ou l'apoptose. Il est maintenant bien accepté que la mitochondrie dérive d'une α-protéobactérie, hypothèse d'abord basée sur des considérations biochimiques (John et Whatley, 1975) et confirmée plus tard par l'analyse

phylogénétique de la petite sous unité de l'ARN ribosomal (PSU ARNr) mitochondriale (Woese, 1987).

Le nombre de gènes contenus dans le génome mitochondrial, entre 5 (chez les apicomplexes) et 94 (chez le jakobids), est très réduit par rapport à celui de son ancêtre α-protéobactérien (Lang, Gray et Burger, 1999). En effet, en évaluant à 1 000 le nombre de gènes du génome pré-mitochondrial, une quantité importante a dû être perdue pendant l'évolution de la mitochondrie (Figure 24). Trois causes principales expliquent cette réduction : (i) la perte des gènes non nécessaires à une vie intracellulaire (p. ex., ceux codant pour des protéines impliquées dans la synthèse de nucléotides, d'acides aminés et de lipides); (ii) la perte des gènes redondants par rapport au génome nucléaire –les produits codés dans le noyau étant utilisés pour les fonctions mitochondriales (p. ex., les ARN de transfert); et (iii) la migration de certains gènes vers le noyau –leurs produits étant ensuite importés dans la mitochondrie (p. ex., gènes codant pour des complexes impliqués dans la phosphorylation oxydative).

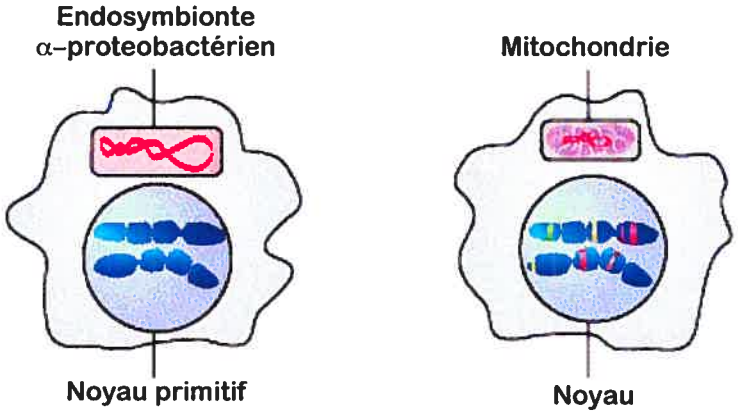

Figure 24 : Réduction du génome mitochondrial

Pendant la transition de bactérie à mitochondrie, le symbionte α-protéobactérien souffre d'une perte massive de gènes, dont certains sont transférés au noyau.

Modifié d'après Burger, Gray et Lang (2003)

Plusieurs groupes d'eucaryotes dits « amitochondriés » ne possèdent pas d'organites qui répondent à la description classique de la mitochondrie. En revanche, ils possèdent des organites dérivés de celle-ci (Boxma *et al.*, 2005; Embley *et al.*, 2003) : soit des hydrogénosomes, des organites entourés de deux membranes produisant de l'ATP et de l'hydrogène (Lindmark et Muller, 1973), soit des mitosomes, des organites cryptiques de fonction inconnue qui contiennent des protéines mitochondriales (Tovar, Fischer et Clark, 1999). Confirmant que l'ancêtre de ces organismes possédait une mitochondrie, des gènes codant pour des protéines d'origine mitochondriale ont été trouvés dans leurs génomes nucléaires (Roger et Silberman, 2002). Ces observations conduisent à la conclusion qu'il n'existe pas d'eucaryote dont l'ancêtre ne possédait pas de mitochondrie, impliquant que l'origine de la mitochondrie a eu lieu avant la divergence de tous les eucaryotes connus à ce jour (Burger, Gray et Lang, 2003).

2.1.2. Le chloroplaste

2.1.2.1. Les plastes primaires, secondaires et tertiaires

Les chloroplastes (ou plastes[8]) sont des organites, généralement photosynthétiques, présents dans certains types d'eucaryotes. Comme la mitochondrie, les plastes dérivent d'une bactérie, plus précisément d'une cyanobactérie (McFadden, 2001; Mereschkowsky, 1905), et leur génome est réduit par rapport à leur ancêtre (Timmis *et al.*, 2004). Les plastes sont présents chez une grande variété d'eucaryotes, en particulier, les plantes terrestres, les algues vertes, rouges et brunes, les diatomées, les dinoflagellés et même certains parasites humains comme, par exemple, *Plasmodium falciparum*, agent responsable de la malaria. Due à cette extraordinaire diversité, l'inférence de l'origine évolutive du plaste ainsi que de son histoire à travers les eucaryotes photosynthétiques modernes s'avère difficile (voir Archibald, 2005 pour une revue récente).

L'ensemble des eucaryotes portant des plastes se divise en deux groupes : ceux qui dérivent de l'endosymbiose primaire entre un hôte eucaryote et une cyanobactérie et ceux qui dérivent des endosymbioses secondaires ou tertiaires, impliquant l'association entre

[8] Le terme plaste sera utilisé de préférence pour désigner les chloroplastes en général, alors que le terme chloroplaste sera réservé seulement aux chloroplastes des plantes vertes.

une algue eucaryote contenant un plaste et un hôte eucaryote (Figure 25). Les plastes peuvent donc être primaires, secondaires ou tertiaires, et ils se distinguent par des caractéristiques structurales, particulièrement par le nombre de membranes qui les entourent (Moreira et Philippe, 2001).

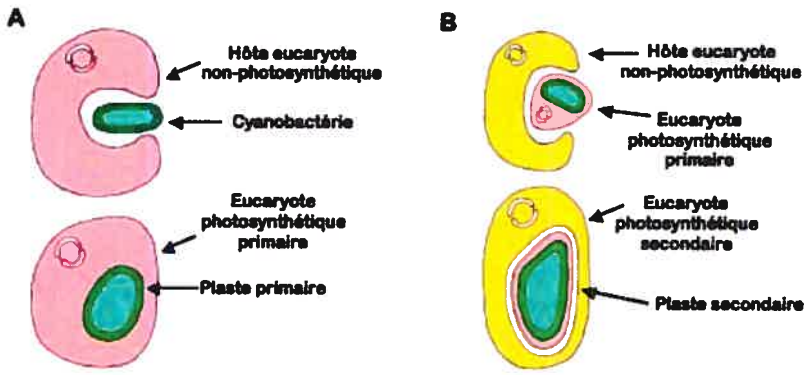

Figure 25 : Endosymbiose primaire et endosymbiose secondaire

A : Endosymbiose entre un eucaryote et une cyanobactérieà pour donner lieu à un photosynthétique primaire contenant un plaste primaire entouré de deux membranes. **B** : Endosymbiose entre un eucaryote hétérotrophe et un photosynthétique primaire pour former un plaste secondaire entouré de quatre membranes.

Modifié d'après Archibald et Keeling (2002)

Parmi les lignées eucaryotes connues, seulement les plantes vertes (incluant les plantes terrestres et les algues vertes), les algues rouges et les glaucophytes possèdent des plastes primaires, tandis que les plastes de toutes les autres lignées ont été acquis par des endosymbioses secondaires ou tertiaires. Il existe sept grands groupes d'eucaryotes contenant des plastes secondaires ou tertiaires. Ceux qui contiennent des plastes dérivés des algues vertes sont les euglénoïdes et les chlorarachniophytes et ceux qui contiennent des plastes dérivés des algues rouges sont les cryptomonadines, les straménopiles, les haptophytes, les dinoflagellés et les apicomplexes (Figure 26). Il est accepté que les euglénoïdes et les chlorarachniophytes, qui ne sont pas phylogénétiquement reliés, ont acquis leurs plastes par deux endosymbioses secondaires indépendantes (Ishida *et al.*, 1997; Keeling, 2001). Toutefois, un grand débat entoure la question du nombre

d'endosymbioses secondaires qui auraient donné lieu aux organismes possédant des plastes secondaires rouges (Moreira et Philippe, 2001; Palmer, 2003).

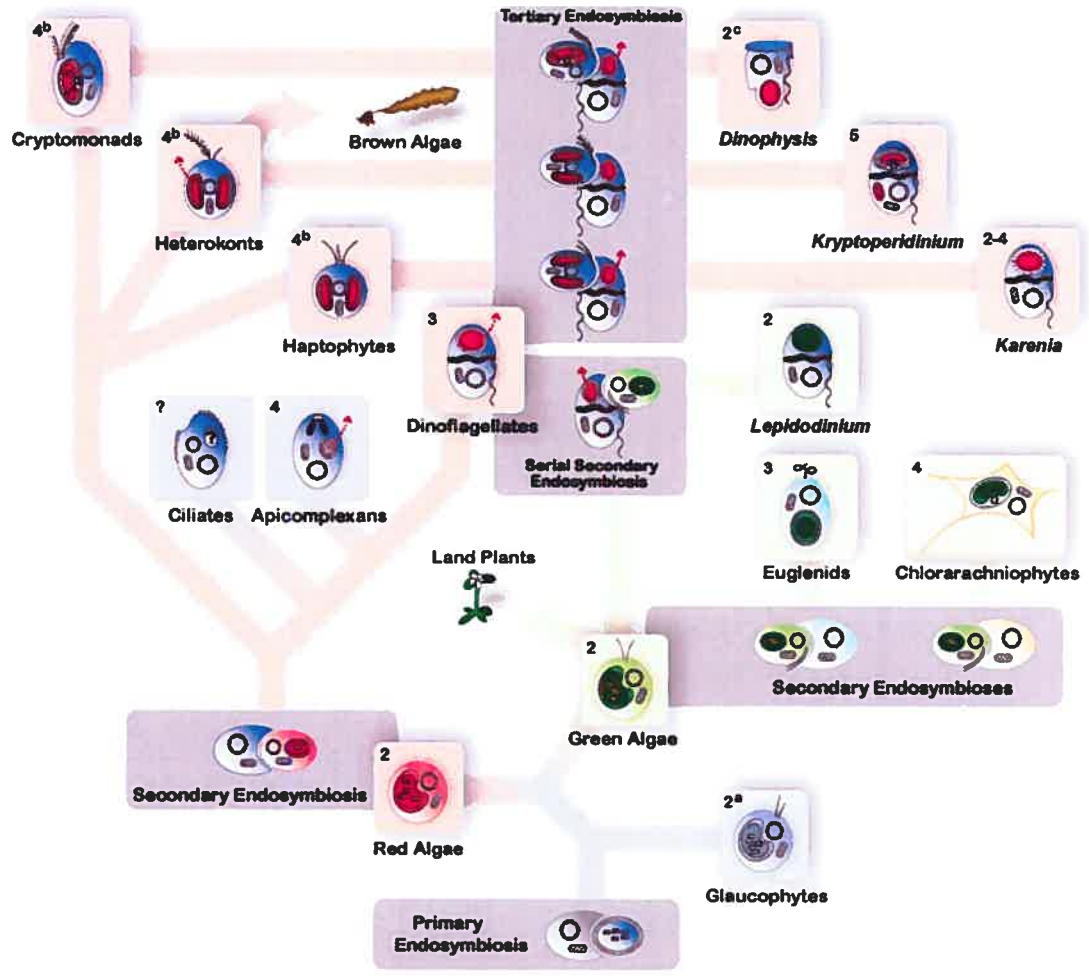

Figure 26 : Hypothèse sur l'origine et l'évolution des plastes

Une seule endosymbiose primaire aurait donné lieu aux plantes vertes, aux algues rouges et aux glaucophytes, et les plastes verts et rouges des autres lignées auraient été acquis par des endosymbioses secondaires ou tertiaires. Dans ce schéma, une seule endosymbiose secondaire rouge et deux vertes sont supposées, mais d'autres hypothèses existent (voir texte). Les lignées qui auraient dû perdre leur plaste selon cette hypothèse sont indiquées par des points d'interrogation. Les flèches rouges indiquent des pertes de plastes dans certains (lignes discontinues) ou tous (lignes continues) les organismes du groupe et les numéros dans le coin en haut à gauche des carrés indiquent le nombre de membranes entourant le plaste de l'organisme correspondant.

D'après Archibald (2005)

2.1.2.2. Combien d'endosymbioses primaires à l'origine des plastes?

Bien que l'origine des plastes des plantes vertes, des algues rouges et des glaucophytes par endosymbiose primaire avec une cyanobactérie ne soit pas discutée, il est encore pour déterminer si ces trois types de plastes dérivent de la même endosymbiose primaire ou s'ils dérivent de plusieurs endosymbioses indépendantes. La première hypothèse implique que les plastes sont monophylétiques et que leur origine remonte à une seule cyanobactérie, tandis que la deuxième implique des cyanobactéries différentes à l'origine des plastes de chacune des lignées. Plusieurs caractères communs aux plastes mais absents des cyanobactéries ont été proposés pour supporter une origine commune des plastes : par exemple, la présence de deux opérons, *psbB-psbN-psbH* et *atp-rps-rpo* (Reith et Munholland, 1993; Stoebe et Kowallik, 1999), et d'une séquence répétée inversée (Turmel, Otis et Lemieux, 1999), ainsi que la similarité en contenu en gènes des génomes plastiques (Martin *et al.*, 1998). Cependant, la validité de ces caractères a été mise en doute alléguant que ces similarités ne sont que des convergences évolutives (Palmer, 2003; Stiller, Reel et Johnson, 2003). Deux caractères additionnels, qui semblent être des inventions post-endosymbiotiques communes aux plantes vertes et aux algues rouges, apportent des évidences plus solides pour la monophylie des plastes : une protéine de triple hélice de liaison à la chlorophylle (Durnford *et al.*, 1999) et les composants TIC110 et TOC34 de l'appareil d'import des protéines nucléaires vers le plaste (McFadden et van Dooren, 2004). Ces trois protéines sont codées dans le noyau; leur présence chez les glaucophytes n'a donc pu être testée par manque de séquences du génome nucléaire de ce groupe.

La majorité des arbres phylogénétiques obtenus à partir des données du plaste sont consistants avec l'hypothèse d'une seule endosymbiose primaire. La seule exception est le gène *rbcL,* codant pour la RUBISCO (ribulose-1,5-bisphosphate carboxylase/oxygénase), qui supporte la relation entre les plastes rouges et les protéobactéries et celle des autres plastes avec les cyanobactéries. Cependant, il est maintenant connu que le comportement de ce gène est dû à des duplications et à des transferts horizontaux entre les différentes lignées (Delwiche et Palmer, 1996). Des analyses de la PSU ARNr incluant un vaste échantillonnage de cyanobactéries (Bhattacharya et Medlin, 1995; Turner *et al.*, 1999) supportent fortement la monophylie des plastes (valeur de bootstrap [VB] = 94-100%). Ce résultat est aussi obtenu avec quelques gènes codant pour des protéines comme *tufA*

(Delwiche, Kuhsel et Palmer, 1995), *atpB* (Douglas et Murphy, 1994) et la concaténation des gènes *rpoB, rpoC1* et *rpoC2* (Cai *et al.*, 2003) (VB = 53-100%). Semant des doutes quant à la monophylie des plastes, il a été suggéré que ces résultats sont dus à des artefacts d'inférence phylogénétique causés par le contenu similaire en A+T des génomes plastiques (Lockhart *et al.*, 1992) et/ou par un processus d'évolution de type covarion (Lockhart *et al.*, 1998). Malgré la disponibilité de génomes complets de plusieurs plastes et cyanobactéries, des analyses phylogénomiques destinées à tester la monophylie des plastes n'ont pas encore été réalisées. Pourtant, ce type d'analyse pourrait être cruciale pour tester la congruence entre les différents marqueurs et pour la détection du signal non phylogénétique.

2.2. Vue historique sur la classification des eucaryotes

2.2.1. La classification selon des caractères morphologiques

Dès le XVIIIème siècle, des caractéristiques morphologiques furent utilisées pour classifier les eucaryotes (Linnaeus, 1758). Les premières classifications distinguaient les organismes motiles, les animaux, de ceux sans capacité de déplacement, les plantes (incluant les algues, les mousses et les champignons), et groupaient toute la diversité des eucaryotes unicellulaires dans un ensemble appelé Protozoa (Owen, 1859), Protoctista (Hogg, 1860) ou Protista (Haeckel, 1866). Pour illustrer l'esprit de l'époque, voici la description d'une telle classification selon Richard Owen (1859) :

> *« When the organism can also move, receive the nutritive matter by a mouth into a stomach, inhale oxygen and exhale carbonic acid, develop tissues the proximate principles of which are quaternary compounds of carbon, hydrogen, oxygen, and nitrogen, it is called an "animal." When the organism is rooted, has no mouth or stomach, exhales oxygen, has tissues composed of "cellulose" or of binary or ternary compounds, it is called a "plant." But the two divisions of organisms called "plants" and "animals" are specialized members of the great natural group of living things, and there are numerous organisms, mostly of minute size and retaining the form of nucleated cells, which manifest the common organic characters, but without the distinctive super-additions of true plants or animals. Such organisms are called "Protozoa," and include the sponges or Amorphozoa, the Foraminifera or Rhizopods, the Polycystineae, the Diatomaceae, Desmidiae, and most of the so-called Polygastria of Ehrenberg, or infusorial animalcules of older authors. »*

En 1969, Robert H. Whittaker proposa sa classification en cinq royaumes, qui est considérée comme la dernière synthèse de l'ère « prémoléculaire » (Adoutte *et al.*, 1996). Selon cette classification, on distingue deux grands groupes, les procaryotes (Monera) et les eucaryotes. Parmi ces derniers, on distingue les protistes, desquels dérivent les plantes, les champignons et les animaux, classés selon leur mode de nutrition, soit l'autotrophie, l'absorption ou l'ingestion (Figure 27). Il est intéressant de remarquer que, dans cette classification, les champignons forment un royaume à part; Whittaker rejeta donc la pensée commune qui dictait que les champignons et les plantes sont apparentés dû à l'absence de motilité et à la possession de parois cellulaires. De plus, dans ce modèle, les protistes occupent déjà une position clé, entre les procaryotes et les eucaryotes multicellulaires, suggérant que leur étude pourrait éclairer des questions comme l'origine de la cellule eucaryote et des grands groupes multicellulaires (Adoutte *et al.*, 1996).

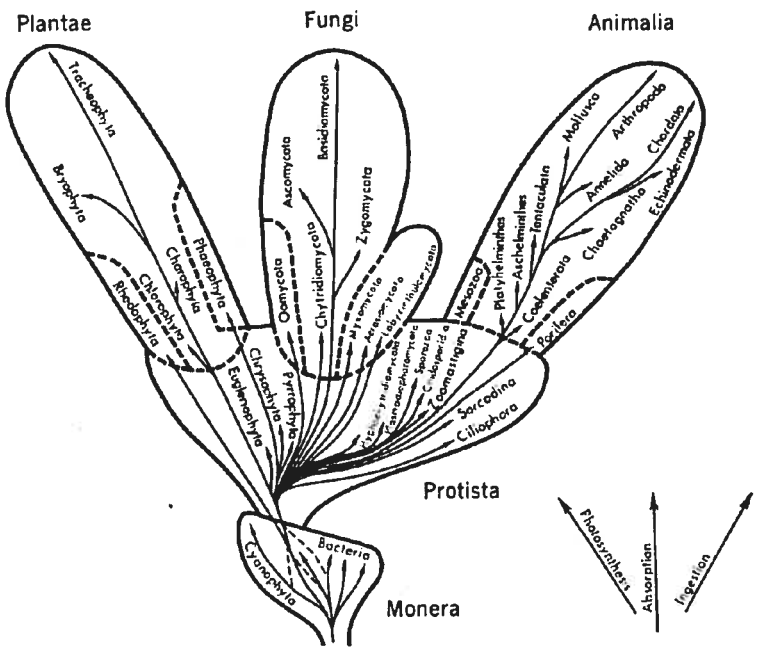

Figure 27 : La classification en cinq royaumes de Whittaker

Les cinq royaumes (Plantae, Fungi, Animalia, Protista et Monera) sont organisés selon leur complexité (du bas vers le haut) et selon leur mode de nutrition (de gauche à droite).

D'après Whittaker (1969)

La première classification à l'intérieur des protistes date du XIXème siècle et reconnaissait quatre grands groupes : les sarcodines (ou rhizopodes), les mastigophores (porteurs de flagelles), les infusoires (ciliés actuels) et les sporozoaires (des parasites intracellulaires). Plus tard, des études de microscopie électronique menèrent à la reconnaissance d'un nombre plus important de groupes de protistes, distinguant déjà des ensembles monophylétiques comme les dinoflagellés, les euglénides, les diatomées ou les apicomplexes (Adoutte *et al.*, 1996). En 1984, Corliss (Corliss, 1984) reconnaissait 45 phylums de protistes et ce nombre a augmenté considérablement jusqu'à une centaine de groupes ayant une identité ultrastructurale propre (Patterson, 1994).

2.2.2. Les premières phylogénies moléculaires

2.2.2.1. Phylogénies basées sur l'ARNr : l'hypotèse Archezoa

Les études comparatives de données morphologiques furent utiles pour différencier les grands groupes d'eucaryotes, mais pas suffisantes pour définir les relations entre les groupes et à l'intérieur de ceux ci. Un grand espoir fut donc placé dans les phylogénies moléculaires, basées initialement sur la PSU ARNr. Dans l'arbre des eucaryotes (raciné avec les procaryotes) obtenu à partir de cette molécule (Sogin, 1989; Sogin *et al.*, 1989; Vossbrinck *et al.*, 1987), on distinguait deux parties : la « couronne », qui contenait les grands groupes –plantes, animaux, champignons, straménopiles, alvéolés et algues rouges–sans que les relations entre eux soient résolues, et la « base », avec le branchement paraphylétique de plusieurs groupes de protistes. Parmi ceux-ci, les plus basales étaient les diplomonadines, les trichomonadines et les microsporidiens, tous des eucaryotes sans mitochondrie (Figure 28).

Le regroupement basal de ces trois lignées d'eucaryotes amitochondriés mena à confirmer l'hypothèse Archezoa (Cavalier-Smith, 1983), selon laquelle plusieurs groupes d'eucaryotes appelés Archezoa (diplomonadines, trichomonadines, microsporidiens et archamibes) émergèrent avant l'origine de la mitochondrie et représentaient donc des reliques vivantes d'une période amitochondriée de l'évolution des eucaryotes (Figure 28). De plus, comme attendu pour des eucaryotes primitifs, ces amitochondriés ont une

ultrastructure simple, ne possèdent pas certains organites comme les peroxisomes et ont un système d'endomembranes peu développé.

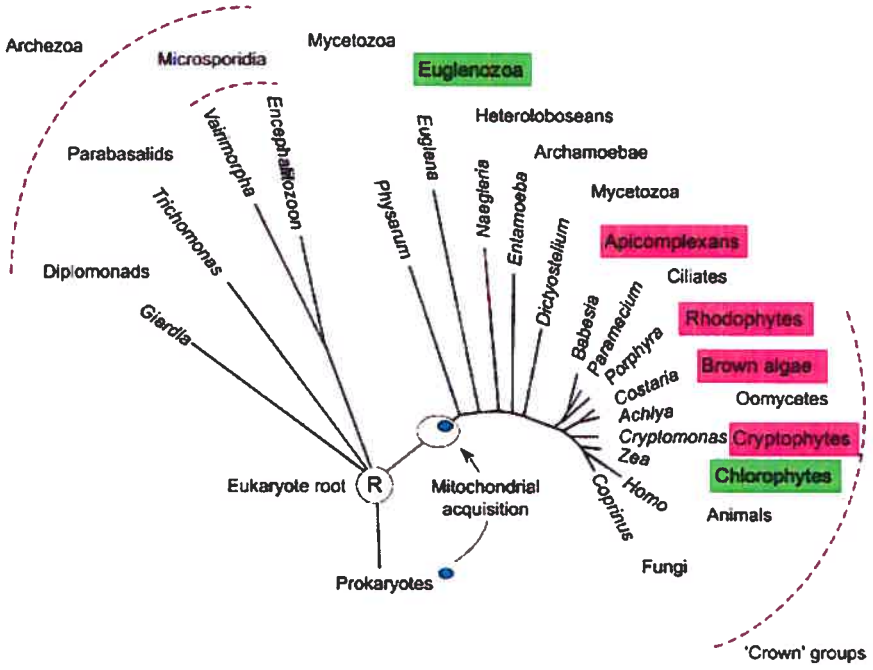

Figure 28 : Phylogénie des eucaryotes basée sur l'ARNr

Les lignées situées près de la racine de l'arbre (diplomonadines, trichomonadines et microsporidiens) furent considérées des eucaryotes amitochondriés primitifs et appelés Archezoa. Les lignées les plus éloignées de la racine furent considérées les plus dérivées et appelées « groupes de la couronne ». Les couleurs indiquent les organismes qui possèdent des plastes rouges ou verts.

D'après Embley et Martin (2006)

Bien qu'apparemment supportée par la morphologie et les données moléculaires, l'hypothèse Archezoa fut vite réfutée par d'autres évidences qui suggéraient une perte secondaire de la mitochondrie chez les amitochondriés (Keeling, 1998; Roger, 1999). Parmi celles-ci on trouve (i) la position non basale de quelques uns des Archezoa (p. ex., l'archamibe *Entamoeba histolytica*) et (ii) la présence de protéines d'origine mitochondriale (p. ex., la protéine de choc thermique HSP70 et la chaperonne Cpn60) chez les archamibes (Clark et Roger, 1995), les trichomonadines (Germot, Philippe et Le Guyader, 1996; Roger, Clark et Doolittle, 1996), les diplomonadines (Roger *et al.*, 1998) et

les microsporidiens (Germot, Philippe et Le Guyader, 1997; Horner *et al.*, 1996). De plus, il a été demontré que ces protéines sont localisées dans des organites dérivés de la mitochondrie (Bui, Bradley et Johnson, 1996; Tovar, Fischer et Clark, 1999). Toutes ces caractéristiques supportent l'hypothèse d'une simplification secondaire chez les Archezoa, leur ancêtre possédant déjà une mitochondrie (Lang, Gray et Burger, 1999) (mais voir aussi Sogin, 1997).

2.2.2.2. Artefacts d'inférence phylogénétique : l'hypothèse du « big bang »

La perte secondaire de la mitochondrie chez les Archezoa n'est pas contradictoire avec la position basale de ces organismes. Cependant, des positions alternatives (souvent mieux supportées) pour ces espèces ont été proposées. En particulier, des arbres inférés à partir de la tubuline (Keeling et Doolittle, 1996) et de l'ARN polymérase II (Hirt *et al.*, 1999) supportent le placement des microsporidiens dans les champignons. Ceci est aussi confirmé par une insertion unique aux champignons, animaux et microsporidiens dans le facteur d'élongation 1α (EF1α) (Baldauf et Doolittle, 1997). Les microsporidiens sont donc des champignons qui, dû à leur style de vie en tant que parasites intracellulaires, ont souffert de simplifications secondaires au niveau de leur ultrastructure et d'une réduction importante de la taille de leur génome (seulement 2,9 Mpb) (Keeling et McFadden, 1998).

Suite à ces observations, il est possible de conclure que le placement basal de certains organismes dans la phylogénie de la PSU ARNr pourrait être causé par des artefacts de reconstruction phylogénétique (Embley et Hirt, 1998). De plus, comme le groupe extérieur utilisé pour enraciner l'arbre des eucaryotes (les procaryotes) est très éloigné, un artefact d'attraction des longues branches entre celui-ci et les eucaryotes avec le plus grand taux d'évolution est attendu (Philippe, Germot et Moreira, 2000). La comparaison de phylogénies basées sur plusieurs marqueurs (ARNr, actine, tubuline, EF1α, etc) supporte cette hypothèse. En effet, comme pour l'ARNr, la majorité des analyses dérivées de gènes codant pour des protéines résultait dans un arbre divisé en deux parties, la couronne et la base. Curieusement, les espèces à la base n'étaient pas les mêmes pour tous les marqueurs. Par exemple, dans les phylogénies basées sur l'ARNr et sur l'actine, les ciliés se positionnent respectivement dans la couronne ou à la base de l'arbre, alors que le phénomène contraire était observé pour les euglénozoaires. Effectivement, il semble que

dans chaque cas les organismes avec le plus haut taux d'évolution pour le gène analysé sont attirés vers la base de l'arbre (Philippe et Adoutte, 1998).

Ces contradictions entre les différents marqueurs peuvent être expliquées par l'hypothèse du *« big bang »* (Philippe et Adoutte, 1998), qui postule que les grands groupes eucaryotes ont divergé à l'intérieur d'un court laps de temps, rendant l'inférence des relations phylogénétiques entre eux difficile ou même impossible. Selon cette hypothèse, tous les groupes d'eucaryotes font partie de la « couronne ». La position basale de certains d'entre eux serait due à des artefacts d'inférence phylogénétique, comme l'attraction des longues branches (Philippe, Germot et Moreira, 2000).

2.2.3. Phylogénies basées sur plusieurs gènes concaténés

Les analyses basées sur une petite quantité de données (un seul gène) ne contiennent pas assez d'information phylogénétique pour résoudre la plupart des relations entre les grands groupes d'eucaryotes, et des études basées sur plusieurs gènes simultanément sont nécessaires. Des analyses de ce type commencèrent à être utilisées dans les années 90, mais dû au manque de données moléculaires disponibles à l'époque, elles n'incluaient que quatre ou cinq espèces (Kuma *et al.*, 1995; Nikoh *et al.*, 1994). Un peu plus tard, des résultats plutôt encourageants furent trouvés par des analyses basées sur la concaténation de plusieurs gènes mitochondriaux ou nucléaires (entre 4 et 13) avec un échantillonnage taxonomique plus intensif (Baldauf *et al.*, 2000; Burger *et al.*, 1999; Moreira, Le Guyader et Philippe, 2000). Entre autres, ces études supportaient des groupes comme les clusters « animaux + champignons » et « plantes vertes + algues rouges ».

La hausse considérable de données moléculaires disponibles pour beaucoup de groupes eucaryotes a mené depuis le début de la présente décennie à un nombre croissant d'analyses basées sur une multitude de gènes (des analyses phylogénomiques). En particulier, ce type d'analyses s'est avéré efficace pour résoudre les relations phylogénétiques entre les angiospermes (Qiu *et al.*, 1999; Soltis, Soltis et Chase, 1999) et entre les mammifères (Madsen *et al.*, 2001; Murphy *et al.*, 2001). À plus grande échelle dans l'évolution des eucaryotes, la phylogénomique a confirmé la monophylie des grands phylums supportés par de caractères morphologiques, corroboré la relation longtemps

suspectée entre les choanoflagellés et les animaux (King et Carroll, 2001; Lang *et al.*, 2002; Philippe *et al.*, 2004), et même proposé des ensembles d'organismes d'apparence très diverse (Bapteste *et al.*, 2002). L'état actuel des connaissances sur la phylogénie des eucaryotes, incluant les contributions de la phylogénomique, est résumé dans la section suivante.

2.3. Vue actuelle de la classification des eucaryotes

Comme il a été dit précédemment, l'arbre des eucaryotes « base + couronne » obtenu à partir de l'ARNr dans les années 80 a subi une période de déconstruction durant les années 90.

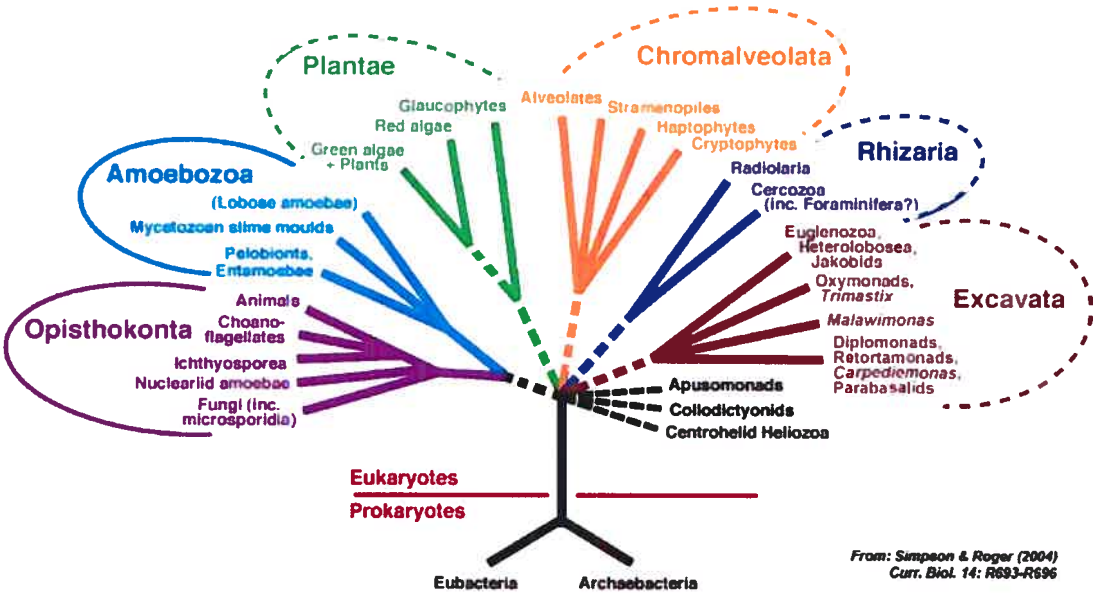

Figure 29 : Hypothèse actuelle sur l'arbre des eucaryotes

Représentation schématique de la classification des eucaryotes en six super ensembles. Les lignes discontinues représentent les ensembles supportés uniquement par des indications préliminaires. Plusieurs groupes d'eucaryotes microscopiques peu étudiés (indiqués en noir) n'ont pas d'affinités avec aucun des six ensembles et semblent former des groupes distincts.

Modifié d'après Simpson et Roger (2004b)

On est maintenant dans une période de reconstruction de l'arbre des eucaryotes grâce (i) à la hausse des données moléculaires disponibles pour une grande variété de groupes, (ii) à l'amélioration des modèles d'évolution de séquences et (iii) à une plus grande puissance de calcul disponible. Nous décrivons ci-dessous l'état actuel de la classification des eucaryotes (Figure 29) en distinguant les groupes dits « solides » de ceux pour lesquels des études plus approfondies sont nécessaires.

2.3.1. L'ensemble Opisthokonta

Le groupe Opisthokonta (Cavalier-Smith, 1987a) comprend les animaux, les champignons et quelques protistes comme les choanoflagellés, les Ichthyosporea et les Nucleariidae (Figure 30). Dès le XIXéme siècle, la relation phylogénétique entre les animaux et les choanoflagellés fut suggérée en se basant sur la similarité de ces derniers avec les choanocytes, une sorte de cellules spécialisées des éponges (James-Clark, 1866). La présence d'un seul flagelle postérieur dans le sperme de la majorité des animaux, dans les spores de plusieurs champignons, ainsi que dans les différents groupes de protistes inclus dans cet ensemble semble être le seul caractère morphologique qui supporte la monophylie des opisthocontes (Cavalier-Smith et Chao, 1995). En revanche, ce groupe apparaît très solide dans les phylogénies moléculaires (p. ex. Baldauf, 1999; Baldauf *et al.*, 2000; Bapteste *et al.*, 2002; Philippe *et al.*, 2004) et est aussi supporté par des insertions dans le EF1α et l'énolase (Baldauf et Palmer, 1993). Les relations entre animaux, champignons et protistes qui composent le groupe restent encore à définir.

2.3.2. L'ensemble Amoebozoa

L'ensemble Amoebozoa (Lühe, 1913) comprend la majorité des protistes se déplaçant et se nourrissant à l'aide de pseudopodes (Figure 30). Il inclut des organismes très distincts comme les classiques amibes du genre *Amoeba*, les myxomycètes, avec une phase amiboïde et une phase multicellulaire de sporulation (p. ex. *Dictyostelium discoideum*), et des parasites amitochondriés comme *Entamoeba histolytica*, qui cause plus de 50,000 morts humaines chaque année. À cause de la phase multicellulaire des myxomycètes, acquérant la forme d'un champignon doté d'une tête constituée d'amibes pleines de spores,

ces organismes furent considérés des champignons, tandis que les autres amoebozoaires furent considérés des animaux. En fin, les évidences moléculaires solides pour la monophylie des Amoebozoa sont apparues récemment (Bapteste *et al.*, 2002; Gray, Lang et Burger, 2004; Brinkmann, Gray et Philippe, communication personnelle).

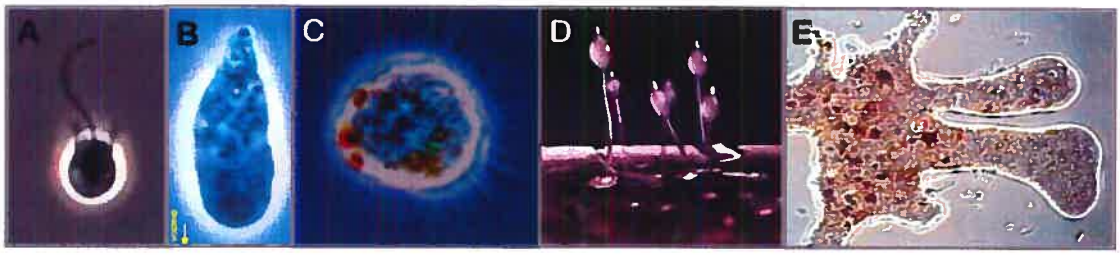

Figure 30 : Exemples d'opisthocontes unicellulaires et d'amoebozoaires
A : *Monosiga ovata* (choanoflagellé). **B :** *Amoebidium parasiticum* (Ichthyosporea). **C :** *Nuclearia simplex* (Nucleariidae), **D :** *Dictyostelium discoideum* (Amoebozoa), **E :** *Amoeba proteus* (Amoebozoa).
Source : http://microscope.mbl.edu/

2.3.3. L'hypothèse Plantae

Le royaume Plantae (Cavalier-Smith, 1981), aussi appelé Archeplastida (Adl *et al.*, 2005), comprend les trois lignées des eucaryotes photosynthétiques primaires, c'est-à-dire ceux qui possèdent des plastes entourés de deux membranes et dérivés directement de l'endosymbiose avec une cyanobactérie : les plantes vertes, les algues rouges et les glaucophytes (Figure 31).

Les **plantes vertes** (Viridiplantae) constituent l'ensemble formé des plantes terrestres et des algues vertes et occupent une grande diversité d'habitats comme les océans, l'eau douce, le sol et même la neige. Leurs plastes sont colorés par les chlorophylles *a* et *b* et c'est grâce à cette caractéristique que la relation de parenté entre les plantes terrestres et les algues vertes a été proposée (Bower, 1908), observation confirmée plus tard par des analyses phylogénétiques (Graham, Delwiche et Mishler, 1991).

Figure 31 : Exemples d'organismes du groupe Plantae

A : *Boxus sempervirens* (plante terrestre). **B :** *Chlorella vulgaris* (algue verte). **C :** *Porphyra sp.* (algue rouge), **D :** *Porphyridium purpureum* (algue rouge), **E :** *Cyanophora paradoxa* (glaucophyte). **F :** *Glaucocystis nostochinearum* (glaucophyte).

Sources : http://www.biol.tsukuba.ac.jp, http://www.botany.wisc.edu, http://www.ucmp.berkeley.edu

Les **algues rouges** (Rhodophyta) prédominent dans les environnements marins, mais peuvent aussi exister dans l'eau douce et même dans des milieux plus hostiles comme des eaux thermales ou des salines. Leurs plastes sont colorés par la chlorophylle *a* et par des phycobiliprotéines organisées en phylcobilisomes, soit des complexes pigment-protéine collecteurs d'énergie lumineuse attachés à la surface de la membrane thylakoïdale. Par cette organisation des thylakoïdes, les plastes des algues rouges (et des glaucophytes; voir plus bas) ressemblent aux cyanobactéries plus que ceux des plantes vertes (Delwiche, 1999).

Les **glaucophytes** (Glaucophyta ou Glaucocystophyta) forment un petit groupe d'algues unicellulaires d'eau douce peu étudié dont les plastes sont aussi colorés par la chlorophylle *a* et par des phycobilisomes. Les plastes des glaucophytes, aussi appelés cyanelles, se distinguent de ceux des deux autres lignées par la présence (i) d'une paroi de peptidoglycane entre les membranes interne et externe qui est le restant de la paroi cellulaire du type gram-négative trouvée chez les cyanobactéries (Löffelhardt et Bohnert, 1994) et (ii) des carboxysomes, des corps polyédriques impliqués dans la fixation de CO_2 présents aussi chez les cyanobactéries (Kaplan et Reinhold, 1999). Ces deux structures d'origine cyanobactérienne et absentes chez les plastes rouges et verts sont à l'origine de l'idée erronée que les cyanelles sont en fait des cyanobactéries non réduites, alors qu'il est connu que le génome du glaucophyte *Cyanophora paradoxa* est aussi réduit que celui des autres plastes (Stirewalt *et al.*, 1995). Toutefois, ces deux caractéristiques ancestrales de leurs cyanelles placent les glaucophytes dans une position clé dans l'évolution des plastes.

En effet, si les plastes primaires sont monophylétiques, il est alors logique que les glaucophytes aient une position basale, minimisant ainsi le nombre de pertes de la paroi de peptidoglycane et des carboxysomes.

L'hypothèse Plantae (Cavalier-Smith, 1981) postule pour une endosymbiose primaire expliquant l'origine des plastes, ce qui implique la monophylie des eucaryotes photosynthétiques primaires ainsi que celle de leurs plastes et dérivés. La monophylie des plastes est supportée par plusieurs caractères et est acceptée par la communauté scientifique (voir section 2.1.2); cependant, à part leurs plastes, il n'existe pas de caractère commun aux algues rouges, aux plantes vertes et aux glaucophytes.

Les premières phylogénies incluant des séquences de plantes vertes, d'algues rouges et de glaucophytes, basées sur la PSU ARNr, la tubuline ou l'actine (Bhattacharya *et al.*, 1995; Bhattacharya et Weber, 1997; Keeling *et al.*, 1999), ne retrouvaient pas la monophylie du groupe, ce qui a mené à postuler une variété d'hypothèses impliquant des origines multiples des plastes pour les trois lignées. Cependant, même si ces analyses ne supportaient pas la monophylie des Plantae, elles ne la rejetaient pas non plus car aucune autre alternative n'était supportée. Il existe deux exceptions à ce manque de résolution : le facteur d'élongation 2 (EF2), qui supporte le groupe « algues rouges + plantes vertes » (Moreira, Le Guyader et Philippe, 2000) et l'ARN polymérase II (RPB1), qui rejette cette relation (Stiller et Hall, 1999). Pour concilier ces deux résultats, il a été proposé (i) que les analyses de la RPB1 sont affectées par des artefacts d'inférence phylogénétiques dû au pauvre échantillonnage taxonomique utilisé (Moreira, Le Guyader et Philippe, 2000) ou (ii) que le signal du EF2 est dû à une dizaine de sites qui supportent la relation algues rouges + plantes vertes par convergence évolutive (Stiller, Riley et Hall, 2001).

Deux analyses basées sur la concaténation de plusieurs gènes supportent la monophylie rouge + vert : la fusion de 13 gènes nucléaires incluant le EF2 et la RPB1 (Moreira, Le Guyader et Philippe, 2000), et la fusion de 4 gènes mitochondriaux (Burger *et al.*, 1999). Cependant, (i) l'exclusion du EF2 de la première analyse supporte un résultat alternatif, suggérant que ce gène est le seul responsable de cette relation (Nozaki *et al.*, 2003), et (ii) l'analyse d'un autre sous-ensemble de gènes du génome mitochondrial (*nad1-6* au lieu de *cox1-3*, *cob*) ne supporte pas la monophylie rouge + vert, suggérant que les données mitochondriales pourraient ne pas être adéquates pour résoudre ce genre de

questions (Stiller, Riley et Hall, 2001). En plus de ne pas donner de résultats concluants quant aux relations entre plantes vertes et algues rouges, l'ensemble de ces analyses ne donnent aucune information sur la position phylogénétique des glaucophytes. En effet, le jeu mitochondrial n'inclut aucun représentatif de ce groupe (Burger *et al.*, 1999; Stiller, Riley et Hall, 2001) et l'analyse d'un jeu de données de six protéines nucléaires incluant le glaucophyte *Cyanophora paradoxa*, ne résout pas leur position (Moreira, Le Guyader et Philippe, 2000).

2.3.4. L'hypothèse Chromalveolata

Les chromalvéolés sont définis comme les algues chromophytes, c'est-à-dire celles dont l'ancêtre possédait des plastes secondaires d'origine rouge avec chlorophylle *c*, ainsi que leurs descendants non photosynthétiques (Christensen, 1989). Ils représentent une grande fraction de la biodiversité des eucaryotes (Figure 32) et ils incluent des organismes photosynthétiques responsables de 70% de la production primaire des océans (haptophytes, dinoflagellés et diatomées), des organismes pathogènes de plantes (oomycètes), des parasites d'animaux (apicomplexes) et des phagotrophes de vie libre (ciliés). D'un point de vue taxonomique, on distingue deux groupes : les chromistes et les alvéolés.

Les chromistes (Cavalier-Smith, 1981) incluent les cryptophytes, les haptophytes et les straménopiles, et sont principalement définis par la localisation du plaste à l'intérieur du réticulum endoplasmique rugueux. Les **cryptophytes** forment un groupe d'algues unicellulaires abondantes qui se caractérisent par la rétention d'un vestige du noyau de l'algue rouge endosymbionte et appelé nucléomorphe. Les **haptophytes** sont des algues unicellulaires de grande importance écologique; la majorité est recouverte de plaques calcaires appelées coccolithes, composantes primaires des boues crayeuses. Les **straménopiles** (ou hétérocontes) constituent un groupe très diversifié qui inclut des organismes photosynthétiques comme les diatomées et les algues brunes (pouvant mesurer jusqu'à 50 mètres de hauteur), ainsi que des organismes pathogènes dépourvus de plaste comme l'oomycète *Phytophthora infestans*, qui attaque la pomme de terre et fut responsable de la grande famine en Irlande entre 1845 et 1849 (Archibald et Keeling, 2004; Cavalier-Smith, 2004).

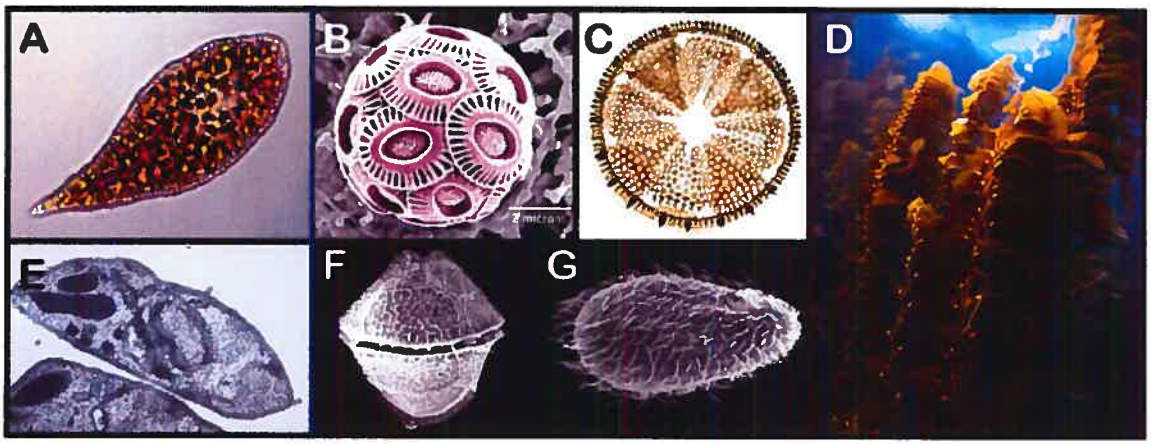

Figure 32 : Exemples d'organismes du groupe Chromalveolata

A : Cryptophyte. **B :** Haptophyte. **C :** Stramenopile (diatomée). **D :** Stramenopile (algue burne). **E :** Apicomplexa. **F :** Dinoflagellé. **G :** Cilié.

Sources : http://www.inria.fr, http://www.aecom.yu.edu, http://www.nielsbio.com, http://www.biocrawler.com

La monophylie des alvéolés (Cavalier-Smith, 1993) est solidement supportée par des phylogénies moléculaires (p. ex., Fast *et al.*, 2002; Harper, Waanders et Keeling, 2005; Wolters, 1991). Ce groupe inclut les dinoflagellés, les apicomplexes et les ciliés, tous trois caractérisés par la présence de vésicules sous-membranaires formant des sortes d'alvéoles. Les **dinoflagellés** sont des algues unicellulaires avec un plaste secondaire rouge à l'origine. Cependant, certaines espèces ont perdu leur plaste et soit sont restées hétérotrophes, soit ont effectué des endosymbioses secondaires avec une algue verte ou tertiaires avec un haptophyte, un stramenopile ou un cryptophyte (Figure 26). Certains sont connus pour causer les marées rouges et l'empoisonnement des mollusques. Les **apicomplexes** sont probablement le groupe de parasites avec le plus de succès sur terre et incluent les agents de la toxoplasmosis, de la cryptosporidiosis et de la malaria (*Plasmodium*). Malgré leur style de vie parasite, la plupart ont retenu un plaste non photosynthétique, témoin du passé photosynthétique de leur ancêtre. Les **ciliés** sont des prédateurs dépourvus de plaste, présents dans une grande diversité d'habitats et incluent des eucaryotes modèles comme *Paramecium* et *Tetrahymena* (Archibald et Keeling, 2004; Cavalier-Smith, 2004).

L'hypothèse Chromalveolata (Cavalier-Smith, 1998) postule une seule endosymbiose secondaire à l'origine des six groupes décrits ci-dessus (Figure 26), ce qui implique la perte

complète du plaste chez certaines lignées comme les oomycètes et les ciliés. Pour tester l'hypothèse Chromalveolata, la monophylie des plastes ainsi que celle des cellules hôtes doivent être obtenues. Des analyses phylogénétiques récentes basées sur la concaténation de cinq gènes du plaste supportent la monophylie des chromistes (Yoon *et al.*, 2002) et des analyses de gènes nucléaires d'origine plastique comme la glycéraldéhyde-3-phosphate déshydrogénase (GAPDH) (Harper et Keeling, 2003) et la fructose-bisphosphate aldolase (FBA) (Patron, Rogers et Keeling, 2004) supportent une origine commune des plastes rouges des dinoflagellés, alvéolés, stramenopiles et cryptophytes. Les données du plaste supportent donc la monophylie des chromalvéolés. En revanche, les phylogénies basées sur des gènes de l'hôte ne donnent pas de résultats concluants (Bhattacharya *et al.*, 1995; Harper, Waanders et Keeling, 2005). Par exemple, l'analyse de la concaténation de six protéines cytosoliques codées dans le noyau incluant tous les groupes de chromalvéolés ne retrouve pas leur monophylie. Cependant, plusieurs analyses phylogénomiques supportent la monophylie des alvéolés + straménopiles (e.g., (Philippe *et al.*, 2004)), qui sont les deux seuls groupes inclus dans ce genre d'analyse à cause du manque de données à échelle génomique pour les haptophytes et les pour cryptophytes. Des études phylogénomiques incluant des représentants de tous les groupes des chromalvéolés sont donc nécessaires pour comprendre la propagation de la photosynthèse chez les eucaryotes.

2.3.5. L'hypothèse Excavata

L'hypothèse Excavata (Cavalier-Smith, 2002; Simpson, 2003) est l'une des plus controversées dans la phylogénie des eucaryotes parce que (i) elle concerne le placement des eucaryotes supposés « primitifs » comme les Archezoa, et (ii) elle implique la monophylie d'un ensemble de protistes qui ne partagent aucun caractère commun à tout le groupe. L'hypothèse Excavata se base plutôt sur des similarités morphologiques et des analyses phylogénétiques qui supportent des sous-ensembles chevauchants formant un réseau qui unifie tout le groupe (Simpson, 2003).

2.3.5.1. D'Archezoa à Excavata

L'histoire de l'hypothèse Excavata est étroitement liée à celle de l'hypothèse Archezoa (exposée dans la section 2.2.2.1). Pendant l'époque de cette dernière, des comparaisons morphologiques montrèrent que certains groupes d'eucaryotes mitochondriés, soit les percolozoaires (Heterolobosea), les jakobides et les malawimonadines, partageaient des similarités ultrastructurales avec les deux lignées d'Archezoa diplomonadines et retortamonadines (Cavalier-Smith, 1992; O'Kelly, 1993; Sleigh, 1989). En particulier, ces espèces partagent un sillon nutritif ventral utilisé pour collecter des particules en suspension. Dans le cadre de l'hypothèse Archezoa, les percolozoaires, les jakobides et les malawimonadines furent considérés les eucaryotes mitochondriés les plus proches des eucaryotes amitochondriés et, par conséquent, de l'origine de la mitochondrie (Cavalier-Smith, 1992; O'Kelly, 1993).

Plus tard, avec la réfutation de l'hypothèse Archezoa dans les années 90, ces similarités ultrastructurales furent considérées, non pas comme des caractères ancestraux, mais plutôt comme des caractères partagés dérivés unifiant les diplomonadines, les retortamonadines, les percolozoaires, les jakobides et les malawimonadines dans un nouvel ensemble : les excavés (Simpson, 2003; Simpson et Patterson, 1999). Plus récemment, le groupe a été élargi (Figure 33) par l'ajout de *Trimaxtix* et de *Carpediemonads*, qui partagent les similarités ultrastucturales typiques des excavés (Simpson, Bernard et Patterson, 2000; Simpson et Patterson, 1999), ainsi des oxymonadines, des parabasaliens (p. ex. *Trichomonas*) et des euglénozoaires (Figure 33) qui, dans les phylogénies moléculaires, semblent être apparentés à l'un ou à plusieurs des groupes mentionnés ci dessus.

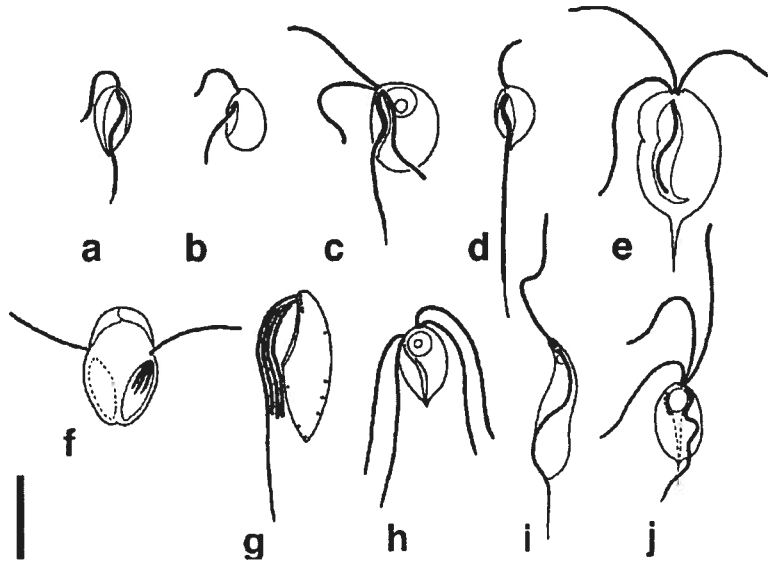

Figure 33 : Apparence des dix groupes d'excavés au microscopie optique
A : *Jakoba incarcerata* (jakobide). **B :** *Malawimonas jakobiformis.* **C :** *Trimastix pyriformis.* **D :** *Carpediemonas membranifera.* **E :** *Chilimastix cuspidata* (retortamonadine). **F :** *Trepomonas agilis* (diplomonadine). **G :** *Percolomonas descissus* (percolozoaire). **H :** *Monocercomonoides hausmanni* (oxymonadine). **I :** *Dimastigella trypaniformis* (euglénozoaire). **J :** *Pseudotrichomonas keilini* (parabasalien).
D'après Simpson (2003)

2.3.5.2. Évidences morphologiques pour l'hypothèse Excavata

L'organisation du cytosquelette des jakobides, des malawimonadines, de *Trimastix*, de *Carpediemonas* et des retortamonadines est très similaire et suggère une origine commune de ces cinq lignées, appelés « excavés typiques » (Simpson, 2003). Les **jakobides** (p. ex., *Reclinomonas, Histiona, Jakoba, Seculamonas*) et les **malawimonadines** (*Malawimonas*) sont des bactériophages flagellés à vie libre qui possèdent une mitochondrie, tandis que ***Trimastix***, ***Carpediemonas*** et les **retortamonadines** ne possèdent pas de mitochondrie classique et habitent dans des environnements pauvres en oxygène (Simpson, 2003). En plus du sillon nutritif ventral, ces cinq groupes partagent au moins sept autres caractéristiques morphologiques uniques aux excavés (Table V). Ces particularités ultrastructurales sont associées à l'appareil flagellaire et incluent les fibres non microtubulaires I, B, C et composée, la division de la racine microtubulaire en deux portions (extérieure et intérieure), une racine simple extra

et un type particulier de manche flagellaire (O'Kelly et Nerad, 1999; Simpson, 2003; Simpson et Patterson, 1999).

Les **diplomonadines** (p. ex., *Giardia*), un groupe d'amitochondriés habitant les intestins des animaux, et les **percolozoa** ou Heterolobosea (p. ex., *Percolomonas*), un groupe composé en majorité d'organismes amiboïdes, possèdent tous les deux un sillon ventral[9], mais manquent de la majorité des autres caractéristiques propres aux excavés (Table V). Par contre, les **oxymonadines**, endosymbiontes des intestins des animaux, partagent plus de caractères structuraux, mais ne possèdent pas le sillon caractéristique. Finalement, les deux groupes qui ne partagent aucun des caractères morphologiques typiques des excavés sont les **parabasaliens** (p.ex., *Trichomonas*) et les **euglénozoaires**, regroupant des organismes photosynthétiques secondaires verts comme les euglénophytes (p. ex., *Euglena*), ainsi que les parasites trypanosomatides (Simpson, 2003).

Tableau V : Distribution des caractéristiques propres aux excavés

Groupe	Sillon ventral	Fibre I	Fibre B	Fibre C	Racine divisée	Racine simple	Manche flagellaire	Fibre composée
Jakobides	●	●	●	●	●	●	●	●
Malawimonas	●	●	●	●	●	●	●	DN
Trimastix	●	●	●	●	●	●	●	●
Carpediemonas	●	●	●	●	●	●	●	●
Retortamonadines	●	●	●	●	●	●	●	●
Diplomonadines	●	●	–	–	●	?	–	–
Heterolobosea	●	●	–	–	●	–	–	–
Oxymonadines	–	●	●	●	–	●	–	–
Parabasala	–	–	–	?	–	–	–	–
Euglenozoaires	–	–	–	–	–	–	–	–

● : Présence du caractère; – : Absence du caractère; ? : homologie douteuse; DN : Données non disponibles. Les groupes soulignés correspondent aux « excavés typiques ». D'après Simpson (2003).

[9] Il n'a pas été trouvé de sillon ventral chez *Giardia* en particulier, mais celui ci est présent chez les autres diplomonadines étudiés.

2.3.5.3. Les phylogénies moléculaires

Plusieurs analyses phylogénétiques incluant une majorité, voire tous, les groupes d'excavés ont été réalisées (p. ex., Archibald, O'Kelly et Doolittle, 2002; Cavalier-Smith, 2003; Edgcomb *et al.*, 2001; Simpson *et al.*, 2002). Cependant, étant basées sur seulement un ou deux gènes, les phylogénies obtenues sont généralement peu résolues et ne rejettent ni ne supportent la monophylie des excavés. Plusieurs sous-ensembles sont toutefois retrouvés. Par exemple, les analyses de la PSU ARNr supportent les relations diplomonadines + retortamonadies + *Carpediemonads* et *Trimastix* + oxymonadies (Dacks *et al.*, 2001; Silberman *et al.*, 2002; Simpson, 2003; Simpson *et al.*, 2002). Des analyses basées sur plusieurs gènes concaténés ont également été réalisées, mais elles incluent seulement entre deux et cinq groupes d'excavés parmi les dix existants. L'analyse non enracinée de la concaténation de six protéines, EF1α, actine, tubulines α et β et HSP70 et 90 (Harper, Waanders et Keeling, 2005) et celle de la concaténation des quatre premières (Baldauf *et al.*, 2000) supportent les relations diplomonadines + parabasaliens et euglénozoaires + Heterolobosea. Deux analyses enracinées avec les archées et basées respectivement sur la concaténation de 13 et 123 protéines incluaient les diplomonadines et les euglénozoaires et ceux groupes émergeaient paraphylétiquement à la base de l'arbre (Bapteste *et al.*, 2002; Moreira, Le Guyader et Philippe, 2000).

En effet, les parabasaliens, les diplomonadines et les euglénozoaires sont parmi les eucaryotes ayant le plus haut taux d'évolution et leur présence cause souvent des artefacts de reconstruction phylogénétique (Embley et Hirt, 1998; Philippe et Germot, 2000; Philippe, Germot et Moreira, 2000). Pourtant, ces deux groupes sont les seuls excavés pour lesquels des données génomiques étaient disponibles au moment de notre étude. Il est donc nécessaire d'obtenir des données génomiques des excavés qui évoluent lentement pour pouvoir réaliser des analyses phylogénomiques incluant des représentants de tous les groupes.

Jusqu'à présent, l'analyse la plus complète pour tester l'hypothèse Excavata se base sur la concaténation de six protéines nucléaires (tubulines α et β, EF1α et 2, HSP70 et 90) et inclut tous les groupes d'excavés sauf les retortamonadines (Simpson, Inagaki et Roger, 2006). Dans cette étude, les excavés ne forment pas un groupe monophylétique, mais quelques-uns des sous-ensembles suggérés par d'autres études (Dacks *et al.*, 2001;

Nikolaev *et al.*, 2004; Silberman *et al.*, 2002; Simpson, 2003; Simpson et Roger, 2004a; Simpson *et al.*, 2002) sont retrouvés: oxymonadines + *Trimastix* (VB = 88 - 100%), diplomonadines + parabasaliens + *Carpediemonads* (VB = 65 - 100%) et jakobides + euglenozoaires + percolozoaires » (VB = 75 - 85%). Curieusement, tous les ensembles retrouvés impliquent des regroupements entre des excavés typiques et non typiques. Ainsi, malgré leurs similarités structurales, les excavés typiques ne forment pas un groupe monophylétique (Figure 34). Le fait que les malawimonadines n'ont pas d'affinité phylogénétique avec aucun des autres excavés est aussi intrigant. Il est cependant important de mentionner que dans l'analyse de Simpson, Inagaki et Roger (2006), les relations entre les groupes d'eucaryotes ne sont pas résolues et subséquemment l'hypothèse Excavata ne peut pas être rejetée.

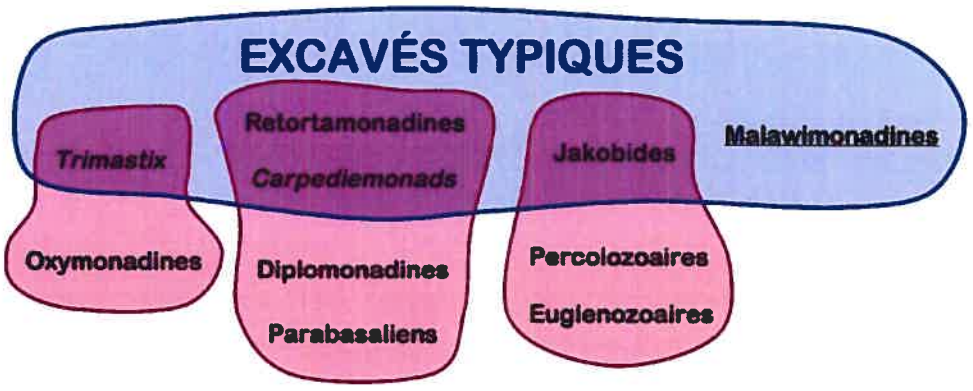

Figure 34 : Relations entre les dix groupes d'excavés

Les groupes encerclés en bleu partagent des caractéristiques ultrastructurales similaires et sont nommés « excavés typiques ». Les groupes encerclés en rose sont ceux trouvés dans les analyses phylogénétiques (voir texte). Il est important de noter que les malawimonadines ne sont apparentés à aucun des autres excavés dans les analyses phylogénétiques.

Adaptation de la Figure 2.6 de Simpson et Roger (2004a)

2.3.5.4. Le génome mitochondrial des jakobides

On a vu plus tôt (section 2.2.2.1) que certaines lignées d'excavés (parabasaliens, diplomonadines) étaient considérées comme les eucaryotes les plus basaux pendant l'essor de l'hypothèse Archezoa, mais plus tard, le séquençage du génome mitochondrial de

Reclinomonas americana (Lang *et al.*, 1997) pointait un autre groupe d'excavés, les jakobides, comme les eucaryotes les plus basaux. En effet, le génome mitochondrial des jakobides est, parmi ceux étudiés, celui qui ressemble le plus au génome des α-proteobactéries : (i) il contient plus de gènes codants qu'aucun autre eucaryote étudié jusqu'à ce jour (Gray, Burger et Lang, 1999; Lang *et al.*, 1997) (Figure 35A), (ii) il présente des motifs Shine-Dalgarno pour l'initiation de la traduction et (iii) il inclut des gènes codant pour des ARNtm, des ARNr et la RNase P, dont la structure secondaire ressemble à celle de leurs homologues bactériens (Gray, Burger et Lang, 1999; Jacob *et al.*, 2004; Lang *et al.*, 1997; Seif, Cadieux et Lang, 2006).

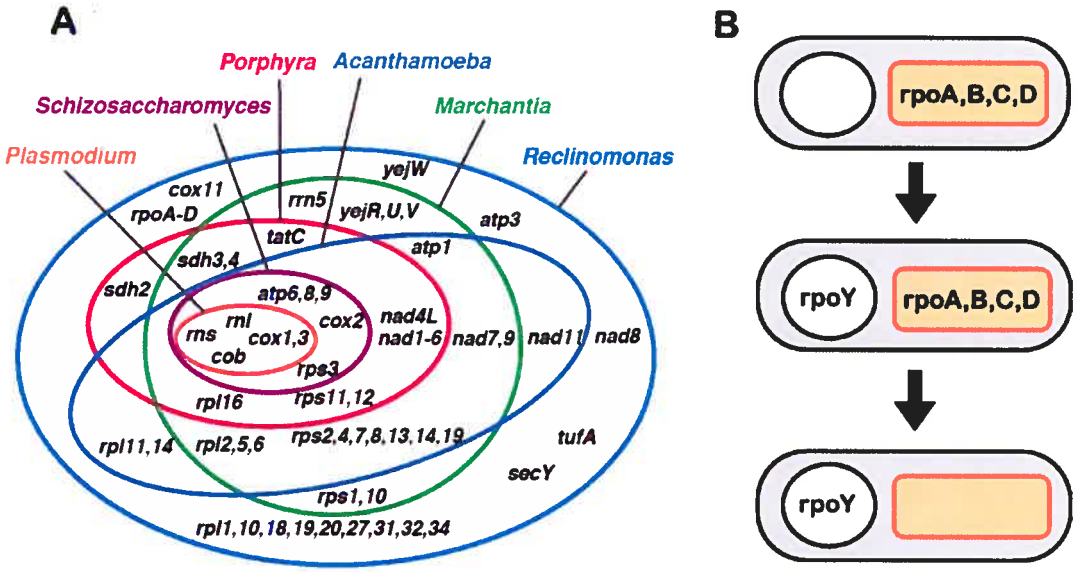

Figure 35: Caractéristiques du génome mitochondrial des jakobides

A : Contenu en gènes de plusieurs génomes mitochondriaux. Les gènes inclus à l'intérieur de chaque ovale correspondent aux gènes présents dans le génome mitochondrial de l'organisme correspondant. Les jakobides contiennent tous les gènes présents dans tous les génomes mitochondriaux connus à ce jour **B :** Modèle illustrant le replacement de l'ARN polymérase mitochondriale d'origine bactérienne (*rpoA, B, C, D*) par une ARN polymérase de type phagique encodée dans le noyau (*rpoY*). Le cercle blanc et le rectangle orange représentent respectivement le noyau et la mitochondrie. Les jakobides semblent être restés dans la première étape du modèle.

D'après Gray, Burger et Lang (1999) et Gray et Lang (1998)

Plus surprenant, le génome mitochondrial des jakobides contient les gènes *rpoA*, *B*, *C* et *D* qui codent pour les quatre sous-unités d'une ARN polymérase de type bactérien (Lang *et al.*, 1997). Chez les autres eucaryotes étudiés, cette enzyme est absente et l'ARN polymérase mitochondriale est une enzyme de type viral codée dans le noyau (Cermakian *et al.*, 1996). Supposant que l'enzyme d'origine virale a remplacé celle d'origine bactérienne très tôt dans l'évolution de la mitochondrie (Figure 35B), l'explication la plus simple pour la présence de l'ARN polymérase bactérienne chez les jakobides est que ceux-ci ont divergé avant les autres eucaryotes et que le remplacement par l'enzyme virale a eu lieu après cette divergence.

2.3.6. L'hypothèse Rhizaria

L'hypothèse Rhizaria est la plus récente des hypothèses concernant la classification des eucaryotes (Cavalier-Smith, 2002) et implique la monophylie d'un groupe très hétérogène de flagellés (les cercomonadines, les foraminifères et des amibes diverses). Il n'existe pas de caractère morphologique unifiant tout le groupe, mais récemment, plusieurs phylogénies moléculaires et des insertions communes ont suggéré la monophylie d'au moins certains de ses membres (Archibald *et al.*, 2003; Burki et Pawlowski, 2006; Keeling, 2001; Longet *et al.*, 2003; Moreira *et al.*, 2006; Nikolaev *et al.*, 2004).

2.4. La racine des eucaryotes

Une des questions les plus controversées de l'évolution des eucaryotes est la position de la racine. Une manière habituelle d'enraciner un arbre phylogénétique est d'utiliser un groupe extérieur, mais, avec les méthodes et les modèles de reconstruction phylogénétique actuellement disponibles, cette méthode n'est pas efficace dans le cas des eucaryotes. En effet, le groupe extérieur (généralement les archées) est tellement éloigné que les inférences sont souvent affectées par des artefacts de reconstruction phylogénétique plaçant les espèces à plus haut taux évolutif à la base de l'arbre (Brinkmann *et al.*, 2005; Philippe et Germot, 2000). Des méthodes alternatives, comme la recherche de

changements rares dans des caractères moléculaires complexes, ont donc été explorées pour définir la position de la racine des eucaryotes.

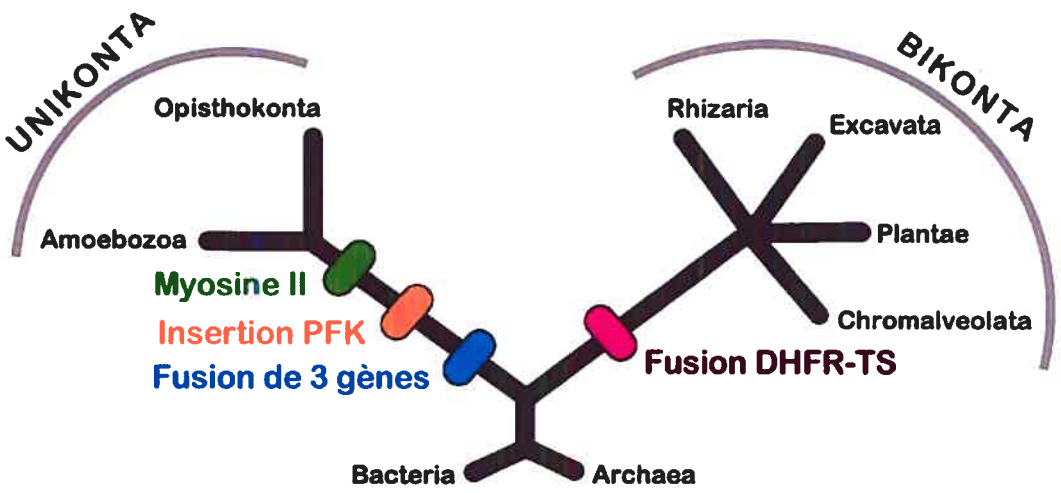

Figure 36: Racine des eucaryotes selon des changements génomiques rares
La fusion de trois enzymes de la voie de synthèse des pyrimidines, la présence d'une myosine de type II et une insertion dans la phosphofructokinase (PFK) chez les unicontes ainsi que la fusion de la dihydrofolate-réductase (DHFR) et de la thymidylate synthase (TS) chez les bicontes supportent tous la position de la racine entre ces deux groupes (voir le texte pour plus de détails).

La combinaison de plusieurs de ces changements rares a été utilisée pour suggérer la racine entre les opisthocontes + Amoebozoa (appelés unicontes) et le reste des eucaryotes (appelés bicontes). Premièrement, chez tous les bicontes les gènes de la thymidylate synthase (TS) et de la dihydrofolate-réductase (DHFR) sont fusionnés, alors qu'ils sont séparés chez les unicontes et chez les procaryotes (Philippe *et al.*, 2000; Stechmann et Cavalier-Smith, 2002). En assumant le scénario le plus parcimonieux, c'est-à-dire, que cette fusion n'est arrivée qu'une seule fois et qu'aucune fission n'a pas eu lieu par la suite, la fusion TS-DHFR serait un caractère dérivé des bicontes supportant la racine à l'extérieur de ce groupe. Deuxièmement, les unicontes partagent (i) la fusion de trois gènes codant pour trois enzymes de la voie de synthèse des pyrimidines (Nara, Hshimoto et Aoki, 2000), (ii) une duplication interne dans le gène de la phosphofructokinase (Stechmann et

Cavalier-Smith, 2003) et (iii) la présence d'une myosine de type II (Richards et Cavalier-Smith, 2005). Puisque ces caractères sont absents chez les bicontes et chez les procaryotes, ils supportent la monophylie des unicontes. Donc, si les bicontes et les unicontes sont respectivement monophylétiques et s'ils incluent tous les eucaryotes, la racine doit forcément être placée entre ces deux ensembles (Figure 36).

La localisation de la racine entre bicontes et unicontes est incompatible avec la position basale des jakobides suggérée par leur ARN polymérase mitochondriale de type bactérien (voir section 2.3.5.3) et implique des scénarios compliqués pour expliquer l'évolution de ce caractère. Indépendamment de la position de la racine (entre les bicontes et les unicontes, entre les jakobides et le reste des eucaryotes, ou ailleurs), des scénarios moins parcimonieux, c'est-à-dire impliquant des gains et des pertes indépendants d'un même caractère chez plusieurs lignées, doivent être considérés pour concilier l'évolution des différents caractères. Ceci suggère que l'homoplasie est aussi envisageable pour des changements génomiques rares et que ce type de caractères doit être interprété avec précaution (Bapteste et Philippe, 2002). La localisation de la racine des eucaryotes reste donc une question ouverte qui nécessite l'étude de la congruence entre les caractères rares et les phylogénies moléculaires.

3. Définition du projet

Le but de ce projet est d'explorer le potentiel de la phylogénomique pour résoudre l'arbre des eucaryotes. D'un côté biologique, deux questions clés dans l'évolution des eucaryotes ont été étudiées : l'origine des plastes et les relations phylogénétiques entre des protistes dits « primitifs ». D'un côté méthodologique, un protocole pour la génération de banques d'ADNc de protistes a été mis en place et des méthodes pour détecter et pour surmonter les erreurs systématiques dans l'inférence phylogénomique ont été explorées.

La première étape du projet a consistée à accumuler des séquences génomiques de trois groupes de protistes peu étudiés, mais qui occupent des positions clés dans la phylogénie des eucaryotes : les glaucophytes, les jakobides et les malawimonadines. Pour faire ceci d'une manière économique et efficace, nous avons opté pour la construction de librairies d'ADNc. Étant donné les particularités de la culture de protistes et de l'extraction de leur ARN, nous avons développé un protocole qui s'adapte spécifiquement à ces organismes. Ce protocole, ainsi que les difficultés associées, ont été présentés dans un manuscrit qui constitue le CHAPITRE I.

Une fois les ESTs du génome nucléaire du glaucophyte *Glaucocystis nostochinearum* séquencés, nous avons initié une collaboration avec un groupe en Autriche qui avait généré des ESTs d'une autre glaucophyte, *Cyanophora paradoxa*. Grâce à ces séquences et à celles déjà disponibles des algues rouges et des plantes vertes, il était possible, pour la première fois, de réaliser des analyses phylogénomiques visant à tester la monophylie des eucaryotes photosynthétiques primaires. Les analyses réalisées ont fait l'objet d'une publication qui est présentée dans le CHAPITRE II.

Le jeux de données qui a été assemblé pour étudier la monophylie des eucaryotes photosynthétiques primaires a été utilisé pour étudier l'impact des erreurs systématiques sur la phylogénomique. Les analyses réalisées ont donné des résultats surprenants qui mettent en évidence de sévères artefacts de reconstruction phylogénétique et qui questionnent la capacité des analyses à échelle génomique pour résoudre la phylogénie des eucaryotes. Nous avons testé diverses méthodes pour surmonter ces artefacts et compilé les résultats obtenus dans un manuscrit qui constitue le CHAPITRE III.

La position de l'algue unicellulaire *Mesostigma viridae* au sein des plantes vertes constitue une question intéressante tant du point de vue biologique que méthodologique. Cette algue partage des caractères morphologiques avec les deux groupes des plantes vertes, soit les streptophytes et les chlorophytes. Cependant, les analyses phylogénétiques donnent des résultats contradictoires quant à sa position. Nous avons réalisé des analyses basées sur des jeux de données nucléaires, mitochondriales et du plaste pour étudier la position de *Mesostigma* ainsi que pour comprendre les artefacts d'inférence phylogénétique responsables des résultats contradictoires obtenus précédemment. Le manuscrit synthétisant cette étude est présenté dans le CHAPITRE IV.

Finalement, nous avons étudié le placement phylogénétique de deux groupes de protistes qui ont occupé des positions clés dans les hypothèses sur l'évolution des eucaryotes : les jakobides et les malawimonadines. Les ESTs séquencés de cinq et deux espèces appartenant respectivement à chacun de ces deux groupes nous ont permis de réaliser les premières analyses phylogénomiques incluant les jakobides et les malawimonadines. Le manuscrit qui décrit cette étude constitue le CHAPITRE V.

CHAPITRE I : LA GÉNÉRATION DE BANQUES D'ADNC

SOUS PRESSE DANS **METHODS IN MOLECULAR BIOLOGY**

CONSTRUCTION OF cDNA LIBRARIES FROM PROTISTS AND FUNGI

NAIARA RODRÍGUEZ-EZPELETA, SHONA TEIJEIRO, LISE FORGET, GERTRAUD BURGER AND B. FRANZ LANG

[1] *Canadian Institute for Advanced Research, Centre Robert Cedergren, Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal, Québec, H3T 1J4, Canada.*

Construction of cDNA libraries from Protists and Fungi

Naiara Rodríguez-Ezpeleta, Shona Teijeiro, Lise Forget, Gertraud Burger and B. Franz Lang

Abstract

Sequencing of cDNA libraries is an efficient and inexpensive approach to analyze the protein coding portion of a genome. It is frequently used for genome survey of poorly studied eukaryotes, and is particularly useful for species that are not easily amenable to genome sequencing, because they are non-axenic and/or difficult to cultivate. In this chapter we describe protocols that have been applied successfully to construct cDNA libraries from numerous protists and fungi, and that require only small quantities of cell material.

Keywords: EST, RNA purification, reverse transcriptase, template-switching, normalisation, DSN

1. Introduction

Sequencing of cDNA libraries has been extensively used to determine the expressed portion of protein coding genes (Expressed Sequence Tags; ESTs) in model eukaryotes. It has also gained importance with the increasing number of eukaryotic genome projects, as the precise inference of exon-intron boundaries relies on substantial training sets of EST data from the respective organisms. In addition, EST sequencing can now also be applied to protists (for most part unicellular eukaryotes that are neither animals, fungi or plants) and fungi (e.g., (1-8)) that are growing poorly, because limitation of cell material can be overcome by PCR-based amplification. There are two major advantages of the EST approach compared to genomics. First, bacterial contamination of eukaryotic cultures can be tolerated, as poly-A tails of bacterial transcripts are too short to be primed by standard oligo-dT primers. In fact, it is even feasible to use total RNA instead of purified mRNA for cDNA library construction, a decisive advantage in the case of non-axenic protists or fungi. The second advantage is that some of these organisms require mechanical methods for cell breakage, because their rigid cell walls resist digestion with commercially available cell wall lysing enzymes. While genomic DNA may become too fragmented through such treatment to be useful for genome sequencing, the significantly smaller mRNAs remain sufficiently intact.

Here, we describe a fast and relatively simple protocol to construct cDNA libraries from protists and fungi. An overview of the procedure is given in Figure 1A. The protocol requires small quantities of cell material, works with both total RNA and purified (poly A^+)

mRNA, and enriches full-length cDNAs. Note, however, that the described protocols involve a PCR amplification step, which is prone to artefacts such as unequal amplification of cDNAs, with a tendency to more efficiently amplify shorter molecules. This potential problem is less relevant in exploratory EST projects. However, in cases where sufficient cell material can be obtained, and where avoiding artefacts is a prime issue, we advise employing cDNA protocols without PCR amplification steps and including prior mRNA purification. In such instances, the readers may follow our procedure from cell culture to mRNA purification, and then continue with one of the protocols described elsewhere in this book.

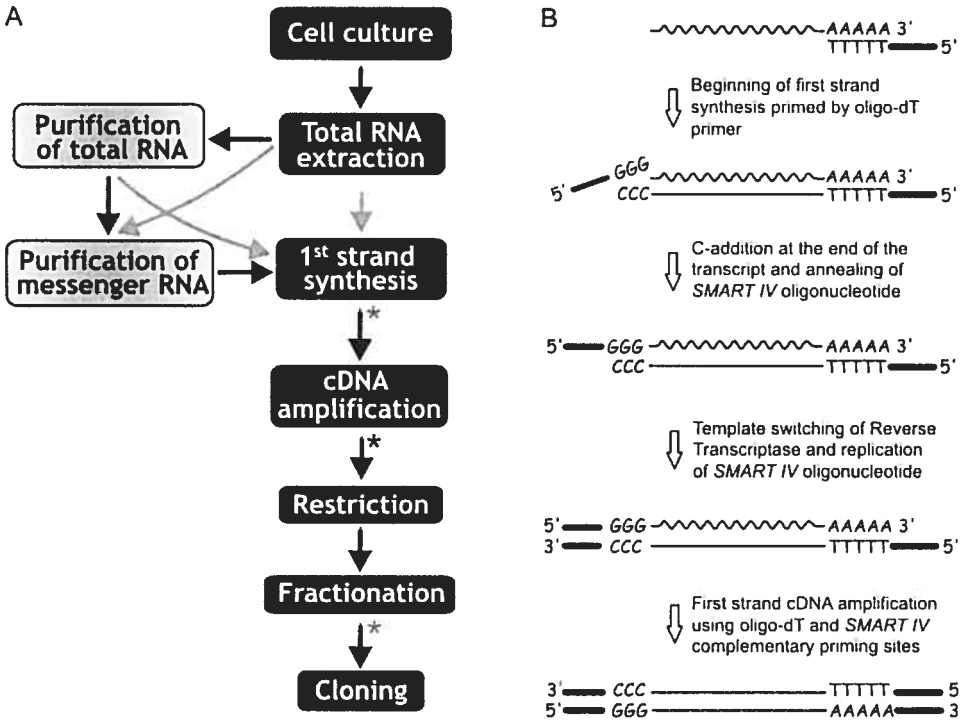

Fig. 1: (A) cDNA library construction. Black boxes, obligatory steps; grey boxes, optional steps. Black arrows point to protocols described in detail. 'Purification of total RNA' stands for elimination of genomic DNA. A black asterisk indicates where a normalisation step is usually introduced; a grey asterisk indicates where alternative normalisation steps can be introduced. (B) First strand cDNA synthesis and amplification. First, an oligo-dT containing oligonucleotide is used to prime first strand cDNA synthesis, catalyzed by a RNase H activity-deficient MMLV reverse transcriptase (RT). When the capped 5' end of the mRNA is reached, the RT adds 2 to 5 C residues to the first-strand cDNA, permitting that the SMART IV primer anneals and that DNA synthesis continues until the end of the oligonucleotide. To our experience, non-capped 5' ends undergo the same reaction, albeit at reduced efficiency. Finally, the oligo-dT and *SMART IV* primers serve for PCR amplification of double-stranded cDNA.

In our standard procedure, first strand cDNA synthesis and amplification are performed using the Creator™ SMART™ cDNA library construction techniques, essentially following the manufacturer's recommendations. The method is based on synthesis of the first cDNA strand with an anchored oligo-dT primer. The terminal C-addition and template switching features of the particular reverse transcriptase allow the second primer to anneal and extend the RNA template and synthesize the first cDNA strand until the end of the primer (Figure 1B) (9-11). Subsequently, the product is amplified by PCR using the same primers as before, cut by *Sfi I* (sites are introduced by the primers), and the asymmetrical restriction sites at both cDNA ends are used for directional cloning. A common problem in first-strand cDNA synthesis is premature termination of reverse transcription at mRNA secondary structures. The SMART™ technique enriches full length cDNAs because template switching to the 5' primer occurs preferentially when the reverse transcriptase has reached the (usually capped) 5' end of the mRNA. Yet, partial cDNAs will also be produced, although less efficiently. In fact, the presence of a certain proportion of partial cDNAs in the cDNA library is desirable, as it reduces the requirement for sequencing long cDNA inserts by primer walking.

Since the abundance of the various transcripts in a cell may vary from thousands of copies to none (12), random EST sequencing becomes usually inefficient after a few thousand readings. The detection of weakly-expressed genes therefore requires normalization of the cDNA libraries, for which we use a simple and efficient procedure involving enzymatic degradation of double stranded DNA or DNA-RNA hybrids (13, 14). We have employed this procedure successfully for several libraries.

2. Materials

Enzymes, buffers, and reagents such as BSA, DTT, dNTP and ATP should be stored frozen at -20°C.

2.1. RNA purification

All solutions must be prepared with RNase-free water, using RNase free chemicals, glass- and plasticware. Gloves should be worn during manipulation of samples (*see* **Note 1**).

1. *Trizol*® reagent (Invitrogen), or preferentially a home-made substitute (*see* **Note 2**): 38% stabilised phenol (to stabilize, add 1 mg of hydroxyl chinoline, 2 µL of mercaptoethanol and 0.5 mL of HPLC water per g of phenol, and incubate at 37°C shaking water), 0.8 M guanidine thiocyanate, 0.4 M ammonium thiocyanate, 0.1 M sodium acetate, 5% glycerol. Note that *Trizol* contains phenol, which causes heavy skin burns and is toxic on contact or by inhalation of vapours. It should therefore be manipulated under a fume hood, using gloves. According to the manufacturer, commercial *Trizol* is stable at 4°C for at least 9 months (but *see* **Note 2**).

2. Mixture (1:1) of 150-212 and 425-600 micron-sized glass beads (Sigma); required for species with a tough cell wall (e.g., most fungi, jakobid flagellates, glaucophytes, red and green algae).
3. Chloroform (add a pinch of bicarbonate for stabilization).
4. Isopropanol.
5. 'Wash ethanol': Ethanol (75%).
6. 'Minikit column', RNeasy® Plus MiniKit (QIAGEN).
7. Oligo-dT cellulose (Amersham Biosciences).
8. '1X Binding Buffer' and '2X Binding Buffer'. Composition of 2X: 20 mM Tris, pH 7.5, 2 mM EDTA, pH 8, 0.1% SDS, 1 M NaCl. This buffer precipitates at room temperature; heat in a water bath at 65°C before use.
9. 'Wash Buffer' (1X): 1 mM Tris, pH 7.5, 1 mM EDTA, pH 8, 0.05% SDS, 0.2 M NaCl. Store at room temperature or heat before use to dissolve precipitated SDS.
10. 'Elution Buffer' (1X): 1 mM Tris, pH 7.5, 1 mM EDTA, pH 8, 0.05% SDS. Store at room temperature or heat before use to dissolve precipitated SDS.
11. NaCl (4 M).
12. 'Ethanol-AmAc': 95% Ethanol; 0.5 M ammonium acetate.

2.2. First strand cDNA synthesis and PCR amplification

1. '*SMART IV* primer', (10 mM) (Clontech):
   5'-AAGCAGTGGTATCAACGCAGACTGGCCATTACGGCCGGG-3'.
2. 'oligo dT-primer': *CDS III/3'* (anchored) PCR primer (10 mM) (Clontech):
   5'-ATTCTAGAGGCCGAGGCGGCCGACATG-d$(T)_{30}N_{-1}$N-3'
   (N_{-1}=A, G or C; N=A, G, C or T).
3. 'RT-buffer': First-Strand Buffer (5X) (Clontech): 250 mM Tris pH 8.3, 30 mM $MgCl_2$, 375 mM KCl.
4. 'Reverse Transcriptase', *PowerScript*™ (Clontech).
5. DTT (20 mM).
6. dNTP mix (10 mM).
7. RNase I (100 µM).
8. '10X PCR Buffer', Advantage®-2 (Clontech): 400 mM Tricine-KOH pH 8.7, 150 mM KOAc, 35 mM $Mg(OAc)_2$, 37 µg/mL BSA, 0.05% Tween-20, 0.05% Nonidet-P40.
9. '5' PCR primer' (10 mM) 5'-AAGCAGTGGTATCAACGCAGAGT-3' (Clontech).
10. 'Polymerase Mix', Advantage®-2 (50X) (Clontech).
11. 'Proteinase K solution', (10 mg/mL).
12. 'Gel Extraction Kit', QIAquick (QIAGEN).

2.3. Normalization

1. DSN enzyme (Evrogen JSC) diluted according to the manufacturer instructions to 1 Kunitz unit/µL.
2. 'DSN Storage Buffer', (Evrogen JSC).

3. '5X Hybridization Buffer': 0.25 M HEPES pH 7.5, 2.5 M NaCl, 1 mM EDTA. This buffer may precipitate; store at room temperature for 20 minutes or incubate at 37°C for about 10 min before use.
4. 'DSN Buffer' for enzyme reaction; 2X composition: 100 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 2 mM DTT.
5. EDTA (5 mM).

2.4. Restriction

1. *Sfi* I restriction endonuclease (20 U/µL) (Clontech).
2. '*Sfi* I Restriction Buffer' (10X) (Clontech).
3. BSA (10 mg/mL).
4. EDTA (0.5 M, pH 8).
5. 'Gel Extraction Kit', QIAquick (QIAGEN).

2.5. DNA Fractionation

1. 'Low melting agarose', SeaPlaque® GTG® ultra-pure (Mandel).
2. Formamide (highest quality, Pharmacia; stored under nitrogen or argon).
3. Electroelution chamber (*see* **Note 3**)
4. 'TAE buffer' (1X): 40 mM Tris-acetate, 20 mM sodium acetate, 1 mM EDTA, pH 8.0.

2.6. Cloning

1. Ligation Buffer (10X): 200 mM Tris pH 7.6, 50 mM $MgCl_2$, 50 mM DTT.
2. pDNRlib vector, cut by *Sfi I* and purified (*see* **Note 4**).
3. ATP (10 mM).
4. T4 DNA ligase (5U/µL).
5. Competent cells (DH5α), and LB agar plates containing 10 mg/mL chloramphenicol and 4 µg/mL tetracycline.

3. Methods

3.1. Cell culture

A large variety of protists require live bacteria as food. To minimize potential RNA degradation by bacterial enzymes, it is important to keep the ratio of eukaryotic *versus* bacterial cells as high as possible. Bacteria can be partially removed from protist cultures through differential centrifugation, but according to our experience, RNA extraction is best performed in the late logarithmic or stationary phase of growth when removal of the

remaining bacteria becomes obsolete. In fungi, cells can be chosen from a wide variety of growth conditions and developmental stages. For initial gene exploration, cells grown under different conditions may be combined.

3.2. RNA purification

3.2.1. Extraction of total RNA

For total RNA extraction, we use a modified *Trizol* protocol.

1. Collect the cells by centrifugation (speed and time varies from one species to another – *see* **Note 5**), or by straining through a fine-mesh nylon coffee filter in case of filamentous fungi.
2. Remove supernatant completely, resuspend cell pellet in *Trizol* and mix. 1 mL of *Trizol* per 5 to 10 x 10^6 cells is recommended. For filamentous fungi or species with rigid cell walls (e.g., algae), add glass beads (*see* **Note 6**) and shake by hand in a glass bottle (15). Once the cells are broken to > 50% (check under light microscope), remove glass beads by repeated rinsing with small volumes of Trizol. The cells in Trizol may be stored at -80°C for at least one month.
3. Leave the cell/Trizol solution for 5 min at room temperature, then add 0.2 mL chloroform per mL of *Trizol*, shake vigorously for 15 seconds and let the mixture settle at room temperature for 2 to 15 min. From here on use 40 mL plastic centrifuge tubes.
4. Centrifuge at 12,000 g at 4°C for 15 min, and collect the colourless aqueous top phase (about 60% of the total volume). Avoid material from the interface (if this occurs, repeat centrifugation).
5. Add 0.3 mL of isopropanol per mL of the collected aqueous phase, mix by inversion and leave 5-10 min at room temperature.
6. Centrifuge at 12,000 g at 4°C for 15 min at and remove the supernatant carefully. The RNA appears as a gel-like or white pellet at the side and bottom of the tube.
7. Wash by adding wash ethanol (1 mL per mL of collected phase) and by inverting the tube a few times. Centrifuge at 12,000 g at 4°C for 10 min and discard supernatant. Repeat the procedure twice. Remove all traces of ethanol by using a Pasteur pipette, and air-dry briefly. The RNA pellet can be stored at -20°C for at least one year.
8. Dissolve the RNA in RNase-free water. The volume of added water will vary with the quantity of recuperated RNA. It is best to start suspending in a small volume and to continue adding until the RNA is perfectly dissolved (final RNA concentration ~ 1 mg/mL). From here on use 1.5 mL Eppendorf tubes.
9. Determine the quality and quantity of the RNA by agarose gel electrophoresis, together with an RNA marker of known size and concentration. Figure 2 (lane A) shows a typical, high quality RNA extracted by the described method.
10. Remove DNA from RNA preparation by purification on a MiniKit column (*see* **Note 7**). Figure 2 (lanes A and B) shows the same material before and after this step.

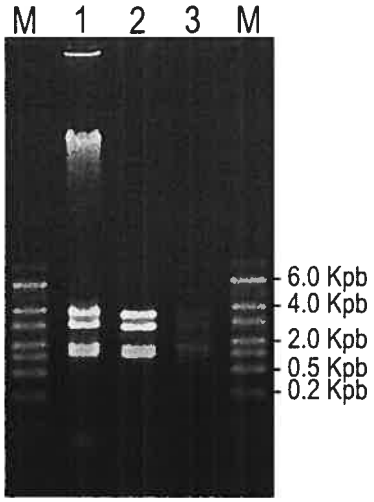

Fig. 2. Agarose gel separation of various stages of a typical RNA purification experiment. 1.5 µL of total RNA were loaded before (lane 1), and after DNA removal (lane 2; step 3.2.1.10 in protocol). Poly A^+ mRNA was then purified from 400 µL of total RNA and recuperated in 15 µL. 3 µL of this mRNA fraction was loaded on the gel (lane 3; step 3.2.2 in protocol). Note a carry-over of rRNAs into the mRNA fraction, which is however neglectable as the amount of loaded mRNA corresponds to ~ 50 times more than that of the total RNA. Lane M; RNA ladder, High Range (Fermentas).

3.2.2. Messenger RNA purification

This protocol applies to ~1 mg of total RNA; the quantity of reagents should be adapted to the actual RNA quantity. Unless otherwise specified, all centrifugations are performed in a microcentrifuge (for Eppendrof tubes) at 14,000 rpm (maximum speed) for 30 sec at room temperature. A correct pH of the solutions (7.5) is critical for high mRNA yield.

1. Fill 40 mg of oligo-dT cellulose into an Eppendorf tube. Wash the cellulose by mixing it with 600 µL Elution Buffer, centrifugation and removal of the supernatant. Repeat this step another two times.
2. Equilibrate the cellulose by mixing it with 600 µL of 1X Binding Buffer, centrifugation and removal of the supernatant. Repeat another two times.
3. Adjust the RNA solution to 600 µL with RNase-free water, and heat it at 65°C for 4 min.
4. Add 600 µL of preheated (65°C) 2X Binding Buffer and incubate at room temperature for 15 min; constantly invert the tube.
5. Centrifuge briefly and discard supernatant.
6. Wash cellulose twice with 1X Binding Buffer and twice with Wash Buffer.
7. Add 250 µL Elution Buffer to cellulose, mix gently and incubate at 37°C for 5 min.
8. Centrifuge for 1 min and transfer the supernatant to a new tube. Add another 250 µL of Elution Buffer to cellulose pellet, mix gently and incubate for 5 min at 37°C.

9. Centrifuge for 1 min and combine the supernatants of step 8 and 9.
10. Using the recovered 500 µl of RNA, repeat the purification cycle at step 3.
11. After two cycles of mRNA purification (steps 3-10), add 40 µl of NaCl (4 M) and 1 mL of Ethanol-AmAc to the RNA and let precipitate at -20°C for 1 h, or over night.
12. Spin down at 14,000 rpm at 4°C for 20 to 30 min, and discard the supernatant.
13. Add 150 µl wash ethanol and centrifuge at 14,000 rpm at 4°C for 10 min.
14. Discard supernatant. Carefully remove ethanol, air dry pellet and resuspend it in RNase-free water. Figure 2 (lane C) shows the result of a typical mRNA purification.

3.3. First strand cDNA synthesis

1. Mix 1-3 µL of RNA solution (~25-500 ng polyA RNA, or 100-1000 ng total RNA), 1 µL *SMART IV* primer, 1 µL oligo dT primer, and adjust volume to 5 µL with HPLC water. Mix well by pipetting up and down.
2. Incubate in a heat block (or PCR machine) at 72°C for 2 min; immediately chill on ice for 2 min; spin down briefly to collect droplets.
3. Add 2 µL 5X RT Buffer and 1 µL Reverse Transcriptase; mix and spin briefly.
4. Incubate in a heat block (or a PCR machine) at 42°C for 1 h.
5. After 1 h, heat again at 72°C for 2 min, chill on ice for 2 min, spin briefly, add 1µL Reverse Transcriptase, and incubate in a heat block at 42°C for 1 h (*see* **Note 8**).
6. Chill on ice. At this point, the first strand cDNA synthesis step is completed.

3.4. cDNA amplification

1. Remove RNA remaining from the first strand cDNA synthesis by adding 0.1 µL of RNase I 100 µM; leave at room temperature for 10 min (*see* **Note 9**).
2. Combine 2 µL first strand cDNA, 80 µL HPLC grade water, 10 µL 10X PCR Buffer, 2 µL 5' PCR primer, 2 µL oligo-dT primer and 2 µL Polymerase Mix. Mix well and PCR-amplify under the following conditions: 20 sec at 95°C followed by ~ 20 cycles: 10 sec at 95°C and 6 min at 68°C. The number of cycles depends on the amount of RNA starting material (*see* **Note 10**).
3. To inactivate the polymerase in the PCR reaction, add 1 µL of Proteinase K solution. Incubate at 45°C for 20 min, followed by 10 min at 65°C.
4. Purify the amplified cDNA with the Gel Extraction kit, and elute the double-stranded cDNA in a final volume of 30 µL (*see* **Note 11** and **Note 12**).

3.5. cDNA normalization

Successful normalization requires an optimized concentration of the DNS enzyme for a given cDNA library. It is recommended to use different DSN concentrations and test the efficiency of normalisation.

1. Prepare three or more dilutions of the original enzyme (1/4, 1/8, 1/16 ...) using DSN Storage Buffer.
2. For each sample to be normalised, mix 4 µL of amplified cDNA (~ 500 ng) with 1 µL of 5X Hybridization Buffer; heat at 98°C for 3 min and at 70°C for 4 h.
3. While keeping the samples at 70°C, add 4 µL of preheated (70°C) 2X DSN Buffer and 1 µL of DSN enzyme (for the dilutions mentioned above, this makes 0.25, 0.125 and 0.0625 Kunitz units); incubate at 70°C for 20 min.
4. Inactivate DSN enzyme by adding 10 µL of 5 mM EDTA.
5. Reamplify the DSN-digested cDNA (as in in section 3.4, points 2 to 4), but elute in 50 µL final volume.
6. Verify the success of the normalisation of each of your samples by gel electrophoresis, and choose that with the desired, even-size distribution for the following step (*see* **Note 13**).

3.6. Restriction

1. Mix 45 µL of amplified and purified cDNA, 40 µL HPLC-grade water, 10 µL *Sfi I* 10 X Restriction Buffer, 1 µL BSA solution, 4 µL (= 80 U) *Sfi I* Restriction enzyme, and incubate for 2 h at 50°C.
2. Stop reaction by addition of 1 µL 0.5 M EDTA.
3. Purify DNA with the Gel Extraction kit by eluting with a final volume of 30 µL (*see* **Note 12**).

3.7. cDNA sizing

1. Depending on the capacity of the wells of the electrophoresis system used, the volume of the sample may have to be reduced from 30 µL to a smaller volume by evaporation in a speed-vac. (see **Note 12**).
2. Add formamide to a final concentration of 10% and loading buffer to the cDNA, incubate for 10 min at 50°C, then chill on ice. This step will reduce aggregation of DNA.
3. Load the sample on a low melting agarose gel (1.2%; TAE buffer) together with a size marker; start by migrating slowly (1.5 V/cm) for a few minutes, then increase to ~3 V/cm. Migrating at higher voltages may overheat and deform the gel matrix. Excise agarose blocks containing DNA fragments of desired size (e.g., 0.5 to 1 kbp; 1 to 5 kbp).
4. Electroelute each cDNA fraction (*see* **Note 14**), and check their yield and size distribution agarose gel electrophoresis loading about 1/10 of the recuperated material (*see* **Note 15**).

3.8. Cloning

Clone each size fraction of cDNA separately.

3.8.1. Ligation

The ration of insert to vector concentrations should be ~ 3:1 (the vector size is 4 kbp). When calculating the size of the inserts, one needs to consider that the fragment sizes are not necessarily uniformly distributed. For example, in the 1 to 5 kbp fraction, there can be overrepresentation of fragments from 1 to 1.5 kbp. The size distribution should be assessed based on agarose gel migration (see step 3.7.4) and the mean size of the fragments estimated accordingly. The final DNA concentration of insert and vector DNA combined should be 5 ng/µL, and the total volume of the reaction should be 3 to 4 µL. See Table 1 for two examples of ligation recipes.

1. For each fraction to be cloned, mix **well** adequate quantities of insert and vector DNA.
2. Heat mix at 50°C for 10 min, place on ice for a few minutes and let stand at room temperature for several minutes.
3. Add Ligation Buffer, ATP (final concentration 0.5 - 1 mM), HPLC grade water and 0.5 U of T4 DNA ligase.
4. Place the ligation mix in a 14°C incubator overnight.

Table 1: Examples for ligation reactions

	0.5 to 1 kbp (mean: 750 bp)	1 to 5 kbp (mean: 1.5 kbp)
Insert[1]	0.5 µL (10 ng/µL)	0.5 µL (20 ng/µL)
Vector (10 ng/µL)	1.00 µL	1.00 µL
Ligation Buffer (10X)	0.30 µL	0.40 µL
ATP (10 mM)	0.27 µL	0.36 µL
Water	0.83 µL	1.64 µL
T4 DNA ligase (5 U/µL)	0.10 µL	0.10 µL
FINAL VOLUME	3.00 µL	4.00 µL

[1]Numbers between parentheses correspond to the initial concentration of insert.

3.8.2. Transformation

The above described ligation mix can be used for 10 transformations (200 µL of competent cells per transformation). Competent cells are prepared and transformation is conducted by standard procedures. Transformed cells should be plated onto chloramphenicol-containing agar plates (about 10 plates per transformation). On average, 1,000 to 2,000 colonies are expected for each transformation with ~20 ng ligation mix (when using high quality competent cells).

4. Notes

1. A major difficulty in handling RNA is the prevention of degradation by contaminant RNases. Autoclaving glassware, tips, tubes and solutions is often insufficient to inactivate RNases. For additional measures, glassware may be baked at 180°C overnight, and plasticware, tubes and solutions be treated with diethylpyrocarbonate (DEPC). DEPC reacts with histidine residues of proteins and thus inactivates RNases. Add DEPC to solutions (water, buffers) at a final concentration of 0.05 - 0.1%, incubate for several hours and autoclave at least 45 min (the characteristic DEPC scent should disappear). Note that DEPC also reacts with RNA; therefore, it has to be completely removed from all materials before use. Moreover, DEPC can react with chemicals containing primary amine groups, such as Tris. Therefore, these chemicals should be added to the solution only once DEPC is removed. DEPC is a suspected carcinogen; take appropriate precautions when handling it (e.g., always wear gloves and handle it under a fume hood). Water purified by a well-maintained MilliQ system is virtually RNase-free, without further treatment. To verify if MilliQ water is indeed RNase-free, dissolve high quality RNA in this water, incubate it at 37°C for several hours, and compare the RNA before and after incubation by gel electrophoresis.
2. Highest quality RNA (and high molecular weight DNA) is regularly obtained with home-made *Trizol*, but not with commercial *Trizol* sources. An apparent reason is that the recipe for home-made *Trizol* calls for phenol of highest purity (supplied in light-protected glass bottles under a protective gas), and that it is protected from oxidation by additives. Advanced phenol oxidation, which is recognizable by a reddish-pink color, causes slight RNA but severe DNA degradation. Stabilized phenol or *Trizol* prepared by our recipe may be stored frozen at -20°C for many years. Once in use, we recommend to keep it for < 1 month at 4°C, in light-protected bottles (brown glass or unstained glass wrapped with aluminium foil).
3. There are numerous techniques for electroeluting DNA from agarose, and diverse devices are commercially available that will not be described here. Relevant information specific to each technique is easily available. The least complicated procedure is electroelution in a closed dialysis tube. Electroelution chambers are available from Schleicher and Schüll (Elutrap), Millipore (Centrilutor), EMD BioSciences (D-tube electroelution) and RPI Research Products (GeneCapsule), to mention only some of the more popular devices.
4. We have noted on several occasions that the 220 bp stuffer fragment is not removed from the commercially distributed, 'ready-to-use' pDNRlib cloning vector. For highest cloning efficiency, the *SfiI*-digested vector should be purified by electrophoresis on low-melting agarose, followed by electroelution of the 4.2 kbp fragment.
5. The centrifugation conditions for pelleting cells depend on multiple factors that are specific to each culture (density of the medium, type of cells, etc). Small cell pellets, in particular those of small and/or flagellated eukaryotes, tend to dissolve quickly and have to be decanted immediately after centrifugation, under close visual control.
6. A number of protists and fungi contain a rigid cell wall that is not (or only for a small fraction of cells) dissolved by *Trizol*. In such cases, cells have to be broken

mechanically. We recommend cell disintegration in the presence of *Trizol*, as RNA will otherwise be degraded by intra-cellular (and if present, bacterial) RNases. Cells of filamentous fungi are broken by grinding together with sand or glass beads in a mortar; cells of unicellular organisms may be disintegrated by manual shaking together with glass beads in a glass bottle (e.g., (15)), or by other suitable disruption methods. Because the volume increases by the addition of glass beads, more *Trizol* has to be used in this case (we use ~ 10 mL of *Trizol* and 10 mL of glass beads for 1 g of cells). Once > 50% of cells are broken (check by microscopy), decant the glass beads and collect the supernatant. Repeatedly (2 to 4 times) rinse the glass beads with small volumes of *Trizol* to collect a maximum of the cell lysate.

7. Total RNA extractions contain variable amounts of genomic DNA (depending on the organism and the extraction conditions), which should be eliminated to avoid undesirable PCR products.
8. The Reverse Transcriptase reaction is repeated once to increase cDNA length. By heating to 72°C after the first reaction cycle, secondary structures in mRNA are destabilized and elongation of the first strand may proceed in the subsequent cycle.
9. RNA should be digested after first strand synthesis to permit optimal synthesis of a second DNA strand, and to avoid interference in the following PCR amplification step.
10. The amount of RNA starting material *versus* the number of PCR cycles recommended by manufacturer is as follows (total RNA/mRNA/number of cycles): 1.0-2.0 µg/0.5-1.0 µg/18-20; 0.5-1.0 µg/0.25-0.5 µg/20-22; 0.25-0.5 µg/0.125-0.25 µg/22-24; 0.05-0.25 µg/0.025-0.125 µg/24-26. We recommend minimizing the number of amplification cycles to avoid PCR artefacts.
11. A Proteinase K digestion prior to the Gel Extraction is recommended. Other PCR purification methods that efficiently remove dNTPs, salts, and long primers may be used as well (note that the longest primer used here is 59 nt long).
12. In order to maximize DNA recuperation, the elution volume may be increased to 100 µL, and subsequently reduced in a speed-vac.
13. The non-normalised and the normalised samples generated with different DSN concentrations should be compared by agarose gel electrophoresis. For best results, the discrete bands of highly expressed mRNAs should have disappeared, and fragment sizes should be evenly distributed and not be smaller than in the non-normalised sample.
14. For the extraction of the sized DNA fragments from the agarose gel we discourage the use of gel extraction kits, because cloning efficiencies may be reduced by 1 to 3 orders of magnitude, compared to electroelution.
15. Size fractionation of cDNA is sometimes difficult because remaining contaminants (polysaccharides?) are carried over despite purification. In such cases, DNA fragments tend to aggregate, causing contamination of the larger-size cDNA with small fragments in electrophoresis. Separation of smaller cDNA quantities (to avoid overloading of the gel) will often help to reduce, although not eliminate, the problem.

Acknowledgements

We wish to thank Jean-François Bouffard, Jung Hwa Seo, Zhang Wang and Yun Zhu for excellent technical assistance. This research has been funded in part by grants from Genome Quebec/Canada and CIHR Canada (BFL). NRE has been supported by the "Programa de Formación de Investigadores del Departamento de Educación, Universidades e Investigación" (Government of Basque Country). Salary and interaction support from the Canadian Institute for Advanced Research, Program in Evolutionary Biology, (BFL and GB) is gratefully acknowledged.

References

1. Crepineau, F., Roscoe, T., Kaas, R., Kloareg, B., and Boyen, C. (2000) Characterisation of complementary DNAs from the expressed sequence tag analysis of life cycle stages of *Laminaria digitata* (Phaeophyceae) *Plant Mol Biol.* **43**, 503-13.
2. Howe, D. K. (2001) Initiation of a *Sarcocystis neurona* expressed sequence tag (EST) sequencing project: a preliminary report *Vet Parasitol.* **95**, 233-9.
3. Nikaido, I., Asamizu, E., Nakajima, M., Nakamura, Y., Saga, N., and Tabata, S. (2000) Generation of 10,154 expressed sequence tags from a leafy gametophyte of a marine red alga, *Porphyra yezoensis DNA Res.* **7**, 223-7.
4. Qutob, D., Hraber, P. T., Sobral, B. W., and Gijzen, M. (2000) Comparative analysis of expressed sequences in *Phytophthora sojae Plant Physiol.* **123**, 243-54.
5. Broeker, K., Bernard, F., Moerschbacher, B. M., Brown, D. W., Cheung, F., Proctor, R. H., et al. (2006) An EST library from *Puccinia graminis* f. sp. tritici reveals genes potentially involved in fungal differentiation *FEMS Microbiol Lett.* **256**, 273-81.
6. Brown, D. W., Cheung, F., Proctor, R. H., Butchko, R. A., Zheng, L., Lee, Y., et al. (2005) Comparative analysis of 87,000 expressed sequence tags from the fumonisin-producing fungus *Fusarium verticillioides Fungal Genet Biol.* **42**, 848-61.
7. Felipe, M. S., Andrade, R. V., Petrofeza, S. S., Maranhao, A. Q., Torres, F. A., Albuquerque, P., et al. (2003) Transcriptome characterization of the dimorphic and pathogenic fungus *Paracoccidioides brasiliensis* by EST analysis *Yeast* **20**, 263-71.
8. Keon, J., Antoniw, J., Rudd, J., Skinner, W., Hargreaves, J., Hammond-Kosack, K., et al. (2005) Analysis of expressed sequence tags from the wheat leaf blotch pathogen *Mycosphaerella graminicola* (anamorph *Septoria tritici*) *Fungal Genet Biol.* **42**, 376-89.
9. Schramm, G., Bruchhaus, I., and Roeder, T. (2000) A simple and reliable 5'-RACE approach *Nucleic Acids Res.* **28**, E96.
10. Schmidt, W. M., and Mueller, M. W. (1999) CapSelect: a highly sensitive method for 5' CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs *Nucleic Acids Res.* **27**, e31.

11. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., and Siebert, P. D. (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction *Biotechniques* **30**, 892-7.
12. Galau, G. A., Klein, W. H., Britten, R. J., and Davidson, E. H. (1977) Significance of rare m RNA sequences in liver *Arch Biochem Biophys.* **179**, 584-99.
13. Shagin, D. A., Rebrikov, D. V., Kozhemyako, V. B., Altshuler, I. M., Shcheglov, A. S., Zhulidov, P. A., et al. (2002) A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas *Genome Res.* **12**, 1935-42.
14. Zhulidov, P. A., Bogdanova, E. A., Shcheglov, A. S., Vagner, L. L., Khaspekov, G. L., Kozhemyako, V. B., et al. (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease *Nucleic Acids Res.* **32**, e37.
15. Lang, B., Burger, G., Doxiadis, I., Thomas, D. Y., Bandlow, W., and Kaudewitz, F. (1977) A simple method for the large-scale preparation of mitochondria from microorganisms *Anal Biochem.* **77**, 110-21.

CHAPITRE II : UNE ORIGINE UNIQUE DES PLASTES

PUBLIÉ DANS **CURRENT BIOLOGY** 2005; 15(14):1325-30

MONOPHYLY OF PRIMARY PHOTOSYNTHETIC EUKARYOTES: GREEN PLANTS, RED ALGAE AND GLAUCOPHYTES

NAIARA RODRÍGUEZ-EZPELETA[1], HENNER BRINKMANN[1], SUZANNE C. BUREY[2], BÉATRICE ROURE[1], GERTRAUD BURGER[1], WOLFGANG LÖFFELHARDT[2], HANS J. BOHNERT[3], HERVÉ PHILIPPE[1] AND B. FRANZ LANG[1]

[1] *Canadian Institute for Advanced Research, Centre Robert Cedergren, Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal, Québec, H3T 1J4, Canada.*

[2] *Max L. Perutz Laboratories, University Departments at the Vienna Biocenter, Department of Biochemistry and Ludwig Boltzmann Research Unit for Biochemistry, 9 Dr. Bohr-Gasse, Vienna, A-1030, Austria*

[3] *Department of Plant Biology and Crop Sciences, University of Illinois, Urbana/Champaign, 1201 W. Gregory Drive, Urbana, IL 61801, USA*

Current Biology, Vol. 15, 1325–1330, July 26, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.cub.2005.06.040

Monophyly of Primary Photosynthetic Eukaryotes: Green Plants, Red Algae, and Glaucophytes

Naiara Rodríguez-Ezpeleta,[1] Henner Brinkmann,[1] Suzanne C. Burey,[2] Béatrice Roure,[1] Gertraud Burger,[1] Wolfgang Löffelhardt,[2] Hans J. Bohnert,[3] Hervé Philippe,[1,*] and B. Franz Lang[1,*]

[1]Canadian Institute for Advanced Research
Centre Robert Cedergren
Département de Biochimie
Université de Montréal
2900 Boulevard Edouard-Montpetit
Montréal, Québec, H3T 1J4
Canada

[2]Max L. Perutz Laboratories
University Departments at the Vienna Biocenter
Department of Biochemistry and Ludwig Boltzmann Research Unit for Biochemistry
9 Dr. Bohr-Gasse
Vienna, A-1030
Austria

[3]Department of Plant Biology and Crop Sciences
University of Illinois Urbana/Champaign
1201 W. Gregory Drive
Urbana, Illinois 61801

Summary

Between 1 and 1.5 billion years ago [1, 2], eukaryotic organisms acquired the ability to convert light into chemical energy through endosymbiosis with a Cyanobacterium (e.g., [3–5]). This event gave rise to "primary" plastids, which are present in green plants, red algae, and glaucophytes ("Plantae" sensu Cavalier-Smith [6]). The widely accepted view that primary plastids arose only once [5] implies two predictions: (1) all plastids form a monophyletic group, as do (2) primary photosynthetic eukaryotes. Nonetheless, unequivocal support for both predictions is lacking (e.g., [7–12]). In this report, we present two phylogenomic analyses, with 50 genes from 16 plastid and 15 cyanobacterial genomes and with 143 nuclear genes from 34 eukaryotic species, respectively. The nuclear dataset includes new sequences from glaucophytes, the less-studied group of primary photosynthetic eukaryotes. We find significant support for both predictions. Taken together, our analyses provide the first strong support for a single endosymbiotic event that gave rise to primary photosynthetic eukaryotes, the Plantae. Because our dataset does not cover the entire eukaryotic diversity (but only four of six major groups in [13]), further testing of the monophyly of Plantae should include representatives from eukaryotic lineages for which currently insufficient sequence information is available.

Results and Discussion

Plastid Genes Significantly Support Plastid Monophyly

The monophyly of plastids is supported by several common features, such as a similar gene content of plastid genomes, the presence of plastid-specific gene clusters that are distinct from those in Cyanobacteria, the conservation of the plastid-protein import machinery and protein-targeting signals, and phylogenies based on plastid and cyanobacterial gene sequences (see [5] and references therein). Yet, some authors have challenged each of these evidences as either weak or inconclusive [14]. In particular, the molecular phylogenies are often based on single or a few genes and are not robust (e.g., [9, 10]). One published multigene phylogeny recovers plastid monophyly, but it includes only a few cyanobacterial taxa [15].

Our analyses are the first to include in a phylogenomic framework data from a broad diversity of Cyanobacteria and other related Bacteria for testing plastid monophyly. Our dataset contains 50 proteins (10,334 amino acid positions) from 16 plastids and 15 Cyanobacteria; 13 additional bacteria (10 Gram-positive bacteria, *Deinococcus*, *Thermus*, and *Chloroflexus*) were added to this dataset for a reduced number of proteins (26 proteins totaling 4,998 amino acid positions) in order to determine the root. Four different phylogenetic inference methods were employed: maximum likelihood (with a concatenate [cML] and a separate [sML] model), Bayesian inference (BI), maximum-likelihood-based distance (Dml), and maximum parsimony (MP). As shown in Figure 1, plastids form a strongly supported monophyletic group (100% bootstrap value [BV]). Within plastids, the relationships among green plants, red algae, and glaucophytes remain unresolved at standard confidence levels. It has been proposed that systematic errors such as long-branch attraction (LBA), compositional bias, and covarion structures are responsible for the recovery of plastid monophyly [16, 17]. We therefore performed analyses with LogDet distances [18], a covarion model [19, 20], and by including only the slowest evolving plastids—*Porphyra* (red alga), *Mesostigma* (green plant), and *Cyanophora* (glaucophyte). All tests of possible artifacts as suggested by Lockhart and coworkers did not affect the strong support for plastid monophyly. Horizontal gene transfer (HGT) is obviously another major concern in cyanobacterial phylogeny [21], but does not seem to affect our results (see the Supplemental Experimental Procedures in the Supplemental Data available with this article online).

Interestingly, with a dataset including 13 additional bacteria as an outgroup (Figure S1), *Gloeobacter* is the deepest branch within Cyanobacteria, consistent with seemingly "primitive" features of the photosynthetic apparatus (see [9] and references therein). However, it cannot be excluded that this is due to LBA and that *Gloeobacter* is highly derived and not early diverging. Similarly, because plastids are fast evolving relative to

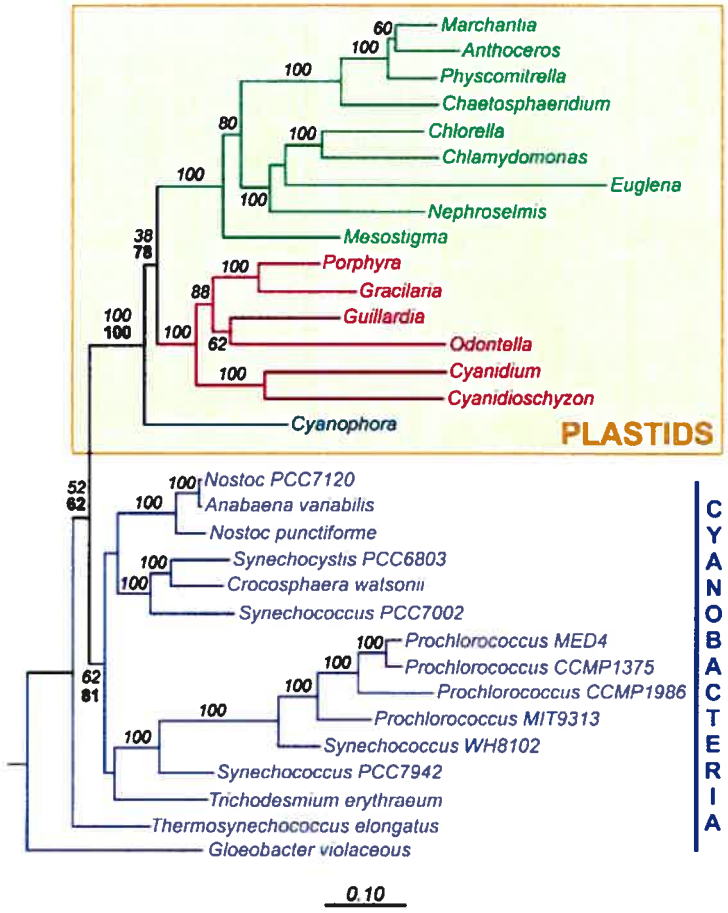

Figure 1. Phylogeny based on plastid and bacterial protein sequences

The analysis is based on the concatenated dataset of plastid-encoded proteins (50 proteins; 10,334 amino acid positions). The tree has been inferred with BI with the WAG+F+Γ model. Numbers in italics represent support values obtained with 100 bootstrap replicates on the concatenated dataset with PhyML (WAG+F+Γ model), and numbers below (in bold) represent bootstrap values based on 10,000 RELL replicates of the sML analysis (see Experimental Procedures for details). The presence of a single value indicates that this branch was constrained in the separate analysis (except for the position of *Euglena*, which was also constrained). The scale bar denotes the estimated number of amino acid substitutions per site. Bootstrap values lower than 50% obtained in both approaches are not shown. The dotted line indicates the position of the root, which was inferred with a dataset of 26 plastid proteins including 13 additional bacteria (see Figure S1). Species names in certain colors denote the following: in green, green plants plus the secondary-plastid-containing *Euglena*; in red, red algae and secondary-plastid-containing *Odontella* and *Guillardia*; and in blue, glaucophytes. Taxon designations are as follows: *Anthoceros formosae*, *Marchantia polymorpha*, *Physcomitrella patens*, *Chaetosphaeridium globosum*, *Euglena gracillis*, *Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Nephroselmis olivacea*, *Mesostigma viride*, *Cyanidioschyzon merolae*, *Cyanidium caldarium*, *Odontella sinensis*, *Guillardia theta*, *Porphyra purpurea*, *Gracilaria tenuistipitata*, and *Cyanophora paradoxa*. *Prochlorococcus* stands for *Prochlorococcus marinus*. Note that we have not included *Synechococcus* PCC6301 because it is closely related if not identical with *Synechococcus PCC7942*.

most Cyanobacteria, they might be attracted toward the root of the tree by LBA. Genome projects on potentially basally diverging Cyanobacteria (e.g., *Pseudanabaena*; [9]) and improved tree-inference methods are required to resolve these questions with confidence.

Nuclear Genes Significantly Support the Monophyly of Plantae

The monophyly of Plantae has been tested with phylogenies that use nuclear and mitochondrial sequences, but support is weak (e.g., [7, 8, 11, 12]). Strong support for the sister-group relationship of green plants and red algae has been obtained in multiprotein phylogenies, one with 13 nuclear proteins [22] and the other with four mitochondrial proteins [23]. However, the nuclear tree has nonsignificant support for the monophyly of Plantae when the then-available six glaucophyte-protein sequences are included, whereas the mitochondrial phylogeny does not include glaucophytes. In addition, the exclusion of only one protein (elongation factor 2) from the nuclear dataset or the use of alternative mitochondrial datasets (*nad* versus *cob/cox* genes) drastically reduces support for the sister-group relationship of green plants and red algae [7, 24]. The use of a limited number of genes is a possible explanation for the lack of significant support for or against the monophyly of Plantae. Indeed, it is well documented that single gene sequences often do not contain sufficient phylogenetic signal to resolve short internal branches, even at moderately deep divergence.

We have therefore performed phylogenetic analyses based on a dataset of 143 orthologous nuclear proteins (30,113 amino acid positions) from 39 species. To overcome the lack of data from glaucophytes, the less-studied group of primary photosynthetic eukaryotes, we have sequenced 4,628 and 8,696 expressed sequence tags (ESTs) from *Cyanophora paradoxa* and *Glaucocystis nostochinearum*, respectively. Our dataset represents all major eukaryotic groups for which sufficient sequence information is available, i.e., not including members of two major, potentially polyphyletic groups Rhizaria (Cercozoa, Radiolaria, Foraminifera, etc.) and excavates (jakobids, malawimonads, Heterolobosea, etc). The monophyly of Plantae remains to be tested with respect to these missing groups. Analyses including diplomonads, parabasalids, and kinetoplastids, the only excavates for which enough data are available, demonstrate that these excavates are fast

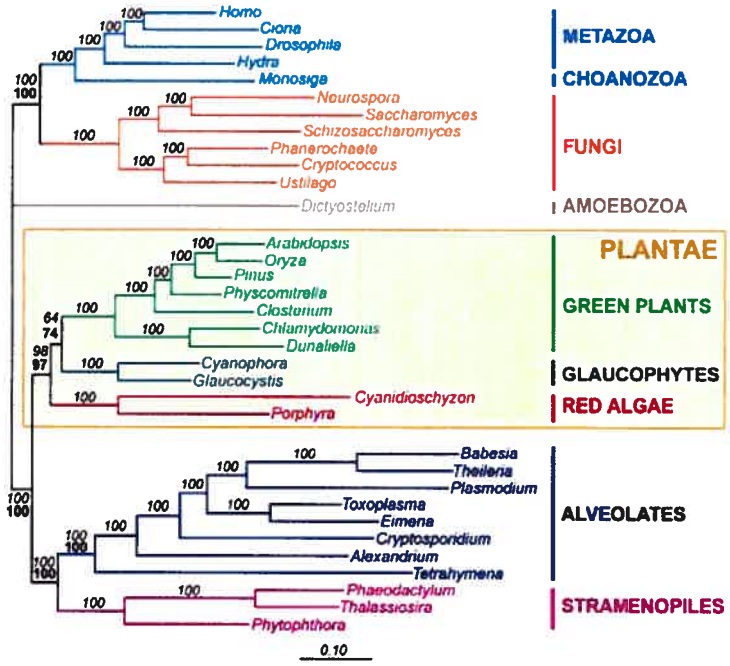

Figure 2. Phylogeny Based on Nuclear-Encoded Protein Sequences

The analysis is based on the concatenated dataset of nuclear encoded proteins (143 proteins; 30,113 amino acid positions). The posterior probabilities for all branches are 1.0. For further details, see Figure 1. The tree is rooted between opisthokonts and other eukaryotes (excluding *Dictyostelium*), a proposal based on the presence/absence of a gene fusion [42, 43]. The position of *Dictyostelium* cannot be deduced with this gene-fusion event because it has lost the corresponding homologous genes. Therefore, a basal trifurcation with *Dictyostelium*, opisthokonts and other eukaryotes is shown. Taxon designations are as follows: *Homo sapiens*, *Ciona intestinalis*, *Drosophila melanogaster*, *Hydra magnipapillata*, *Monosiga brevicollis*, *Candida albicans*, *Saccharomyces cerevisiae*, *Neurospora crassa*, *Phanerochaete chrysosporium*, *Cryptococcus neoformans*, *Ustilago maydis*, *Arabidopsis thaliana*, *Oryza sativa*, *Pinus taeda*, *Physcomitrella patens*, *Closterium peracerosum-strigosum-littorale complex*, *Chlamydomonas reinhardtii*, *Dunaliella salina*, *Cyanidioschyzon merolae*, *Porphyra yezoensis*, *Cyanophora paradoxa*, *Glaucocystis nostochinearum*, *Babesia bovis*, *Theileria annulata*, *Plasmodium falciparum*, *Eimeria tenella*, *Toxoplasma gondii*, *Cryptosporidium parvum*, *Alexandrium tamarense*, *Phaeodactylum tricornutum*, *Thalassiosira pseudonana*, and *Phytophthora sojae*.

evolving (Figure S4). Because the inclusion of fast-evolving taxa causes phylogenetic artifacts [25, 26], only analyses without these lineages are shown in the following (note, however, that there is no difference in tree topology; Figure 2, Figure S4, and Table S1).

Analyses with the remaining 34 species (cML, sML, and BI, Figure 2; and MP and DmI, not shown) significantly support the monophyly of Plantae and of all other relationships except one (see below), indicating the presence of a strong signal in our dataset. In the cML analysis, the monophyly of all major lineages (e.g., animals, fungi, green plants, alveolates, and stramenopiles [13]) are confirmed with bootstrap values of 100%, which corroborates numerous previous analyses. In addition, the superkingdom Opisthokonta, including Fungi and Holozoa (Metazoa and the choanoflagellate *Monosiga brevicollis*), is recovered at 100%, as are the superensemble Alveolata uniting Apicomplexa, Ciliophora, and Dinoflagellata (100%) and a clade uniting stramenopiles and alveolates (100%). Finally, in the cML analysis, the support value for the monophyly of Plantae is significant (98%). Our inferences with sML, which fits the data best (Table S2) also recover the monophyly of Plantae with high confidence (97% BV; Figure 2). However, the relationships among the three groups of Plantae remain unsupported (64% and 74% BV for the sister-group of green plants and glaucophytes, with the cML and sML approaches, respectively).

To assess the confidence level for the monophyly of Plantae, we retained the best 25 topologies from the exhaustive sML analysis (see Experimental Procedures) and performed several statistical tests. All tests gave essentially the same results; the least-biased and most-rigorous test available to date, the "approximately unbiased" (AU) test, is shown [27, 28] for seven of the most relevant topologies tested (Table 1). The AU test rejects all scenarios in which Plantae are not monophyletic (significance level = 0.05). The relationships within the Plantae remain unresolved, although the sisterhood of glaucophytes and green plants has the highest probability, in agreement with the results of the bootstrap analyses shown in Figure 2. Interestingly, removal of the fast-evolving *Cyanidioschyzon* renders the three alternative arrangements among the lineages of Plantae almost identical (see column AU-33 in Table 1). A detailed discussion of the impact of taxon sampling in phylogenomics will be presented in a separate study (N.R.-E. et al., unpublished data).

How Many Genes Does It Take to Resolve the Monophyly of Plantae?

The above-presented analyses are the first that strongly support the monophyly of Plantae. To verify whether a large number of genes is indeed required to obtain this result, we calculated for each internal branch in Figure 2 the bootstrap values as a function of the number of amino acid positions used (Figure 3). With ~8000 amino acid positions, all internal branches but two are recovered with a BV > 90%. The monophyly of Plantae is supported with only 70% BV, with the same number of amino acid positions (Figure 3, thick

Table 1. Likelihood Tests of Alternative Tree Topologies

Rank	Tree topology	Δ ln L[a]	AU[b]	AU-33[c]
1	Best tree; glaucos with greens	−27.5	0.892	0.575
2	Glaucos with reds	27.5	0.297	0.567
3	Glaucos basal to (reds + greens)	42.8	0.147	0.412
4	Reds basal to (alveos + strams)	84.5	0.044	0.018
5	Glaucos basal to (dicts + opis)	137.6	0.006	0.007
11	Greens basal to (alveos + strams)	235.2	3e-07	2e-04
12	Three Plantae lineages unrelated	238.6	3e-61	8e-30

Comparison of alternative trees with CONSEL [40], inferred from separate maximum-likelihood analyses of 143 proteins and 34 species, with the same model as the analysis in Figure 2. The 25 best topologies from the sML analysis were retained. In the three best topologies, Plantae are monophyletic. All other 22 topologies are rejected at a significance level of 0.05. Topologies 4, 5, and 11 are the best in which only two Plantae lineages are sister groups, and topology 12 is the best in which the three Plantae lineages are unrelated. The following abbreviations are used: glaucos, glaucophytes; reds, red algae; greens, green plants; alveos, alveolates; strams, stramenopiles; dicts, *Dictyostelium*; and opis, opisthokonts.

[a] Log likelihood difference.

[b] Approximate Unbiased test.

[c] When removing *Cyanidioschyzon* from the dataset (AU-33 column), AU values for the best and the second-best tree become almost identical (0.575 and 0.567), eliminating the marginal support for the sister-group relationship of glaucophytes and green plants.

black line). In fact, the support value of this branch increases slowly but regularly with the addition of positions, finally reaching 90% BV at >20,000 amino acid positions. The sisterhood of green plants and glaucophytes (Figure 3; thick dotted line) also increases slowly with the addition of more data, but it reaches only 74% BV with the complete dataset (Figure 2).

In summary, our results show that 30,000 amino acid positions are necessary to recover the monophyly of Plantae with significant support. This explains why other studies, which all used much fewer sequence positions, did not obtain statistically significant support for this clade (e.g., [8, 22, 24]).

Conclusions

Our phylogenomic analyses support the idea that Plantae are monophyletic and that plastids form a monophyletic group to the exclusion of Cyanobacteria, providing compelling evidence for a single origin of primary photosynthesis in eukaryotes. Still, our large datasets are insufficient to resolve the branching order within Plantae and are thus unable to support or reject the common assumption that glaucophytes emerged prior to the divergence of green plants and red algae. Addressing this issue requires analyses that include more taxa and/or more genes of the three Plantae lineages, in particular red algae and glaucophytes. Furthermore, the monophyly of Plantae remains to be tested after addition of several major eukaryotic groups not included here because of the lack of gene sequences or of their fast rate of evolution. The eukaryotic groups from which data are most urgently required, preferentially from slow-evolving species, are the Rhizaria (including Cercozoa, Radiolaria, Foraminifera, etc.), Amoebozoa (Lobosa and Conosa), and the potentially nonmonophyletic Excavata (Euglenozoa, Heterolobosea, jakobids, malawimonads, diplomonads, parabasalids, retortamonads, etc.).

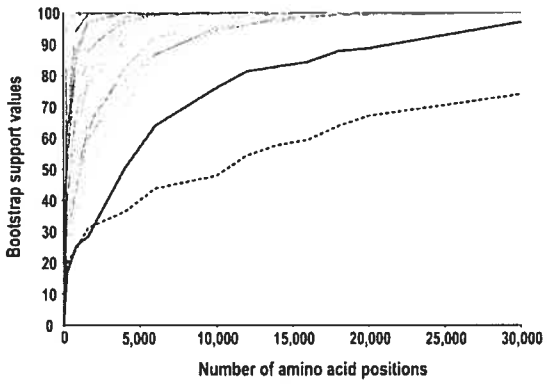

Figure 3. More than 100 Genes Are Required to Recover the Monophyly of Plantae

Evolution of the bootstrap support values (BV) for each internal branch, as a function of the number of amino acid positions. Y and X axes refer to bootstrap values (in %) and number of amino acid positions, respectively. Thick line represents monophyly of Plantae; the dotted line represents green plants + glaucophytes; and the thin lines represent other internal branches in Figure 2.

As we show here, the high support for the monophyly of Plantae critically relies on the use of a large collection of protein sequences. Whereas few relationships (e.g., the sisterhood of animals and fungi) can be convincingly demonstrated already with a small number of sequences (e.g., 13 mitochondrial proteins), we posit that the resolution of other ancient events in the history of eukaryotes will require massive datasets in the order of 100 or more genes [29].

An efficient way to obtain data from many organisms is EST sequencing, which requires limited amounts of cell material and is therefore useful for the exploration of underrepresented eukaryotes, many of which are difficult to grow and unavailable in axenic culture. On the basis of phylogenetic analyses with these data, key species can then be selected for genome projects. Glaucophytes clearly belong to the taxa of prime interest because they contain minimally derived plastids and appear as the slowest-evolving eukaryotes in our phylogenomic analyses (Figure 2). Complete glaucophyte genome sequences would allow for a better understanding of the origin of eukaryotic photosynthesis

while providing deeper insight into the biogenesis of plastids.

Experimental Procedures

Construction of cDNA Libraries, Sequencing, Selection of Orthologous Proteins, and Data Extraction from Multiple Alignments

A detailed description of cDNA-library constructions and sequencing is available with the Supplemental Data. Sequences are available at http://amoebidia.bcm.umontreal.ca/public/pepdb/welcome.php. The nuclear dataset is based on an available alignment [12]. Data from *Cyanophora*, *Glaucocystis* (this study), and additional sequences retrieved from GenBank (http://www.ncbi.nlm.nih.gov) and other sources (see Supplemental Data) were added to the alignment as described [12]. Species evolving at highly accelerated rates (Microsporidia, Euglenozoa, Parabasalida, and Diplomonadida) were not included (but see Figure S4), and only representative (preferentially slowly evolving) members of fungi, animals, and embryophyte plants were used. Unambiguously aligned sequence blocks were extracted with Gblocks [30]; after manual verification, potential paralogs were identified and removed as described [31]. When all orthologous proteins that are available from at least 23 out of the 34 used species are included, the dataset contains 143 proteins (see Supplemental Data for a detailed list), totaling 30,113 amino acid positions. On average, 19% of the amino acids are missing.

The plastid dataset consists of 50 proteins (a total of 10,334 amino acid positions; see Supplemental Data for a detailed list) from 16 plastids and 15 cyanobacteria that were publicly available. The number of land-plant plastids in this data collection was restricted to three slowly evolving species. An alternative plastid dataset including 26 proteins (a total of 4,998 amino acid positions) from 13 additional bacteria was used to root the tree. Sequences were aligned with CLUSTALW [32] and refined manually with MUST [33], and ambiguously aligned positions were removed with Gblocks [30]. The two resulting datasets are available upon request.

Phylogenetic Analyses

The concatenated datasets of nuclear and plastid/cyanobacterial sequences were analyzed by maximum likelihood (ML) with PhyML 2.4 [34], maximum parsimony (MP) with PAUP* 4.0 b10 [35], bayesian inference (130,000 and 120,000 generations for nuclear and plastid dataset respectively, repeated three times with identical results) with MrBayes 3.0 b4 [19], and distance methods with TREE-PUZZLE 5.2 [36] and BIONJ [37]. The reliability of each internal branch was evaluated on the basis of 100 (ML) or 1000 (MP and distance approach) bootstrap replicates. Subsequently, separate ML analyses (sML) were conducted as described [31]. In brief, relationships that are undisputed and supported by 100% bootstrap values (e.g., the monophyly of animals, fungi, and green plants) were constrained, and all resulting tree topologies were exhaustively analyzed independently for each protein to identify the tree topology with the best overall likelihood value (for more details, see Supplemental Data). Site-wise likelihood values were calculated by PAML [38]. The support for each internal branch was evaluated by the RELL method [39], with 10,000 replicates. For likelihood tests of competing tree topologies, p values were calculated with CONSEL [40].

Number of Amino Acid Positions and Bootstrap Support

For sML analysis, the relationship between the number of sequence positions and the bootstrap value was calculated for various internal branches as described [41]. In order to do so, the constraints were adapted to permit testing of groups within Apicomplexa, Fungi, Holozoa, green plants, and stramenopiles. In brief, variable fractions of amino acid positions of the complete dataset (e.g., 1,000; 2,000; 3,000; ...; 30,000) were randomly drawn from the dataset, each 100 times. RELL bootstrap analysis was then performed on each of the 100 samples for each size fraction. The average of the bootstrap values for each size fraction was plotted against its size.

Supplemental Data

Supplemental Data include Supplemental Experimental Procedures, four figures, and two tables and are available with this article online at http://www.current-biology.com/cgi/content/full/15/14/1325/DC1/.

Acknowledgments

We wish to thank Frédéric Delsuc, Nicolas Rodrigue, and three anonymous reviewers for helpful comments on the manuscript; David To for his assistance in configuring the parallel version of MrBayes; and the PEP sequencing team at the Université de Montréal for the supply of *Glaucocystis* data. This work has been supported by operating and equipment funds from Genome Quebec/Canada. B.F.L., G.B., and H.P. are members of the Program in Evolutionary Biology of the Canadian Institute for Advanced Research (CIAR), whom we thank for salary and interaction support. The authors H.P. and B.F.L. are grateful to the Canada Research Chairs Program and the Canadian Foundation for Innovation (CFI) for salary and equipment support. H.J.B. acknowledges support from the National Science Foundation, USA, DBI-9813360 and DBI-0223905, and thanks the W.M. Keck Center for Comparative and Functional Genomics at the University of Illinois, Urbana-Champaign, for DNA sequencing services (ESTs of *Cyanophora*). We are also grateful to the Austrian Research Fund for financing grant P15438 (to W.L.). N.R.E. has been supported by "Programa de Formación de Investigadores del Departamento de Educación, Universidades e Investigación" (Government of Basque Country), and B.R. is a CIHR strategic training fellow in Bioinformatics.

Received: February 13, 2005
Revised: June 7, 2005
Accepted: June 9, 2005
Published: July 26, 2005

References

1. Douzery, E.J., Snell, E.A., Bapteste, E., Delsuc, F., and Philippe, H. (2004). The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? Proc. Natl. Acad. Sci. USA *101*, 15386–15391.
2. Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G., and Bhattacharya, D. (2004). A molecular timeline for the origin of photosynthetic eukaryotes. Mol. Biol. Evol. *21*, 809–818.
3. Mereschkowsky, C. (1905). Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biologisches Centralblatt *25*, 593–604.
4. McFadden, G.I. (2001). Primary and secondary endosymbiosis and the origin of plastids. J. Phycol. *37*, 951–959.
5. Palmer, J.D. (2003). The symbiotic birth and spread of plastids: how many times and whodunit? J. Phycol. *39*, 4–11.
6. Cavalier-Smith, T. (1981). Eukaryote kingdoms: seven or nine? Biosystems *14*, 461–481.
7. Stiller, J.W., Riley, J., and Hall, B.D. (2001). Are red algae plants? A critical evaluation of three key molecular data sets. J. Mol. Evol. *52*, 527–539.
8. Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. Science *290*, 972–977.
9. Turner, S., Pryer, K.M., Miao, V.P., and Palmer, J.D. (1999). Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J. Eukaryot. Microbiol. *46*, 327–338.
10. Douglas, S.E., and Turner, S. (1991). Molecular evidence for the origin of plastids from a cyanobacterium-like ancestor. J. Mol. Evol. *33*, 267–273.
11. Bhattacharya, D., Helmchen, T., Bibeau, C., and Melkonian, M. (1995). Comparisons of nuclear-encoded small-subunit ribosomal RNAs reveal the evolutionary position of the Glaucocystophyta. Mol. Biol. Evol. *12*, 415–420.
12. Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W., and Casane, D. (2004). Phylogenomics of eukaryotes: impact

of missing data on large alignments. Mol. Biol. Evol. *21*, 1740–1752.
13. Simpson, A.G., and Roger, A.J. (2004). The real 'kingdoms' of eukaryotes. Curr. Biol. *14*, R693–R696.
14. Stiller, J.W., Reel, D.C., and Johnson, J.C. (2003). A single origin of plastids revisited: convergent evolution in organellar genome content. J. Phycol. *39*, 95–105.
15. Cai, X., Fuller, A.L., McDougald, L.R., and Zhu, G. (2003). Apicoplast genome of the coccidian *Eimeria tenella*. Gene *321*, 39–46.
16. Lockhart, P.J., Howe, C.J., Bryant, D.A., Beanland, T.J., and Larkum, A.W. (1992). Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol. *34*, 153–162.
17. Lockhart, P.J., Steel, M.A., Barbrook, A.C., Huson, D.H., Charleston, M.A., and Howe, C.J. (1998). A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. *15*, 1183–1188.
18. Lockhart, P.J., Steel, M.A., Hendy, M.D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. *11*, 605–612.
19. Huelsenbeck, J.P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *17*, 754–755.
20. Fitch, W.M. (1976). The molecular evolution of cytochrome c in eukaryotes. J. Mol. Evol. *8*, 13–40.
21. Zhaxybayeva, O., Lapierre, P., and Gogarten, J.P. (2004). Genome mosaicism and organismal lineages. Trends Genet. *20*, 254–260.
22. Moreira, D., Le Guyader, H., and Philippe, H. (2000). The origin of red algae and the evolution of chloroplasts. Nature *405*, 69–72.
23. Burger, G., Saint-Louis, D., Gray, M.W., and Lang, B.F. (1999). Complete sequence of the mitochondrial DNA of the red alga *Porphyra purpurea*. Cyanobacterial introns and shared ancestry of red and green algae. Plant Cell *11*, 1675–1694.
24. Nozaki, H., Matsuzaki, M., Takahara, M., Misumi, O., Kuroiwa, H., Hasegawa, M., Shin-i, T., Kohara, Y., Ogasawara, N., and Kuroiwa, T. (2003). The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids. J. Mol. Evol. *56*, 485–497.
25. Sanderson, J.S., and Shaffer, H.B. (2002). Troubleshooting molecular phylogenetic analyses. Annu. Rev. Ecol. Syst. *33*, 49–72.
26. Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. *6*, 361–375.
27. Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. Syst. Biol. *51*, 492–508.
28. Whelan, S., Lio, P., and Goldman, N. (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. *17*, 262–272.
29. Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M., et al. (2002). The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Natl. Acad. Sci. USA *99*, 1414–1419.
30. Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540–552.
31. Philippe, H., Lartillot, N., and Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. Mol. Biol. Evol. *22*, 1246–1253.
32. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). 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*, 4673–4680.
33. Philippe, H. (1993). MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res. *21*, 5264–5272.
34. Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*, 696–704.
35. Swofford, D.L. (2002). PAUP: Phylogenetic Analysis using Parsimony (and Other Methods). Version 4 (Sunderland, Massachusetts: Sinauer Associates).
36. Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*, 502–504.
37. Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. *14*, 685–695.
38. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556.
39. Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. J. Mol. Evol. *31*, 151–160.
40. Shimodaira, H., and Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics *17*, 1246–1247.
41. Lecointre, G., Philippe, H., Van Le, H.L., and Le Guyader, H. (1994). How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. Mol. Phylogenet. Evol. *3*, 292–309.
42. Philippe, H., Lopez, P., Brinkmann, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Muller, M., and Le Guyader, H. (2000). Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc. R. Soc. Lond. B. Biol. Sci. *267*, 1213–1221.
43. Stechmann, A., and Cavalier-Smith, T. (2002). Rooting the eukaryote tree by using a derived gene fusion. Science *297*, 89–91.

Supplemental Data

S1

Monophyly of Primary Photosynthetic Eukaryotes: Green Plants, Red Algae, and Glaucophytes

Naiara Rodríguez-Ezpeleta, Henner Brinkmann, Suzanne C. Burey, Béatrice Roure, Gertraud Burger, Wolfgang Löffelhardt, Hans J. Bohnert, Hervé Philippe, and B. Franz Lang

Supplemental Experimental Procedures

Construction of cDNA Libraries

Total RNA was purified from exponentially growing *Cyanophora paradoxa* (LB1555UTEX) cells via the Qiagen-tips 500 (Qiagen, Chatsworth, CA; protocol for plant tissues). Three micrograms of PolyA$^+$ RNA (isolated with the Oligotex-dT kit; Qiagen) was used for cDNA synthesis (cDNA Synthesis Kit; Stratagene, La Jolla, CA), and fragments larger than 500 bp were selected by DNA size-fractionation columns (Invitrogen, Carlsbad, CA) and ligated into the plasmid pBluescript SKII (+) (Stratagene) for directional cloning. Transformed ElectroMAX™ DH10B™ cells (Invitrogen) were plated onto Sgal/IPTG/Amp plates for black/white selection.

Glaucocystis nostochinearum total RNA was extracted after breaking of cells with glass beads [S1], with TRIZOL (Invitrogen) instead of the extraction buffer and following the manufacturer's instructions for the RNA purification steps. A cDNA library was constructed with 100 ng polyA$^+$ RNA (isolated with Oligotex-dT Cellulose Type 7; Amersham Biosciences; Piscataway, NJ) and primers, plasmid pDNRlib, and reverse transcriptase from the Creator™ Smart™ cDNA Library Construction kit (BD Biosciences Clontech; Palo Alto, CA). Plasmids were isolated from bacterial cultures with the QIAprep 96 Turbo Miniprep Kit (Qiagen), and sequencing reactions were performed with the ABI Prism BigDyeTM Terminators version 3.0/3.1 (Perkin-Elmer, Wellesley, MA). The purified sequencing reactions were separated and analyzed either on an ABI 3700 (*Cyanophora* ESTs) or an MJ BaseStation automatic sequencer (*Glaucocystis* ESTs). Normalization of the *Glaucocystis nostochinearum* cDNA library was performed with size-selected (>500 bp) cDNA, following a published procedure [S2]. Trace files were interpreted with PHRED, and sequences were quality- and vector-trimmed before clustering with PHRAP [S3, S4], followed by automatic annotation [S5] and incorporation into the PEPdb database.

Data Sources for Some Organisms

Candida albicans: Stanford Genome Technology Center (http://www-sequence.stanford.edu/group/candida).

Cryptococcus neoformans: *C. neoformans* cDNA Sequencing Project (http://www.genome.ou.edu/cneo.html) and *C. neoformans* Genome Project, Stanford Genome Technology Center, and The

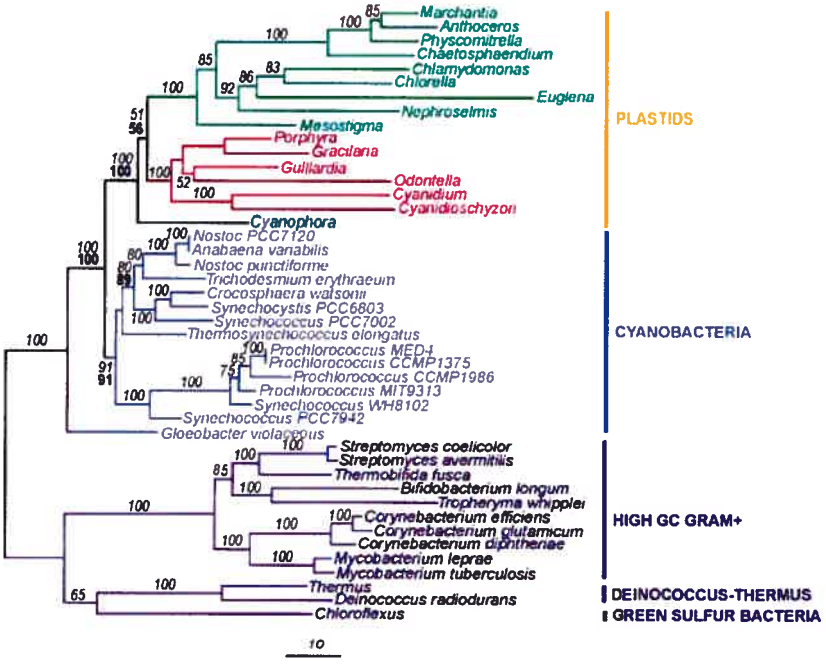

Figure S1. Phylogeny Based on Plastid and Bacterial Protein Sequences

The analysis is based on the concatenated dataset of plastid-encoded proteins (26 proteins; 4,998 amino acid positions). The tree has been inferred with BI with the WAG+F+Γ model. Numbers in italics represent support values obtained with 100 bootstrap replicates on the concatenated dataset with PhyML (WAG+F+Γ model), and numbers below (in bold) represent bootstrap values based on 10,000 RELL replicates of the sML analysis (see Experimental Procedures for details). The presence of a single value indicates that this branch was constrained in the separate analysis [except the sister group of (*Porphyra* + *Gracilaria*) and (*Guillardia* + *Odontella*), which is also constrained]. The scale bar denotes the estimated number of amino acid substitutions per site. Bootstrap values lower than 50% are not shown.

Table S1. Likelihood Tests of Alternative Topologies Including the Excavates

Excavates Sister Group of:	Δ ln *L*a	AU
bMetazoa/Choanozoa	188.1	0.006
Fungi	608.2	4e-70
Metazoa/Choanozoa/Fungi	113.9	1e-47
Dictyostelium	−0.1	0.532
Metazoa/Choanozoa/Fungi/*Dictyostelium*	10.4	0.345
Alveolates/Stramenopiles	0.1	0.606
Plantae	79.1	3e-05
Glaucophytes/Green plants	246.9	4e-74
Green plants	356.3	4e-38
Glaucophytes	406.4	4e-51
Red algae	349.6	3e-76
Alveolates	86.0	0.015
Stramenopiles	658.3	0.001

Starting from the tree topology of Figure 2 in the main text, the three excavate lineages (*Trichomonas*, *Giardia*, and kinetoplastids) were placed as sister group to all well-established monophyletic groups, and the likelihood values of the resulting 13 topologies were calculated with TREE-PUZZLE [S8] with a WAG+F+Γ model. The rows in which Plantae are polyphyletic are marked in grey. Rejections at significance level of 0.05 are in bold and underlined. All tests were performed with CONSEL.

[a]Log likelihood difference

[b]AU test.

Institute for Genomic Research (http://www-sequence.stanford.edu/group/c.neoformans).

Thalassiosira pseudonana: The DOE Joint Genome Institute (http://genome.jgi-psf.org/thaps1/thaps1.download.ftp.html).

Phytophthora sojae: The DOE Joint Genome Institute (http://genome.jgi-psf.org/sojae1/sojae1.download.ftp.html).

Monosiga brevicollis: [S6].

Phylogenetic Analyses Based on Plastid Sequences

Proteins Used

The following proteins were used: Acetyl-CoA carboxylase β subunit; ATP synthetase α chain; ATP synthetase β chain; ATP synthetase C chain; ATP synthetase epsilon chain; ATP synthetase subunit 6; Photosystem II 44Da reaction center protein; Photosystem II apoprotein; Cytochrome b6; Apocytochrome f; cytochrome b6/f complex subunit IV; cytochrome b6/f complex subunit V; Photosystem I P700 apoprotein A1; photosystem I subunit IX; Photosystem I assembly protein ycf3; Photosystem I assembly protein ycf4; Photosystem II protein D1; Photosystem II protein D2; photosystem II protein I; photosystem I P700 apoprotein A2; Photosystem I iron-sulfur center; Cytochrome b-559 α subunit; Cytochrome b-559 β subunit; photosystem II phosphoprotein; photosystem II complex subunit J; photosystem II K-protein; photosystem II L-protein; 50S ribosomal proteins 14, 16, 20, 22, 2, 32 and 37; 30S ribosomal proteins 11, 12, 14, 16, 18, 19, 2, 3, 4, 7 and 8; RNA polymerase α subunit; RNA polymerase β subunit; DNA-directed RNA polymerase β′ chain; and cytochrome c biogenesis protein.

Gblocks Parameters

A minimum of 50% of the sequences for conserved positions, a minimum of 75% for flanking positions, a maximum of five contiguous nonconserved positions, and a minimum of five positions per block were used.

Constraints Applied in a Separate ML Analysis

- Green plastids (((((*Anthoceros formosae*, *Marchantia polymorpha*), *Physcomitrella patens*), *Chaetosphaeridium globosum*), ((*Euglena gracillis*, (*Chlamydomonas reinhardtii*, *Chlorella vulgaris*)), *Nephroselmis olivacea*)), *Mesostigma viride*)
- Red plastids ((*Cyanidioschyzon merolae*, *Cyanidium caldarium*), ((*Odontella sinensis*, *Guillardia theta*), (*Gracilaria Tenuistipitata Porphyra purpurea*)))
- Cyanobacterial group 1 (((((*Prochlorococcus marinus CCMP1375*, *Prochlorococcus marinus MED*), *Prochlorococcus marinus CCMP1986*), *Prochlorococcus marinus MIT9313*), *Synechococcus sp WH*), *Synechococcus elongatus PCC7942*)
- Cyanobacterial group 2 ((*Anabaena variabilis ATCC29413*, *Nostoc sp*), *Nostoc punctiformis*)
- Cyanobacterial group 3 (*Synechococcus PCC7002*, (*Synechocystis PCC*, *Crocosphaera watsonii*))
- Noncyanobacterial outgroup (((((*Corynebacterium glutamicum*, *Corynebacterium efficiens*), *Corynebacterium diphtheriae*), (*Mycobacterium tuberculosis*, *Mycobacterium leprae*)), ((*Thermobifida fusca*, (*Streptomyces avermitilis*, *Streptomyces coelicor*)), (*Tropheryma whipplei*, *Bifidobacterium longum*))), ((*Thermus*, *Deinococcus radiodurans*), *Chloroflexus*))

In the dataset including the noncyanobacterial outgroup, one additional constraint was introduced: a sister-group relationship of *Trichodesmium* with the group (*Anabaena variabilis ATCC29413*, *Nostoc sp*), *Nostoc punctiformis*). The above-described constraints reduce the number of operational taxonomic units (OTUs) to nine (135,135 possible topologies). Likelihood values were calculated with PROTML [S7] and a JTT+F model, for each protein and for each topology. Subsequently, the sum of log likelihood values for all proteins was computed. The 4000 best topologies were selected and analyzed with TREE-PUZZLE version 5.2 [S8] and a WAG+F+Γ (eight categories) model.

Phylogenetic Analyses Based on Nuclear Sequences

Proteins Used

The following proteins were used: Actin; Cytosolic chaperonin complex subunits α, β, γ, delta, epsilon, zeta, eta and theta; Heat shock protein HSP 60 kDa mitochondrial; Heat shock protein HSP 70 kDa cytosol; Heat shock protein 5 HSP 70 kDa glucose regulated; Heat

Table S2. Akaike's Information Criterion

	AIC	Concatenated Model	Separate Model
Nuclear Dataset			
	No gamma	1,559,930 (−779,882; 84)	1,567,790 (−772,963; 10,930)
	Gamma	1,476,540 (−738,186; 85)	1,470,950 (−724,401; 11,073)
Plastid Dataset			
	No gamma	444,772 (−222,308;78)	437,436 (−214,818; 3,900)
	Gamma	419,794 (−209,794;79)	412,954 (−202,527; 3,950)

The log likelihood value for the best topology (see Figures 1 and 2 in the main text) was calculated according to four different models (concatenated and separate—both with and without rate variation among sites—gamma, and no gamma). All models use the WAG amino acid substitution matrix. The Akaike's Information Criterion (AIC) values for each model were calculated as *AIC* = $-2 \times$ *log likelihood* + $2 \times$ *number of free parameters*. The number of free parameters to be estimated for each model was calculated as: *number of branches* + *number of amino acid frequencies (19)* + *one alpha parameter (when a gamma distribution was considered)*. The AIC values for each model are shown in bold and underlined, and numbers in parenthesis correspond to the log likelihood values and the number of free parameters.

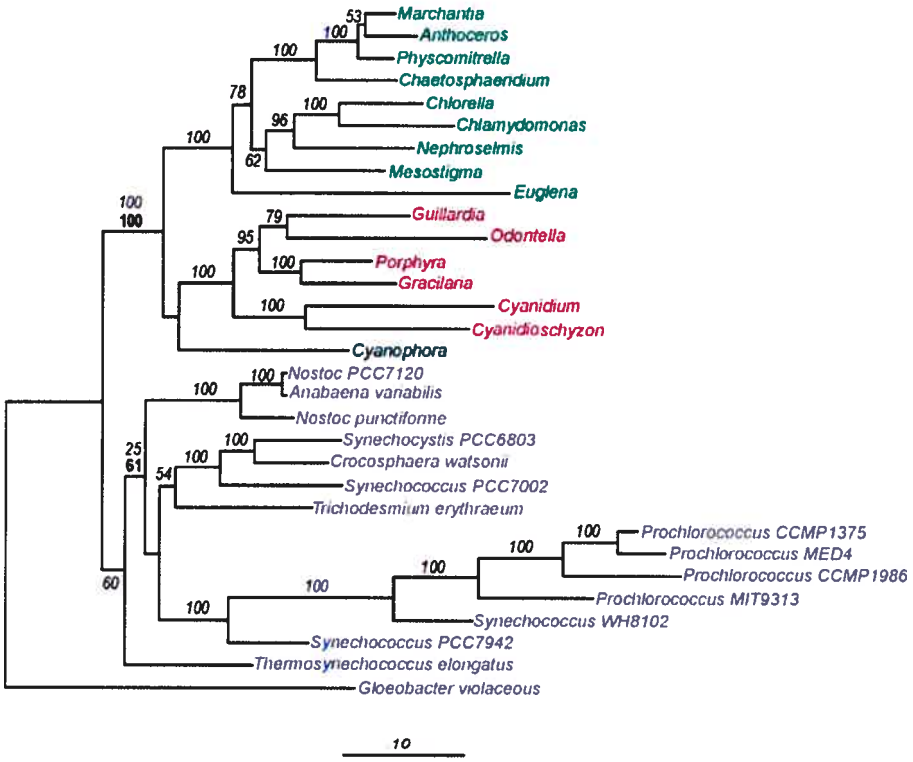

Figure S2. Phylogeny Based on Plastid and Bacterial Operational Genes

The analysis is based on the concatenated dataset of plastid-encoded proteins (22 proteins; 6380 amino acid positions). See Figure 1 in the main text for details.

shock protein HSP 70 kDa mitochondrial; 26S proteasome AAA-ATPase regulatory subunits 4, 6 6A, 6A′, 7, and 8; Transitional endoplasmic reticulum ATPase TER ATPase; Vesicular fusion protein nsf2; Elongation factor EF1 α; EF1 α related GTP binding protein polypeptide chain release factor RF3; Elongation factor EF2; Elongation factor Tu family U5 snRNP specific protein; Eukaryotic initiation factor 5a; Eukaryotic peptide chain release factor subunit 1; 40S ribosomal protein SA 40 kDa laminin receptor 1; Signal recognition particle receptor α subunit SR α; Signal recognition particle 54 kDa protein; Seryl tRNA synthetase; TATA box binding protein related factor 2; Nucleolar GTP binding protein 1; Fibrillarin; Farnesyl pyrophosphate synthase; 20S proteasome β subunit macropain zeta chain; 20S proteasome α 1 chain; 20S proteasome α 2 chain; 20S proteasome α 3 chain; 20S proteasome α 6 chain; 20S proteasome α v chain; 20S proteasome α w chain; 20S proteasome α x chain; 20S proteasome α y chain; 20S proteasome α z chain (because no standard nomenclature does exist for some 20S proteasome chains, names v, w, x, y, and z were arbitrarily chosen); proteasome β 4; proteasome β 5; proteasome β 6; proteasome β 7; 60 ribosomal protein L10 QM protein; Histone H4; Heat shock 70 kDa protein C, E, SSE and mitochondrial forms; initiation factors 1a, 2b, 2g, 2p, and 6; Inositol-3-phosphate synthatse isozyme 1; minichromosome family maintenance protein 5; S-adenosyl-methionine synthetase; Ribosome biogenesis protein NEP1 C2F protein; ATP binding protein subunit B and C; Protein Chromosome 2 orf4 CGI-27 kDa; Shwachman-Bodian-Dyamond syndrome protein; UV excision repair protein RAD23; DNA repair protein RAD51; 60S acidic ribosomal protein P0 L10E; Succinyl-CoA ligase α chain mitochondrial precursor; DNA topoisomerase I, mitochondrial precursor; Vacuolar ATP synthase catalytic subunit A; Vacuolar ATP synthase catalytic subunit B; Vacuolar ATP synthase catalytic subunit C; Vacuolar ATP synthase catalytic subunit E; TGF β inducible nuclear protein; High mobility group like nuclear protein 2 NHP2; High mobility group like nuclear protein 2 NHP2 protein 1; 60S acidic ribosomal protein P2; 60S ribosomal proteins L7a, 11b, 12b, 13, 14a, 15a, 16b, 17, 18, 19a, 1, 20, 21, 22, 23a, 24a, 24b, 25, 26, 27, 2, 30, 31, 32, 33a, 34, 35, 371, 38, 39, 3, 42, 43, 4b, 5, 6, 7a, and 9; 40S Ribosomal proteins 10, 11, 12, 13a, 14, 15, 16, 19, 1, 20, 22a, 23, 25, 26, 27, 28a, 29, 2, 3, 4, 5, 6, and 8; Tubulins α and β; and threonyl-tRNA synthetase.

Gblocks Parameters

A minimum of 50% of the sequences for conserved positions, a minimum of 75% for flanking positions, a maximum of five contiguous nonconserved positions, and a minimum of five positions per block were used.

Constraints Applied in a Separate ML Analysis:

- Holozoa ((((*Homo sapiens*, *Ciona intestinalis*), *Drosophila melanogaster*), *Hydra magnipapillata*), *Monosiga brevicollis*)
- Fungi (((*Candida albicans*, *Saccharomyces cerevisiae*), *Neurospora crassa*), ((*Phanerochaete chrysosporium*, *Cryptococcus neoformans*), *Ustilago maydis*))
- Green plants (((((*Arabidopsis thaliana*, *Oryza sativa*), *Pinus taeda*), *Physcomitrella patens*), *Closterium peracerosum-strigosum-littorale complex*), (*Chlamydomonas reinhardtii*, *Dunaliella salina*))
- Red algae (*Cyanidioschyzon merolae*, *Porphyra yezoensis*)
- Glaucophytes (*Cyanophora paradoxa*, *Glaucocystis nostochinearum*)
- Apicomplexa + Dinoflagellata (((((*Babesia bovis*, *Theileria annulata*), *Plasmodium falciparum*), (*Eimeria tenella*, *Toxoplasma gondii*)), *Cryptosporidium parvum*), *Alexandrium tamarense*)
- Stramenopiles ((*Phaeodactylum tricornutum*, *Thalassiosira pseudonana*), *Phytophthora sojae*)

The above-described constraints reduce the number of operational

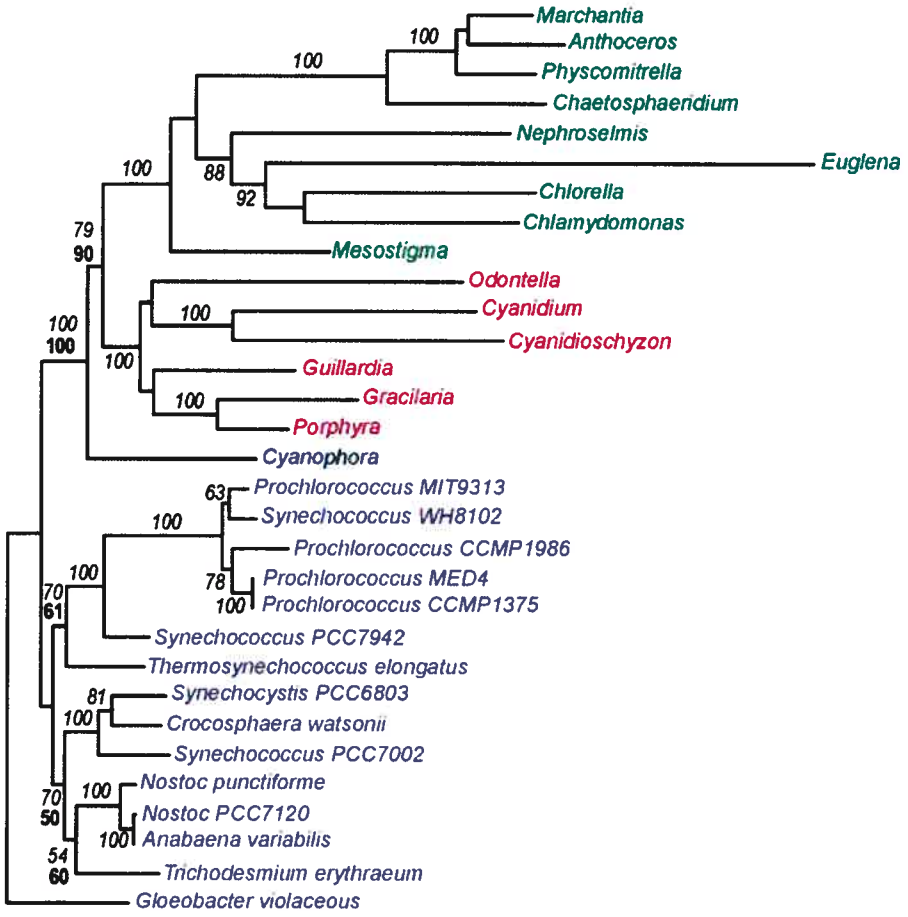

Figure S3. Phylogeny Based on Plastid and Bacterial Informational Genes

The analysis is based on the concatenated dataset of plastid-encoded proteins (28 proteins; 3954 amino acid positions). See Figure 1 in the main text for details.

taxonomic units (OTUs) to nine (135,135 possible topologies). Likelihood values were calculated with PROTML [S7] and a JTT+F model, for each protein and for each topology. Subsequently, the sum of log likelihood values for all proteins was computed. The 1500 best topologies were selected, as well as 500 topologies covering the spectrum of likelihood values of the remaining topologies. The resulting 2000 topologies were analyzed with TREE-PUZZLE version 5.2 [S8] and WAG+F+Γ (eight 8 categories) model.

Horizontal Gene Transfer In Cyanobacterial Phylogenies

Horizontal gene transfer (HGT) is a major concern in cyanobacterial phylogeny [S9], although a core of "informational" genes (coding for proteins involved in transcription, translation, and replication) are less affected by transfers than operational genes (all other proteins) [S10]. HGT is expected to decrease tree resolution in multi-gene analyses, and might in part explain the weak resolution of the cyanobacterial tree (Figure 1 in the main text). To test this hypothesis, we have performed phylogenetic analyses based on two subsamples of the plastid dataset, one including only informational genes, and the other one operational genes (Figures S2 and S3). Trees inferred from both datasets are similar to the one in Figure 1 in the main text, recovering the monophyly of plastids with strong support, but failing to resolve the relationships within Cyanobacteria. Accordingly, HGT does not explain the difficulty to place plastids within the cyanobacterial tree.

Concatenated Versus Separate Model

The sML model allows branch length and the α parameter of the γ distribution (modeling among-site rate variation) to be independently estimated for each gene [S11]. According to the Akaike information criterion [S12] (defined as AIC = $-2 \times$ log likelihood $+ 2 \times$ number of free parameters), the sML model, although requiring many more free parameters, provides significantly higher likelihood values than the concatenated model (Table S2). In fact, for all the models tested (including different amino acid replacement matrices; not shown), sML with a WAG substitution matrix combined with a γ distribution fits the data best.

Supplemental References

S1. Lang, B., Burger, G., Doxiadis, I., Thomas, D.Y., Bandlow, W., and Kaudewitz, F. (1977). A simple method for the large-scale

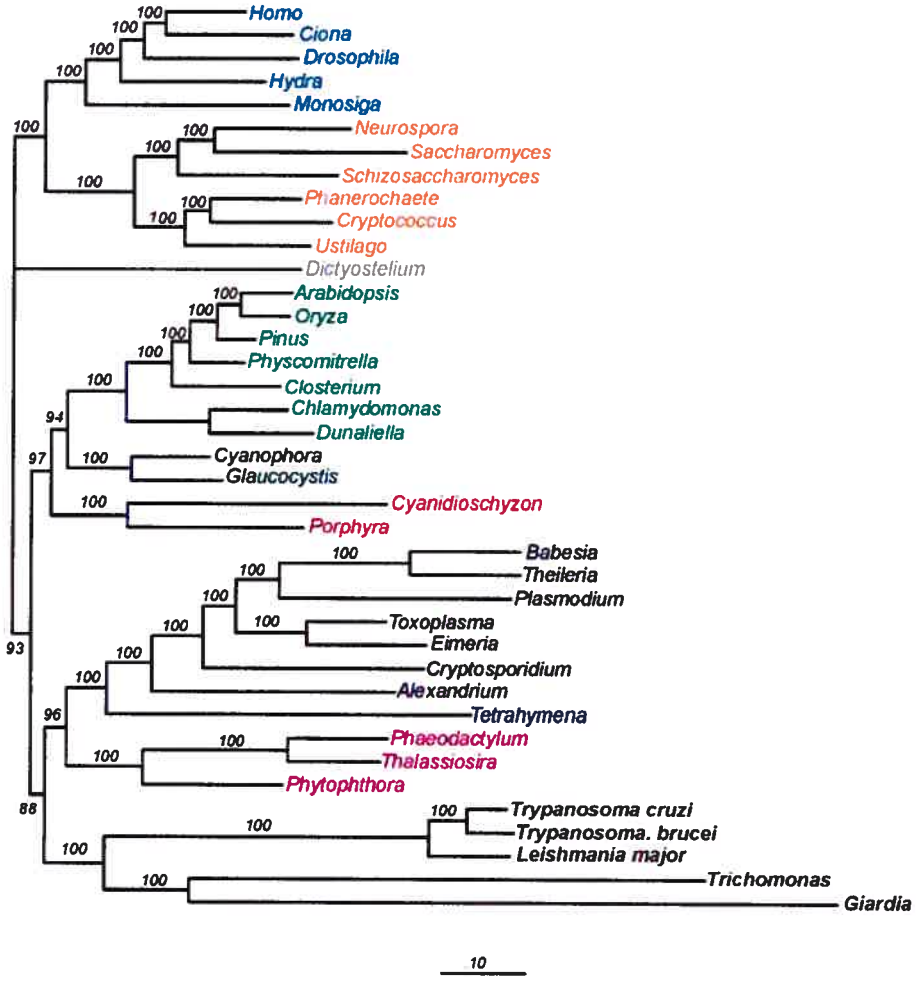

Figure S4. Phylogeny Based on Nuclear-Encoded Protein Sequences Including Excavates

The analysis is based on the concatenated dataset of nuclear-encoded proteins (143 proteins; 30,113 amino acid positions). The tree was inferred by BI with the WAG+F+Γ model. Numbers represent support values obtained with 100 bootstrap replicates on the concatenated dataset with PhyML. The scale bar denotes the number of amino acid substitution per site. Taxon designations are as follows: *Trichomonas*, *Trichomonas vaginalis*; *Giardia*, *Giardia lamblia*. For other taxa, see Figure 2 in the main text.

preparation of mitochondria from microorganisms. Anal. Biochem. *77*, 110–121.

S2. Zhulidov, P.A., Bogdanova, E.A., Shcheglov, A.S., Vagner, L.L., Khaspekov, G.L., Kozhemyako, V.B., Matz, M.V., Meleshkevitch, E., Moroz, L.L., Lukyanov, S.A., et al. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res. *32*, e37.

S3. Ewing, B., Green, P., Hillier, L., and Wendl, M.C. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. *8*, 186–194.

S4. Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. *8*, 175–185.

S5. Koski, L.B., Gray, M.W., Lang, B.F., and Burger, G. (2005). AutoFACT: An Automatic Functional Annotation and Classification Tool. BMC Bioinformatics *6*, 151 Published online June 16, 2005. 10.1186/1471-2105-6-151.

S6. King, N., Hittinger, C.T., and Carroll, S.B. (2003). Evolution of key cell signaling and adhesion protein families predates animal origins. Science *301*, 361–363.

S7. Adachi, J., and Hasegawa, M. (1996). MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. *28*, 1–150.

S8. Strimmer, K., and von Haeseler, A. (1996). Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. *13*, 964–969.

S9. Zhaxybayeva, O., Lapierre, P., and Gogarten, J.P. (2004). Genome mosaicism and organismal lineages. Trends Genet. *20*, 254–260.

S10. Brochier, C., Bapteste, E., Moreira, D., and Philippe, H. (2002). Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. *18*, 1–5.

S11. Yang, Z. (1996). Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42, 587–596.

S12. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Proceedings 2nd International Symposium on Information Theory, Csaki, ed. (Budapest: Akademia Kiado), 267–281.

CHAPITRE III : COMMENT DÉTECTER ET SURMONTER LES ERREURS SYSTÉMATIQUES

SOUS PRESSE DANS **SYSTEMATIC BIOLOGY**

DETECTING AND OVERCOMING SYSTEMATIC ERRORS IN GENOME-SCALE PHYLOGENIES

NAIARA RODRÍGUEZ-EZPELETA[1], HENNER BRINKMANN[1], BÉATRICE ROURE[1], NICOLAS LARTILLOT[2], B. FRANZ LANG[1] AND HERVÉ PHILIPPE[1]

[1] *Canadian Institute for Advanced Research, Centre Robert Cedergren, Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal, Québec, H3T 1J4, Canada.*

[2] *Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS-Université de Montpellier 2, 161, rue Ada, 34392 Montpellier Cedex 5, France.*

Detecting and overcoming systematic errors in genome-scale phylogenies.

Naiara Rodríguez-Ezpeleta[1], Henner Brinkmann[1], Béatrice Roure[1], Nicolas Lartillot[2], B. Franz Lang[1] and Hervé Philippe[1]

[1]*Canadian Institute for Advanced Research, Centre Robert Cedergren, Département de Biochimie, Université de Montréal, 2900 Boulevard Édouard-Montpetit, Montréal, Québec, H3T 1J4, Canada; E-mail:* [redacted]

[2] *Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS-Université de Montpellier 2, 161, rue Ada, 34392 Montpellier Cedex 5, France.*

Keywords: Phylogenomics; eukaryotic phylogeny; long-branch attraction; compositional heterogeneity; non-phylogenetic signal; systematic error; inconsistency; data removal
Running head: DETECTING AND OVERCOMING SYSTEMATIC ERRORS

Abstract.--- Genome-scale datasets result in an enhanced resolution of the phylogenetic inference by reducing stochastic errors. However, there is also an increase of systematic errors due to model violations, which can lead to erroneous phylogenies. Here, we explore the impact of systematic errors on the resolution of the eukaryotic phylogeny using a dataset of 143 nuclear-encoded proteins from 37 species. The initial observation was that, despite the impressive amount of data, some branches had no significant statistical support. To demonstrate that this lack of resolution is due to a mutual annihilation of phylogenetic and non-phylogenetic signals, we created a series of datasets with slightly different taxon sampling. As expected, these datasets yielded strongly supported but mutually exclusive trees, thus confirming the presence of conflicting phylogenetic and non-phylogenetic signals in the original dataset. To decide on the correct tree, we applied several methods expected to reduce the impact of some kinds of systematic error. Briefly, we show that (i) removing fast-evolving positions, (ii) recoding amino acids into functional categories, and (iii) using a site-heterogeneous mixture model (CAT), are three effective means of increasing the ratio of phylogenetic to non-phylogenetic signal. Finally, our results allow us to formulate guidelines for detecting and overcoming phylogenetic artefacts in genome-scale phylogenetic analyses.

The use of large multi-gene datasets to infer phylogenetic trees (phylogenomics) has been successfully applied to resolve evolutionary questions for which single-gene phylogenies failed (Bapteste et al., 2002; Delsuc et al., 2006; Delsuc et al., 2005; Madsen et al., 2001; Murphy et al., 2001; Philippe et al., 2005a; Qiu et al., 1999; Rodríguez-Ezpeleta et al., 2005; Soltis et al., 1999). This increase in resolution results from the reduction of sampling error through the addition of phylogenetically informative positions. However, higher statistical support does not necessarily lead to more accurate results, because the potential for systematic errors also grows with the increasing size of datasets, which in some cases may lead to strongly supported but incorrect phylogenies (Brinkmann et al., 2005; Jeffroy et al., 2006; Philippe et al., 2005b; Phillips et al., 2004; Stefanovic et al., 2004).

In the probabilistic framework (maximum likelihood and bayesian inference), systematic errors can be traced back to mis-specifications in the model of sequence evolution (model violations). Known causes of model violations are across-site rate variation (Yang, 1994), heterotachy (the across-site rate variation through time) (Kolaczkowski and Thornton, 2004; Philippe et al., 2005c; Spencer et al., 2005), site-interdependent evolution (Robinson et al., 2003; Rodrigue et al., 2005), compositional heterogeneity (Foster, 2004; Galtier and Gouy, 1995; Lockhart et al., 1992), and site-heterogeneous nucleotide/amino-acid replacement (Lartillot and Philippe, 2004; Pagel and Meade, 2004). In the following, we will call the apparent signal arising from such model violations "non-phylogenetic" signal, as opposed to genuine phylogenetic signal that corresponds to *bona fide* shared-derived characters.

The impact of model violations on phylogenetic accuracy is greatly exaggerated when multiple substitutions occur at given sites (mutational saturation). In the absence of model violation, mutational saturation would result in random sequences simply leading to poorly resolved trees (but see (Susko et al., 2005)). In contrast, when the model is violated, systematic error becomes manifest. Because long branches (either due to fast evolutionary rate or long time span) accumulate more multiple substitutions, they are most affected by long branch attraction (LBA), the well-known case of systematic error that provokes the clustering of fast-evolving species regardless of their true phylogenetic relationship (Felsenstein, 1978). Several complementary approaches have been applied to overcome systematic errors such as LBA: (i) increased taxon sampling and improved models of sequence evolution, allowing a more efficient detection of multiple substitutions, and (ii) removal of fast-evolving species (Aguinaldo et al., 1997), genes (Brinkmann et al., 2005; Philippe et al., 2005b) or sequence positions (Brinkmann and Philippe, 1999; Burleigh and Mathews, 2004; Hirt et al., 1999; Ruiz-Trillo et al., 1999).

In this paper, we study the relative contribution of phylogenetic and non-phylogenetic signal to genome-scale phylogenies and explore different methods to overcome systematic error. We use the global eukaryotic phylogeny as a case study for two reasons. First, the eukaryotic diversification is difficult to resolve, possibly because of closely spaced speciation events (Knoll, 1992; Philippe and Adoutte, 1998), implying that the phylogenetic signal would be limited, and second, multiple substitutions are expected given the long time span of eukaryotic evolution, most likely making non-phylogenetic signal significant.

Using a dataset of 143 nuclear encoded protein sequences from 37 eukaryotic species, we show that slight deviations in the evolutionary rate or amino acid composition of the sequences can lead to strongly supported but incorrect phylogenies. This occurs when the phylogenetic signal for a given branch is significantly weaker than the non-phylogenetic signal. Alternatively, when both signals are of equivalent strength, they may counterbalance each other, leading to unresolved trees, even with large datasets. We demonstrate that (i) variations in taxon sampling, (ii) removal of fast-evolving sites, (iii) use of a site-heterogeneous mixture model (Lartillot and Philippe, 2004), and (iv) amino acid coding into functional categories have the potential to overcome some types of systematic errors in genome-scale datasets.

MATERIALS AND METHODS

Phylogenetic analyses

The analyses were performed on a previously described dataset of 143 nuclear-encoded proteins (30,244 amino acid positions) from 39 eukaryotic species (Rodríguez-Ezpeleta et al., 2005), excluding the two fastest evolving lineages (*Trichomonas vaginalis* and *Giardia lamblia*). Trees were inferred using maximum parsimony (MP), bayesian inference (BI) and maximum likelihood (ML) methods. The alignments (including corresponding trees) have been submitted to TreeBASE under accession numbers SN3166-13372 to SN3166-13377.

Heuristic analyses

MP analyses were performed using PAUP* (Swofford, 2000), with tree bisection and reconnection search and 10 random additions of species. The support was evaluated based on 1,000 bootstrap replicates. BI analyses were conducted using MrBayes 3.0 b4 (Ronquist and Huelsenbeck, 2003) or PhyloBayes (http://www.lirmm.fr/mab/article.php3?id_article=329). MrBayes analyses were performed with the WAG amino acid replacement matrix (Whelan and Goldman, 2001), gamma distributed rates across sites (4 discrete categories), and stationary amino acid frequencies estimated from the dataset (WAG+F+ Γ4 model). Three independent analyses with 120,000 generations gave identical results. PhyloBayes analyses were performed with the CAT mixture model, which accounts for across-site heterogeneities in the amino-acid replacement process (Lartillot and Philippe, 2004). Two independent runs were performed with a total length of 2,500 cycles (250 topological moves per cycle) with the same operators as in Lartillot et al. (2006). The first 500 points were discarded as burn-in, and the posterior consensus was computed on the 2,000 remaining trees. Preliminary ML analyses were performed on the concatenated dataset using heuristic searches with PhyML 2.4 (Guindon and Gascuel, 2003) and TreeFinder (Jobb et al., 2004) with the WAG+F+Γ4 model. The support was evaluated based on 100 bootstrap replicates.

Exhaustive analyses

The probability of getting trapped in a local minimum during heuristic topology searches is high for large datasets (Salter, 2001), but an exhaustive search is impossible in our case given the large number of possible topologies for 37 species (10^{49}). This problem was addressed by constraining relationships supported by consistently more than 95% bootstrap values (MP and ML) and 1.0 posterior probability (BI) (opisthokonts –animals, choanoflagellates and fungi, red algae, green plants, glaucophytes, apicomplexans, stramenopiles and kinetoplastids). This reduces the number of topologies to be exhaustively analysed to 135,135. To further alleviate computational cost and memory usage, we proceeded in two steps. First, exhaustive ML analyses without taking rates across-sites variation into account were performed with PROTML (Adachi and Hasegawa, 1996) and the JTT amino acid replacement matrix (Jones et al., 1992) for each protein separately (for details see Rodriguez-Ezpeleta et al., 2005). The resulting 135,135 tree topologies were sorted by likelihood value, and the top 1,733 trees were selected. These trees were augmented by sampling every 500th subsequent topology, for a total of 2,000 trees. For these 2,000 trees, likelihood values were calculated with TREE-PUZZLE (Schmidt et al., 2002) and the concatenated WAG+F+Γ4 model (all parameters estimated for the concatenated dataset). We verified that retention of the 1,733 top ranking topologies was sufficient: first, the correlation between the likelihood values obtained with the separate JTT+F and the concatenated WAG+F+Γ4 models for the 2,000 selected topologies is excellent (R^2=0.9693; Fig. S1); second, the order of topologies is almost identical with and without considering rates across sites; and third, the 9 best topologies from the separate JTT+F analysis receive a total of 98% of the RELL bootstrap support (Kishino et al., 1990) (the 83 best topologies receive 100% of the RELL bootstrap support). Indeed, retaining 100 topologies gives virtually identical results (not shown). In order to estimate statistical support for each branch the RELL bootstrap method (Kishino et al., 1990) was used. In brief, site-wise likelihood values were calculated with PAML (Yang, 1997) with the concatenated WAG+F+Γ4 model and used to perform RELL bootstrap analyses with 10,000 replicates.

The relationship between the number of sequence positions and the bootstrap support values (BVs) was calculated as described (Lecointre et al., 1994). Briefly, different numbers of positions (3,000; 6,000, etc.) were randomly drawn from the complete dataset 100 times. RELL bootstrap values (100 replicates) were then computed for each of the 100 samples and for each size fraction (site-wise likelihoods were not recomputed for each sample for obvious computation time reasons, but are expected to be similar with this large number of positions; see below). The average of the BV of all branches for each size fraction was plotted against its size.

Removal of fast evolving sites

Fast evolving sites were identified using a modification of the method proposed by Ruiz-Trillo *et al.* (1999) and Burleigh and Mathews (2004). Instead of eliminating sites according to the discrete gamma category to which they most likely belong, they were eliminated according to their site-wise rates calculated by PAML (i.e., weighted average rates over all categories with the weights given by the posterior probabilities of each

category)on the concatenated dataset for each topology. Sites were then sorted according to (i) the rates estimated on a given topology or to (ii) the mean of the rates estimated on all topologies. Then, fast-evolving sites were progressively removed in steps of 1,000. RELL bootstrap analyses (1,000 replicates) were performed after each step, and the resulting values plotted against the alignment size.

The computational burden associated with site removal is only circumvented if the BVs are computed using the RELL method. However, two important assumptions of this method may be violated if too many sites are removed: (i) the parameters of the model estimated on the complete dataset (in particular, branch lengths) should remain similar for the reduced dataset and (ii) the topological constraints imposed should remain valid. First, the constraints were verified after the removal of 15,000 and 20,000 sites by performing heuristic analyses with TreeFinder; and second, the parameters and the site-wise likelihoods were re-estimated on these two datasets. After the removal of 15,000 sites (half of the dataset) all constraints are still respected, and the results obtained with and without re-estimating site-wise likelihood values are similar (the correlation coefficient between BVs is 0.86). However, after removal of 20,000 sites, some of the constraints are no longer supported (e.g., the sister group of apicomplexans and ciliates), and the RELL bootstrap values obtained before and after parameter re-estimation differ substantially. We thus stopped after the removal of 15,000 sites in all analyses.

Testing for saturation

The saturation of the alignments was measured by plotting the number of observed differences (p distances) against the number of substitutions that are computed as patristic distances (in our case, derived from the ML tree) using TREEPLOT of the MUST package (Philippe, 1993; Philippe et al., 1994). Both distance matrices were compared, and the slope of the graph was calculated using the COMP_MAT program in the MUST package. The greater the number of inferred substitutions with respect to the number of observed differences (small slope), the greater the saturation of the data (Jeffroy et al., 2006).

Compositional heterogeneity

The amino acid composition bias of the species in the dataset was visualized by assembling a 37 x 20 matrix containing the percentage of each amino acid per species using the NET program from the MUST package (Philippe, 1993). This matrix is displayed as a two dimensional plot in a Principal Components Analysis (PCA), as implemented in the SAS program (SAS, 1999). To calculate the overall compositional bias in the data, the Bowker's test for compositional symmetry (Ababneh et al., 2006; Bowker, 1948) was applied. Bowker's values were calculated for each pair of sequences and the median value was computed. Large Bowker's values indicate strong heterogeneity in the dataset, whereas lower Bowker's values indicate that the sequence composition is homogeneous (note that the phylogenetic dependency among all Bowker's values is not corrected for here).

Two attempts to reduce the potential impact of compositional bias were performed, by (i) constructing neighbour-joining trees based on LogDet + Γ pair-wise distances, calculated with the LDDist perl module (Thollesson, 2004) and using the rate categories estimated by TREE-PUZZLE; and (ii) recoding the data using the common six groups of amino acids that usually replace one another (Hrdy et al., 2004). To allow for a general-time-reversible (GTR) matrix implemented in most programs, the dataset was recoded to four categories instead of six, by combining aromatic (FYW) and hydrophobic (MVIL) amino acids and coding the rare cysteine as missing data. The four amino acid categories were named A, T, G and C, respectively, and the parameters of the GTR matrix were estimated by PAUP. The 2,000 best topologies from the exhaustive search were analyzed by TREE-PUZZLE with a GTR+F+Γ4 model. RELL bootstrap (10,000 replicates) analyses were performed as described above. The constraints were verified after the recoding with heuristic ML analyses using TreeFinder.

Results and Discussion

Phylogenomic analyses do not resolve every branch

Fig. 1 shows the ML tree based on 143 nuclear protein-coding genes (30,244 amino acid positions) from 37 eukaryotic species. The monophyly of all major eukaryotic groups and the relationships within them are recovered with 100% bootstrap support value (BV) and are in agreement with current knowledge of eukaryotic evolution (Baldauf et al., 2000; Simpson and Roger, 2004), underlining that the use of a large number of genes notably improves overall statistical support. Only four branches receive BVs below 100%. Among them, the monophyly of primary photosynthetic eukaryotes or Plantae (green plants, rhodophytes and glaucophytes) requires special attention. This grouping has already been suggested based on genomic features and molecular phylogenies of plastid and nuclear proteins (Cai et al., 2003; Huang and Gogarten, 2006; McFadden and van Dooren, 2004; Moreira et al., 2000; Rodríguez-Ezpeleta et al., 2005); however, with a particular taxon sampling (Fig. 1), it only receives statistically non-significant support (64% BV).

Unsupported trees are usually attributed to a lack of phylogenetic information in the data, suggesting that the addition of more genes or positions will increase resolution (Bapteste et al., 2002; Rodríguez-Ezpeleta et al., 2005; Rokas et al., 2003; Saitou and Nei, 1986). Therefore, we studied the variation of the BVs obtained for the monophyly of Plantae with respect to the number of amino acid positions considered. As shown in Fig. 2 (open triangles), the BVs rapidly increase as more positions are added. But when more than about 10,000 amino acid positions are considered, the BVs attain a plateau, suggesting that the addition of more data (even of complete genome sequences) will most likely not lead to a statistically significant support for the monophyly of Plantae, given this taxon sampling and this tree reconstruction method. In fact, an alternative grouping, the sister group relationship of red algae and kinetoplastids (Fig. 2; close circles), displays very similar behaviour, raising rapidly to a plateau of less than 40% BV.

The shape of the curves obtained in Fig. 2 suggests that the unsupported monophyly of Plantae is not due to a lack of phylogenetic signal. Rather, it seems as if two competing

signals exist in the data: one that supports the monophyly of Plantae and another one that supports a sister-group relationship between red algae and kinetoplastids.

Coexistence of phylogenetic and non-phylogenetic signal in the data

Since kinetoplastids present the longest unbroken branch in the dataset (Fig. 1), the hypothesis of a LBA artefact as the cause for their clustering with red algae can be advanced. To test if the two red algae (*Cyanidioschyzon* and *Porphyra*, both have moderate evolutionary rate differences) are differently affected by this artefact, two datasets were created, either including *Porphyra* or *Cyanidioschyzon* as the single representative of the red algae. Surprisingly, the use of one or the other red algae has drastic effects on the outcome. With *Porphyra* as the sole red algal representative, the BV for Plantae raises from 64% to 99% (Fig. 3a), whereas with *Cyanidioschyzon* alone, the support for Plantae drops to 0% and the support for the sisterhood of red algae and kinetoplastids raises to 100% (Fig. 3b). Since we share the view with others that red algae are indisputably monophyletic (e.g., Ragan and Gutell (1995); see also Fig. 1), one of the two trees in Fig. 3 has to be wrong. Because *Cyanidioschyzon* evolves somewhat faster than *Porphyra*, our working hypothesis is that the monophyly of Plantae observed in Fig. 3a is the product of genuine phylogenetic signal, whereas the grouping of red algae and kinetoplastids (Fig. 3b) is an LBA artefact.

As *Cyanidioschyzon* evolves only 1.25 times faster than *Porphyra* (Fig. 1), the radical difference in the resulting tree topologies (Figs. 3a,b) may seem surprising. We posit that this can be explained by the large number of amino acid positions in this dataset. More than 15,000 amino acid positions (Fig. 3d) are required to recover the sister-group of *Cyanidioschyzon* and kinetoplastids with BV > 95%, suggesting that the non-phylogenetic signal is weak; however, the phylogenetic signal for the monophyly of Plantae is as weak (Fig. 3c).

Testing the LBA hypothesis using differences in taxon sampling

If the grouping of *Cyanidioschyzon* and the kinetoplastids is due to LBA, this artefact should be reproduced with other long unbroken branches in this dataset. To explore this hypothesis, three combinations of taxa were created that induce long unbroken branches. Starting from the dataset of Fig. 3b, the kinetoplastids were removed and (i) *Saccharomyces* was kept as the only representative of the *Dictyostelium*/opisthokont clade, (ii) either *Theileria* and *Phytophthora* or (iii) *Plasmodium* and *Phytophthora* were kept as the only representatives of alveolates and stramenopiles, respectively.

In all cases, only *Cyanidioschyzon* is attracted to the longest unbroken branch (Fig. 4). Importantly, Plantae remain monophyletic in these three cases when *Porphyra* is used (Fig. S2). This confirms that the grouping of kinetoplastids and *Cyanidioschyzon* is due to LBA. Surprisingly, the grouping of *Plasmodium* and *Cyanidioschyzon* receives only 66% BV (Fig. 4c), whereas the grouping of *Theileria* and *Cyanidioschyzon* (Fig. 4b) has 90% BV. Interestingly, in a Principal Components Analysis (Fig. 5), the amino-acid composition of *Cyanidioschyzon* is most similar to that of *Saccharomyces* and kinetoplastids, less to *Theileria*, and least to *Plasmodium* - the species with the most extreme genomic A+T

content (80.6%). Therefore, even if the two alveolates (*Theileria* and *Plasmodium*) have almost the same evolutionary rate (Fig. 1), the extreme compositional bias in *Plasmodium* appears to have an additional effect on the bootstrap support (Fig. 4b,c).

Extracting phylogenetic signal by removing fast evolving sites

Because fast evolving sites are more likely to be saturated and prone to accumulation of non-phylogenetic signal, a progressive removal of such sites should decrease artefacts caused by model violations (Brinkmann and Philippe, 1999; Burleigh and Mathews, 2004; Olsen, 1987; Ruiz-Trillo et al., 1999). We studied the impact of the fast sites in our dataset by progressively removing blocks of the fastest evolving sites.

The estimation of site-specific rates requires the knowledge of a tree topology. To avoid circularity, we used the best (ML) topology obtained with a dataset that does not include the red algae (which we cannot place with confidence with all the data). The experiment was performed on the datasets from Figs. 3b, 4a-c. In three cases, the removal of the fast evolving sites strengthens the support for the monophyly of Plantae and lowers the one for the alternative position (Fig. 6a, c, d), confirming that the removal of the fast evolving sites increases the ratio of phylogenetic to non-phylogenetic signal (Brinkmann and Philippe, 1999; Brochier and Philippe, 2002). With *Saccharomyces* as the only representative of opisthokonts and Amoebozoa, the removal of the fastest evolving sites is insufficient to recover the monophyly of Plantae, although a small increase in the BV is observed (Fig. 6b). The number of sites that need to be removed to recover this relationship is different in each case, which may result from different levels of non-phylogenetic signal in various datasets.

For the experiments described above, a tree topology without red algae was used to calculate site-wise rates (to avoid introduction of bias). The procedure is justified in this special case, where a single taxon is added to an otherwise unquestioned topology, but should probably not be applied when more complex changes are expected. To test if the choice of tree topology significantly effects the estimation of site-wise rates, results were compared for the red algae+kinetoplastids (Fig. 3b) and the Plantae topology (Fig. 3a). When the rates were estimated on the red algae+kinetoplastids topology, the removal of the fastest evolving sites does not improve phylogenetic accuracy (Fig. S3); in contrast, if the rates were estimated on the Plantae topology, the removal of even fewer sites than in Fig. 6 leads to recovery of the correct topology (Fig. S4). Evidently, the specific topology used to estimate the rates heavily influences the results. As a solution to this problem, we propose to use the mean site-wise rates estimated for a given set of best topologies. In our specific example, with the 2,000 topologies, results are virtually identical to the experiment in which a tree without red algae was used (Fig. S5). This "mean rate approach" is an interesting avenue that deserves further investigation.

Fast evolving sites are mutationally saturated and compositionally biased

For each of the non-overlapping windows of 1,000 sites that have been progressively removed, the mutational saturation and the compositional bias were studied. As expected, the mutational saturation (grey line in Fig. 7) is tightly correlated to the evolutionary rates, confirming that the fast-evolving sites are the most saturated. Since the effects of model violations are more evident in mutationally saturated sites, the removal of the fastest evolving sites efficiently overcomes systematic errors. We also measured a well-known source of model violation, the compositional heterogeneity among lineages. For each of the successively removed blocks of 1,000 positions, the Bowker's test for compositional symmetry was computed (black line in Fig. 7). Interestingly, the compositional heterogeneity is tightly correlated with the rate of the sites: the most saturated sites are the most compositionally biased. Therefore, by removing the fast evolving sites, we most likely overcome systematic error due to compositional heterogeneity. Accordingly, other sources of model violations may also be decreased by fast evolving site removal, and this question deserves further studies.

The effects of model violations

Another kind of model violation that may result in phylogenetic artefacts is the heterogeneity of the amino-acid replacement process across sites (Baurain et al., 2006; Koshi and Goldstein, 2001; Lartillot et al., 2006; Lartillot and Philippe, 2004; Pagel and Meade, 2004). Most sites of a protein show substitutions among a small set of 2 to 4 biochemically equivalent amino acids (Miyamoto and Fitch, 1996). However, homogeneous models inherently assume that, under maximal saturation, all 20 amino acids are likely to be observed at any given site with probabilities equal to the equilibrium frequencies. As a result, the probability of convergence is strongly under-estimated by standard models of evolution (Chang, 1996; Felsenstein, 2004).

The site-heterogeneous mixture model, CAT (Lartillot and Philippe, 2004), was applied to the various taxon samplings previously studied. The monophyly of Plantae was recovered in all but two cases even when *Cyanidioschyzon* is the only red alga. In particular, in two cases where the homogeneous model fails (Figs. 4b, c), the inference with a more complex model is not sensitive to systematic error: when *Theileria* or *Plasmodium* are the only representatives of alveolates, Plantae where supported by a posterior probability (pp) of 0.98 and 0.85, respectively. Nevertheless, the monophyly of Plantae was not recovered (pp=0) in the presence of kinetoplastids or of *Saccharomyces* as LBA attractors. Therefore, site-specific substitution pattern is not the only cause of the observed artefacts.

An alternative potential source of model violation is the non-stationarity of the amino-acid replacement process, known to affect our dataset (Fig. 5). Under a stationary model, where the same amino acid or nucleotide composition is assumed along the tree, compositional heterogeneity may drastically mislead phylogenetic reconstruction (Hasegawa and Hashimoto, 1993; Hendy and Penny, 1989; Lockhart et al., 1992; Phillips et al., 2004). Although models have been developed to overcome this violation (e.g. Foster,

2004; Galtier and Gouy, 1995; Yang and Roberts, 1995; Blanquart and Lartillot, 2006), they are computationally demanding, and have implementation limitations, and are therefore of limited value. Other ways to overcome non-phylogenetic signal due to compositional heterogeneity have been reported, such as the LogDet method (Lake, 1994; Lockhart et al., 1994) and the RY (Woese et al., 1991) or Dayhoff (Hrdy et al., 2004) coding for nucleotides and amino acids.

Interestingly, amino acid coding into functional categories has an impact on both kinds of model violations mentioned above, i.e., compositional effects and the expected number of amino acids per position. First, it alleviates compositional bias (Phillips et al., 2004; Woese et al., 1991). For example, Lysine (K) and Arginine (R) are two easily interchangeable amino acids whose codons differ at a single position (AAR and AGR, respectively), and that are preferred in AT- and GC- rich genomes respectively. Hence, coding pairs or groups of amino acids such as K and R as a single character state should compensate for these biases. Second, the recoding will also alleviate the problem of homoplasies that occur in peaked biochemical profiles, by reducing the number of character states from 20 to 4.

Applied to our dataset, the LogDet method failed to recover the expected tree topology, and a strong LBA artefact unites alveolates and kinetoplastids to the exclusion of stramenopiles, a grouping that attracts *Cyanidioschyzon*. In fact, it has already been suggested that the LogDet method may fail in the presence of rate heterogeneity among sites or lineages (Conant and Lewis, 2001). Instead, a modification of the Dayhoff coding (Hrdy et al., 2004; see Material and Methods) increases the support for Plantae while decreasing the attraction of *Cyanidioschyzon* with long unbroken branches (Fig. 8), in all four cases (Figs. 3b, 4a-c). Importantly, with amino acid recoding, Plantae monophyly was recovered with *Saccharomyces* as the only representative of Opisthokonts, when all other methods failed.

Altogether, the overall pattern suggests that the artefacts observed in this dataset are mainly caused by a combination of compositional bias and site -heterogeneity that operate at different levels, depending on the attractor: a site-heterogeneity violation, dominant in the case of *Plasmodium* and *Theileria*, and possibly compositional bias with kinetoplastids and *Saccharomyces*. As discussed above, recoding is efficient in alleviating both sources of systematic error simultaneously, although it reduces the phylogenetic signal considerably.

Conclusion

The common view that using genome-scale datasets is a universal remedy for resolving phylogenetic questions (e.g. Rokas et al., 2003) is inaccurate. Tree reconstruction artefacts that are invisible in single-gene phylogenies may become dominant in large datasets (Jeffroy et al., 2006). Depending on the relative contribution of phylogenetic and non-phylogenetic signal, certain genome-scale datasets may either lead to predicting incorrect tree topologies with confidence, or one or more branches remain unresolved whatever the data size.

The identification of 'misbehaving' species that contribute an unproportional fraction of non-phylogenetic signal is possible through variations in taxon sampling. Removal of these species from the dataset has been common practice to overcome some phylogenetic artefacts. Alternatively, more general approaches include data recoding, removal of fast-evolving sites, or the use of more realistic models of sequence evolution. Yet, current implementations of these procedures will either eliminate much of the phylogenetic signal, or are impracticable in terms of computational load. In practical terms, we therefore recommend a combined application of all methods that will overcome, at least, some of the well-known types of systematic errors. Evidently, these approaches cannot address all kinds of systematic error present in a dataset; for example, none of the techniques applied here detect or overcome heterotachy (rate heterogeneity across sites through time).

Ultimately, the development of more sophisticated models of sequence evolution that address simultaneously the different kinds of systematic biases will reduce the requirement for intense user intervention by making best use of phylogenetic signal.

ACKNOWLEDGEMENTS

We wish to thank Denis Baurain for assistance with the tests for amino acid composition and helpful comments on a previous version of the manuscript, and Pablo Vinuesa, Frank (Andy) Anderson, Andrew J. Roger and two anonymous reviewers for useful suggestions. This work has been supported by operating and equipment funds from Genome Quebec/Canada. B.F.L. and H.P. acknowledge the program in Evolutionary Biology of the Canadian Institute for Advanced Research (CIAR) for salary and interaction support, and the Canada Research Chairs Program and the Canadian Foundation for Innovation (CFI) for salary and equipment support. N.R.E. has been supported by 'Programa de Formación de Investigadores del Departamento de Educación, Universidades e Investigación' (Government of Basque Country) and B.R. by 'Bourses d'Excellence biT' (CIHR).

REFERENCES

Ababneh, F., L. S. Jermiin, C. Ma, and J. Robinson. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225-1231.

Adachi, J., and M. Hasegawa. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. 28:1-150.

Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489-93.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972-7.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A 99:1414-9.

Baurain, D., H. Brinkmann, and H. Philippe. 2006. Lack of Resolution in the Animal Phylogeny: Closely Spaced Cladogeneses or Undetected Systematic Errors? Mol Biol Evol.

Blanquart, S., and N. Lartillot. 2006. A bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol Biol Evol 23:2058-71.

Bowker, A. H. 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc. 43:572-574.

Brinkmann, H., and H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817-25.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743-57.

Brochier, C., and H. Philippe. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature 417:244.

Burleigh, J. G., and S. Mathews. 2004. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. Am J Bot 91:1599-1613.

Cai, X., A. L. Fuller, L. R. McDougald, and G. Zhu. 2003. Apicoplast genome of the coccidian *Eimeria tenella*. Gene 321:39-46.

Chang, J. T. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math Biosci 134:189-215.

Conant, G. C., and P. O. Lewis. 2001. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. Mol Biol Evol 18:1024-33.

Delsuc, F., H. Brinkmann, D. Chourrout, and H. Philippe. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965-8.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361-75.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401-410.

Felsenstein, J. 2004. Inferring phylogenies, Sunderland, Massachusetts.

Foster, P. G. 2004. Modeling compositional heterogeneity. Syst Biol 53:485-95.

Galtier, N., and M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92:11317-21.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696-704.

Hasegawa, M., and T. Hashimoto. 1993. Ribosomal RNA trees misleading? Nature 361:23.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst Zool 38:297-309.

Hirt, R. P., J. M. Logsdon, Jr., B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci U S A 96:580-5.

Hrdy, I., R. P. Hirt, P. Dolezal, L. Bardonova, P. G. Foster, J. Tachezy, and T. M. Embley. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432:618-22.

Huang, J., and J. P. Gogarten. 2006. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. Trends Genet 22:361-6.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends Genet.

Jobb, G., A. von Haeseler, and K. Strimmer. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol 4:18.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-82.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. J. Mol. Evol. 31:151-160.

Knoll, A. H. 1992. The early evolution of eukaryotes: a geological perspective. Science 256:622-7.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980-4.

Koshi, J. M., and R. A. Goldstein. 2001. Analyzing site heterogeneity during protein evolution. Pac Symp Biocomput:191-202.

Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci U S A 91:1455-9.

Lartillot, N., H. Brinkmann, and H. Philippe. 2006. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol In press.

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-109.

Lecointre, G., H. Philippe, H. L. Van Le, and H. Le Guyader. 1994. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. Mol Phylogenet Evol 3:292-309.

Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, and A. W. Larkum. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol 34:153-62.

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605-612.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610-4.

McFadden, G. I., and G. G. van Dooren. 2004. Evolution: red algal genome affirms a common origin of all plastids. Curr Biol 14:R514-6.

Miyamoto, M. M., and W. M. Fitch. 1996. Constraints on protein evolution and the age of the eubacteria/eukaryote split. Syst Biol 45:568-75.

Moreira, D., H. Le Guyader, and H. Philippe. 2000. The origin of red algae and the evolution of chloroplasts. Nature 405:69-72.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, M. S. Springer, D. J. Kao, R. W. DeBry, R. Adkins, and H. M. Amrine. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348-51.

Olsen, G. J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harb Symp Quant Biol 52:825-37.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571-81.

Philippe, H. 1993. MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res 21:5264-72.

Philippe, H., and A. Adoutte. 1998. The molecular phylogeny of Eukaryota: solid facts and uncertainties. Pages 25-56 *in* Evolutionary relationships among Protozoa (G. Coombs, K. Vickerman, M. Sleigh, and A. Warren, eds.). Kluwer, Dordrecht.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005a. Phylogenomics. Annual Review of Ecology, Evolution, and Systematics 36:541-562.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005b. Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. Mol Biol Evol.

Philippe, H., U. Sörhannus, A. Baroin, R. Perasso, F. Gasse, and A. Adoutte. 1994. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. J. Evol. Biol. 7:247-265.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005c. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5:50.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455-8.

Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402:404-7.

Ragan, M., and R. Gutell. 1995. Are red algae plants? Bot J Linnean Soc 118:81-105.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20:1692-704.

Rodrigue, N., N. Lartillot, D. Bryant, and H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347:207-17.

Rodríguez-Ezpeleta, N., H. Brinkmann, S. C. Burey, B. Roure, G. Burger, W. Loffelhardt, H. J. Bohnert, H. Philippe, and B. F. Lang. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol 15:1325-30.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-4.

Ruiz-Trillo, I., M. Riutort, D. T. Littlewood, E. A. Herniou, and J. Baguna. 1999. Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science 283:1919-23.

Saitou, N., and M. Nei. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J Mol Evol 24:189-204.

Salter, L. A. 2001. Complexity of the likelihood surface for a large DNA dataset. Syst Biol 50:970-8.

SAS. 1999. SAS/STAT User's guide, version 8.12. SAS institute Inc.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-4.

Simpson, A. G., and A. J. Roger. 2004. The real 'kingdoms' of eukaryotes. Curr Biol 14:R693-6.

Soltis, P. S., D. E. Soltis, and M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402-4.

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol 22:1161-4.

Stefanovic, S., D. W. Rice, and J. D. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol Biol 4:35.

Susko, E., M. Spencer, and A. J. Roger. 2005. Biases in phylogenetic estimation can be caused by random sequence segments. J Mol Evol 61:351-9.

Swofford, D. L. 2000. PAUP*: Phylogenetic analysis using parsimony and other methods, version 4b10. Sinauer Associates, Sunderland, Massachusets.

Thollesson, M. 2004. LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. Bioinformatics 20:416-8.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691-9.

Woese, C. R., L. Achenbach, P. Rouviere, and L. Mandelco. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. Syst Appl Microbiol 14:364-71.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306-14.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-6.

Yang, Z., and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol Biol Evol 12:451-8.

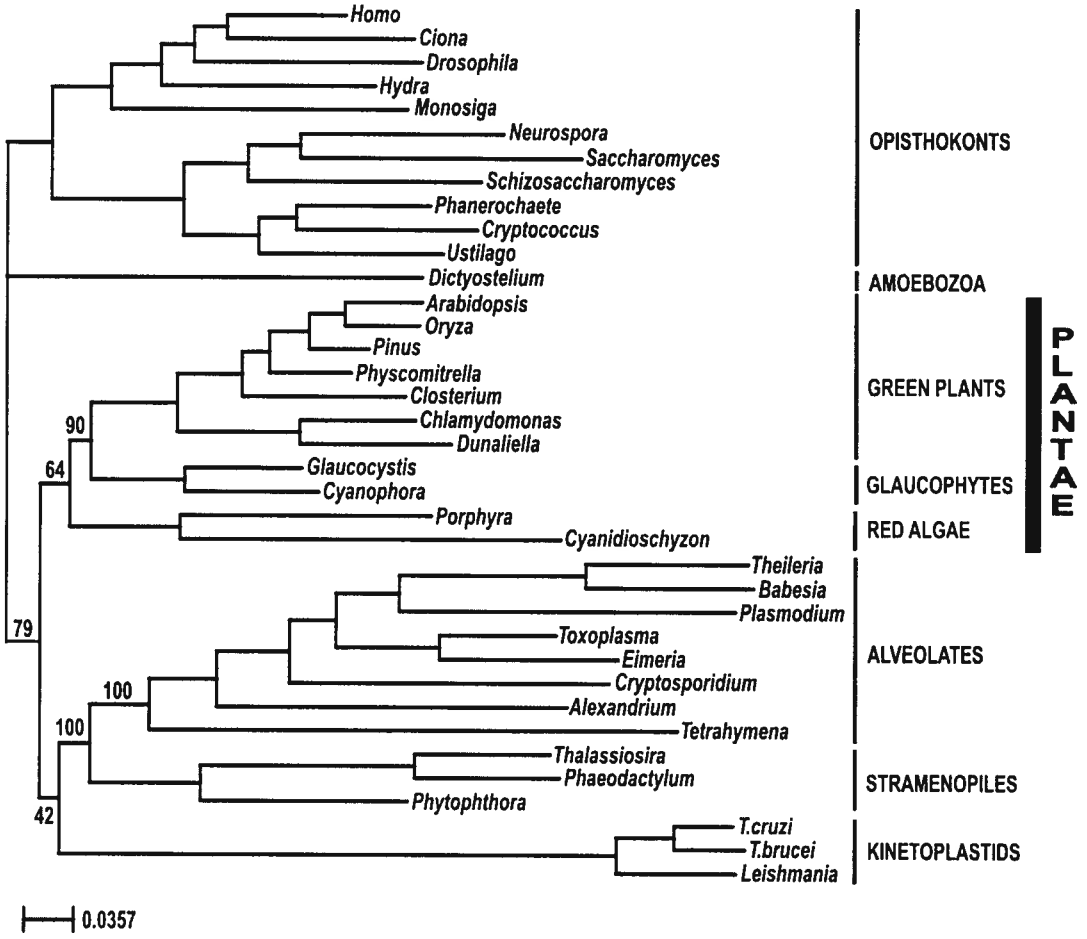

FIGURE 1. Eukaryotic phylogeny based on 143 nuclear-encoded proteins (30,244 amino acid positions) inferred by exhaustive ML analysis with the concatenated WAG+F+Γ4 model. The same topology is obtained with PhyML, TreeFinder, and MrBayes. Numbers indicate bootstrap values obtained by analysing 10,000 RELL replicates on the exhaustive ML analysis. Branches without values are supported by BVs of 100 and posterior probabilities of 1.0 in the ML (PhyML and TreeFinder) and BI (MrBayes) analyses respectively, and were constrained in the exhaustive analysis. The scale bar denotes the estimated number of amino acid substitution per site.

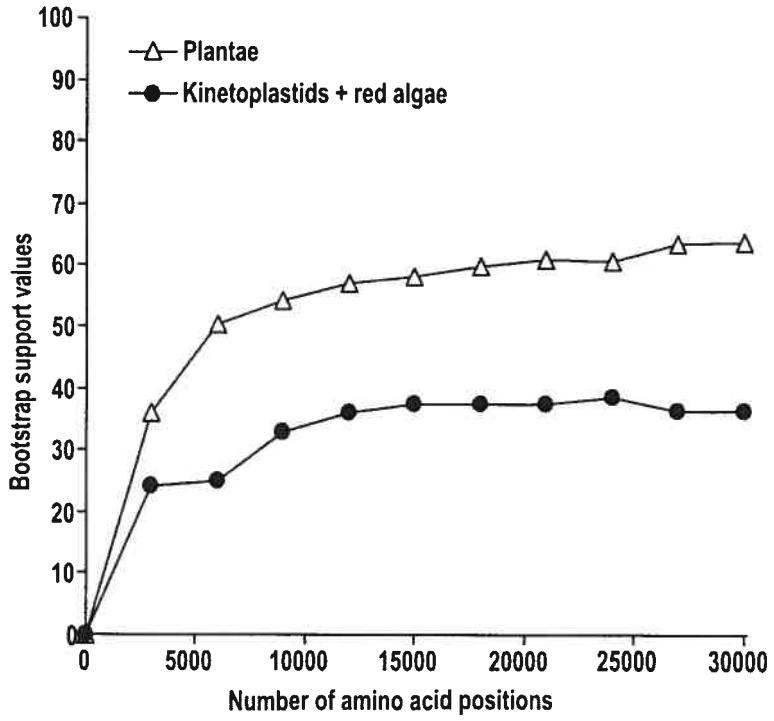

FIGURE 2. Bootstrap values for the monophyly of Plantae (triangles) and the sisterhood of red algae and kinetoplastids (dots), as a function of the number of amino acid positions. Bootstrap values were obtained by sampling different numbers of positions (3,000; 6,000, etc) 100 times, and by averaging the RELL bootstrap values (100 replicates) for samples of the same size.

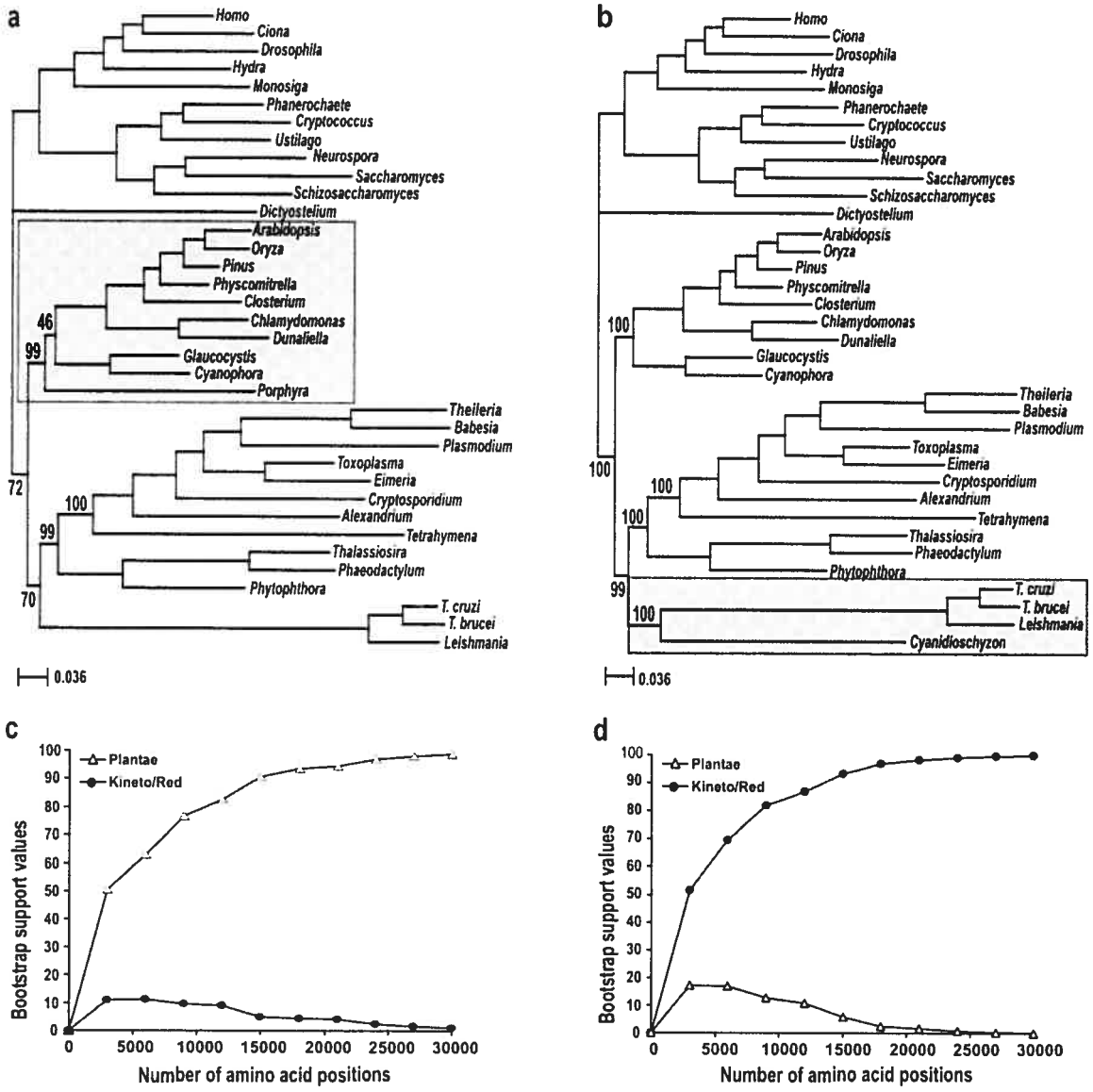

FIGURE 3. Alternative topologies obtained as described in Fig. 1 when only *Porphyra* (a) or only *Cyanidioschyzon* (b) was used to represent the red algae. No value above branch indicates that the corresponding node was supported at 100% BV in the ML analyses with PhyML and TreeFinder, and was constrained in the exhaustive analysis. Grey shaded areas indicate the alternative positions of red algae. For each dataset, the bootstrap values of the two alternative positions for red algae were plotted against the number of amino acid positions (c and d).

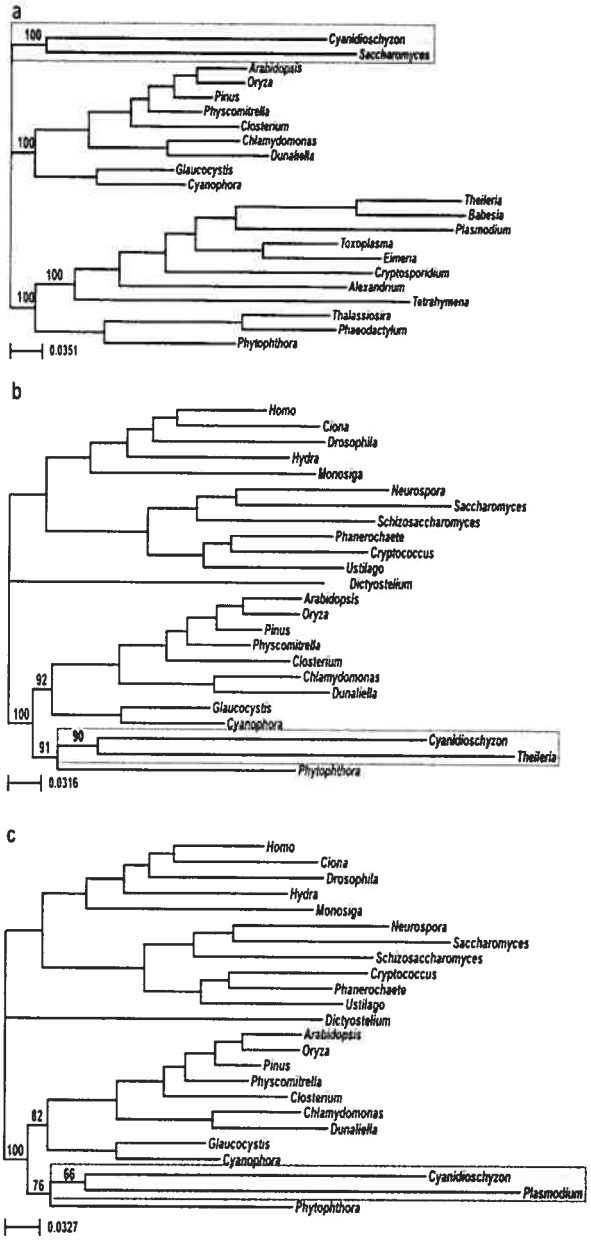

FIGURE 4. Same analyses as in Fig. 1, but with selected combinations of taxon samplings that are likely to induce an LBA artefact. In all cases, only *Cyanidioschyzon* was used to represent the red algae, and the kinetoplastids were excluded from the dataset; (a) using *Saccharomyces* as the only representative of the opisthokonts and Amoebozoa; using either (b) *Theileria* and *Phytophthora,* or (c) *Plasmodium* and *Phytophthora* as the representatives of alveolates and stramenopiles respectively. No value above branches indicates that the corresponding node was supported at 100% BV in the ML analyses with PhyML and TreeFinder, and was constrained in the exhaustive analysis. Grey shaded areas indicate the position *Cyanidioschyzon*.

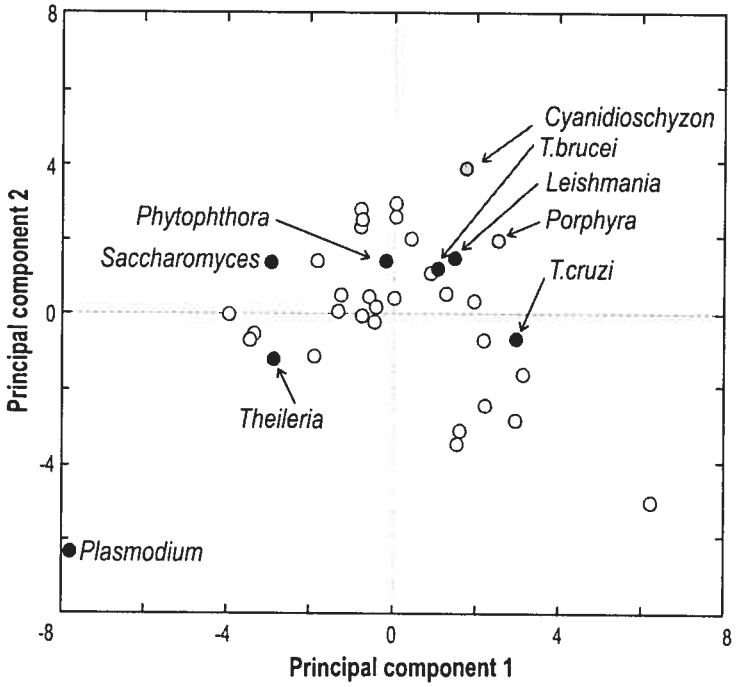

FIGURE 5. Reduced dimensionality plot showing the main principal components of the global amino acid compositions. The variances that explain the two first axes are respectively 32% and 25%. Grey circles denote the two red algae and black circles are other relevant species used in previous analyses.

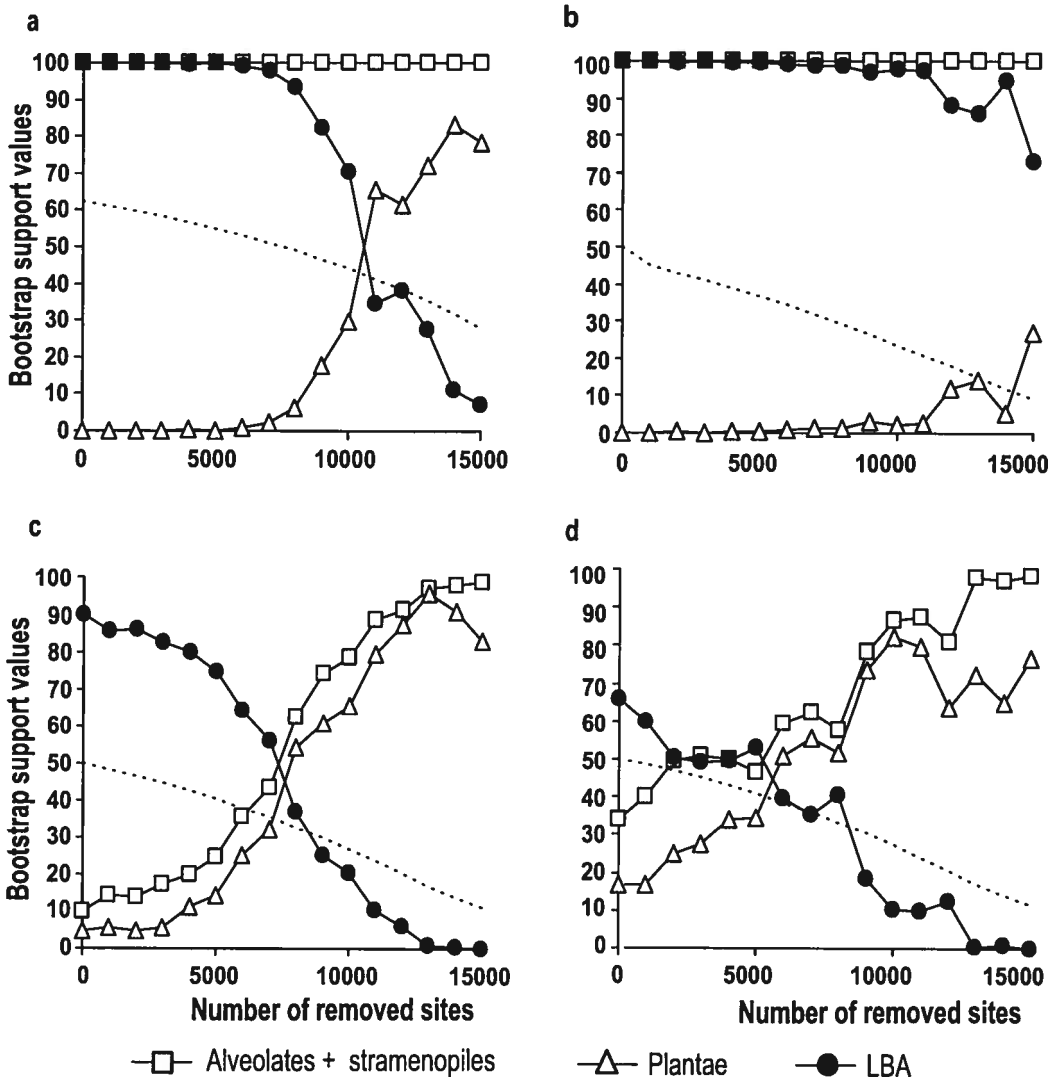

FIGURE 6. Progressive removal of fast-evolving sites from the datasets of Figs. 3b (graph a), 4a (graph b), 4b (graph c) and 4c (graph d). The site-specific rates were calculated with the best ML topology from which *Cyanidioschyzon* was excluded. Dotted line represents the number of parsimony informative positions.

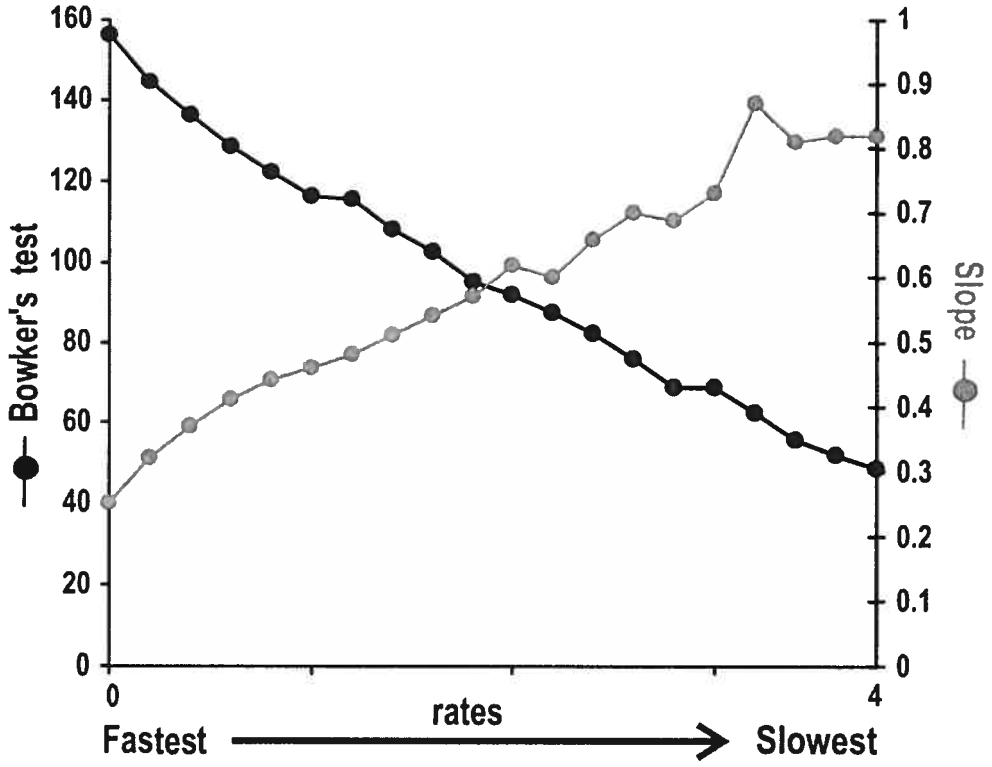

FIGURE 7. Amino acid compositional bias and level of saturation observed in blocks of 1,000 positions when progressively removing sites from fast to slow. For each block, the Bowker's test for amino acid composition (black) and the correlation between the observed differences and estimated substitutions (grey) were performed.

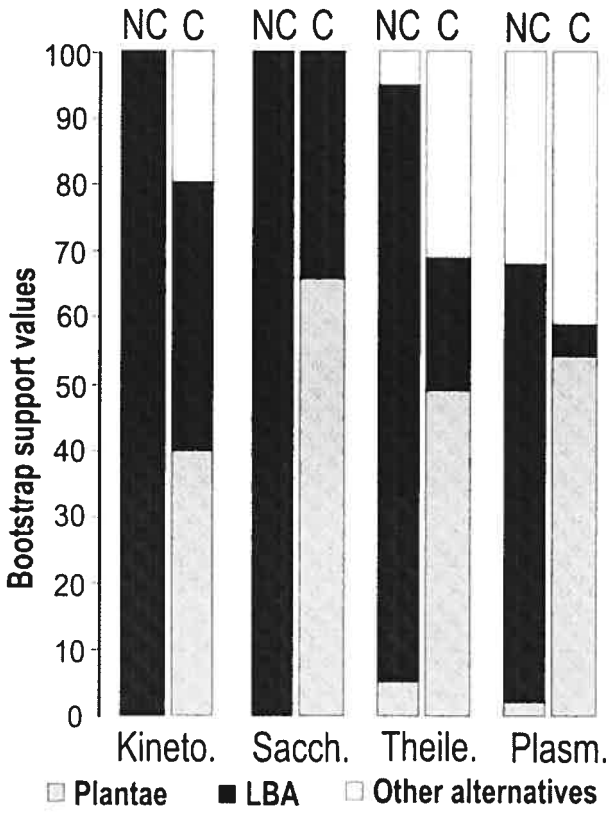

FIGURE 8. Differences in bootstrap support without (WAG+F+Γ4 model) and with amino acid recoding (GTR+F+Γ4 model). Four datasets including *Cyanidioschyzon* as the only red alga were analyzed before (NC) and after the coding (C). Support for the monophyly of Plantae, grey; misplacement of the red algae as shown in Figs. 3b (Kineto.), 4a (Sacch.), 4b (Theile.) and 4c (Plasm.), black.

Supplementary Figures

Detecting and overcoming systematic errors in genome-scale phylogenies

Rodríguez-Ezpeleta et al.

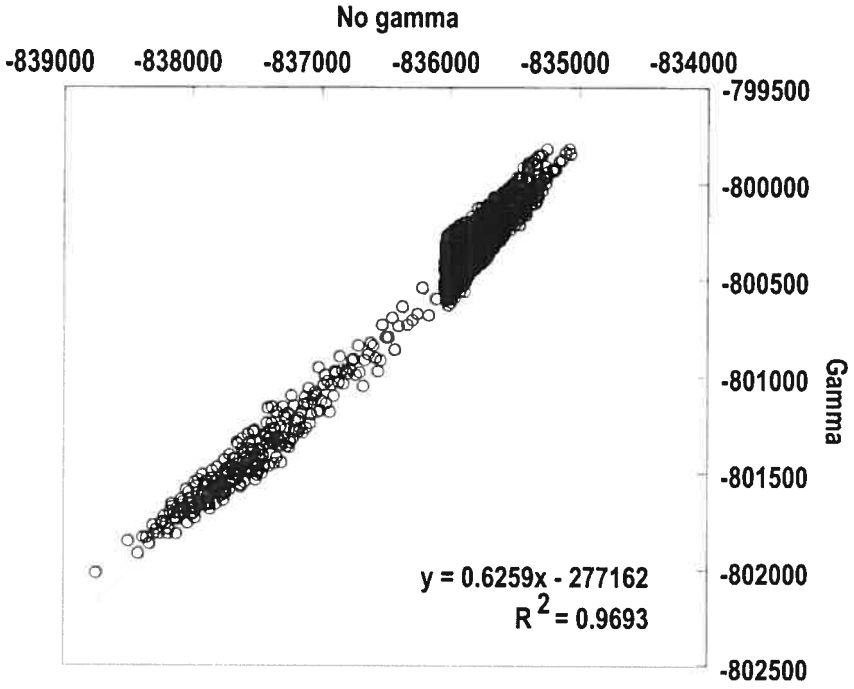

FIGURE S1: Correlation between the log likelihood values for the same topologies analyzed with (Y axis) and without (X axis) gamma distributed rates across sites. The trees correspond to the 1,733 best ranking topologies and 267 others sampled with a step of 500 from the analysis without gamma distributed rates across sites.

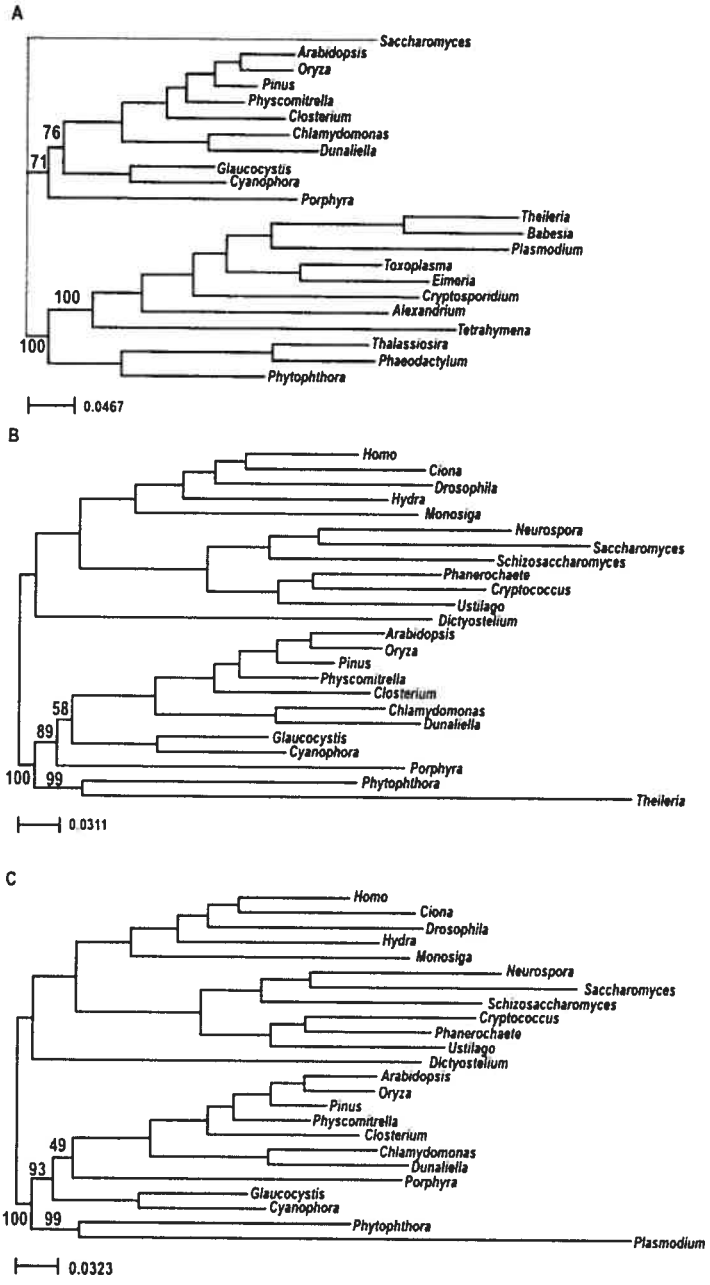

FIGURE S2: Same analyses as in Fig. 1, but using different combinations of taxon sampling that are likely to induce an LBA artefact. In all cases, only *Porphyra* was used to represent the red algae and the kinetoplastids were excluded from the dataset; (A) using *Saccharomyces* as the only representative of the opisthokonts and Amoebozoa; (B) using *Theileria* and *Phytophthora* as the only representatives of alveolates and stramenopiles respectively; and (C) using *Plasmodium* and *Phytophthora* as the only representatives of alveolates and stramenopiles respectively.

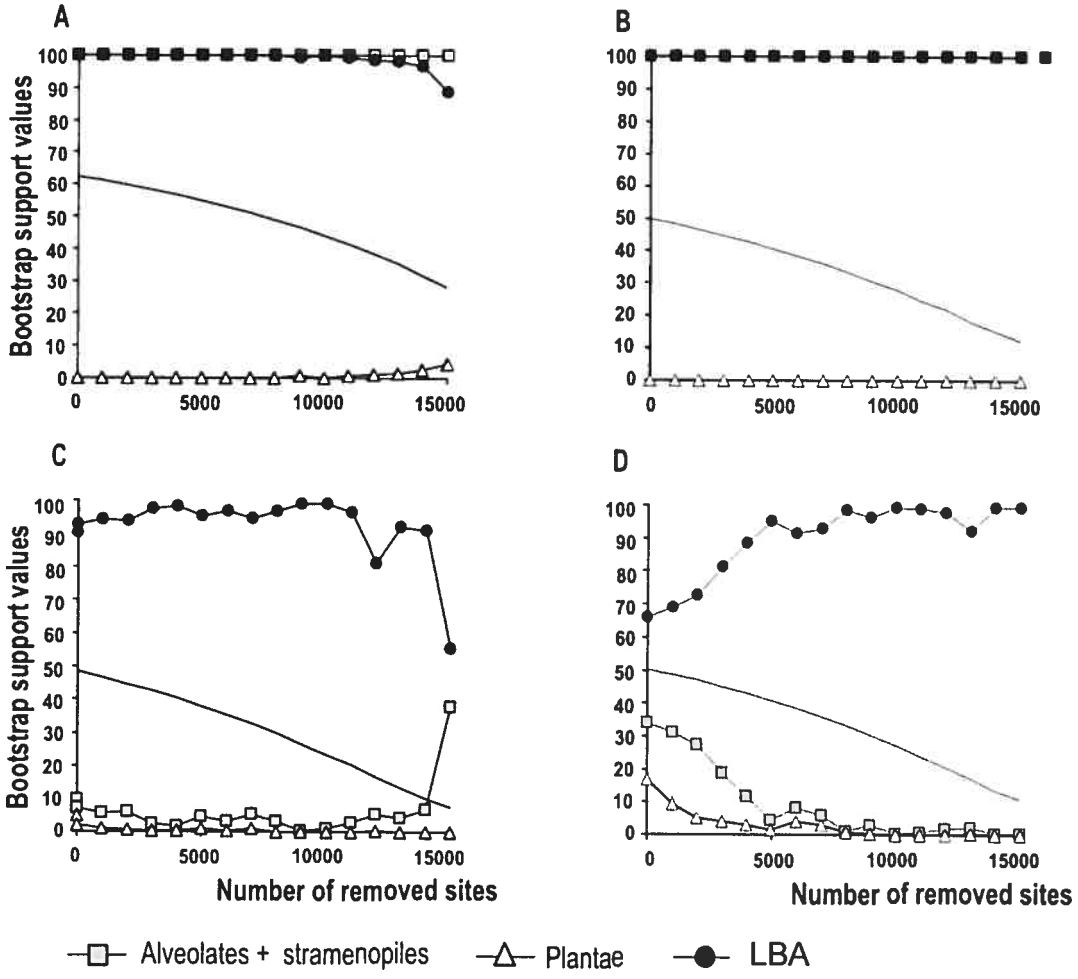

FIGURE S3: Progressive removal of the fastest evolving sites from the datasets of Figs. 3B, 4A, 4B and 4C. The rates of each site were calculated on the incorrect topology in which *Cyanidioschyzon* is attracted to a long branch. Y and X axes refer to bootstrap values (in %) and number of amino acid positions removed, respectively. Plain line represents the number of parsimony informative positions.

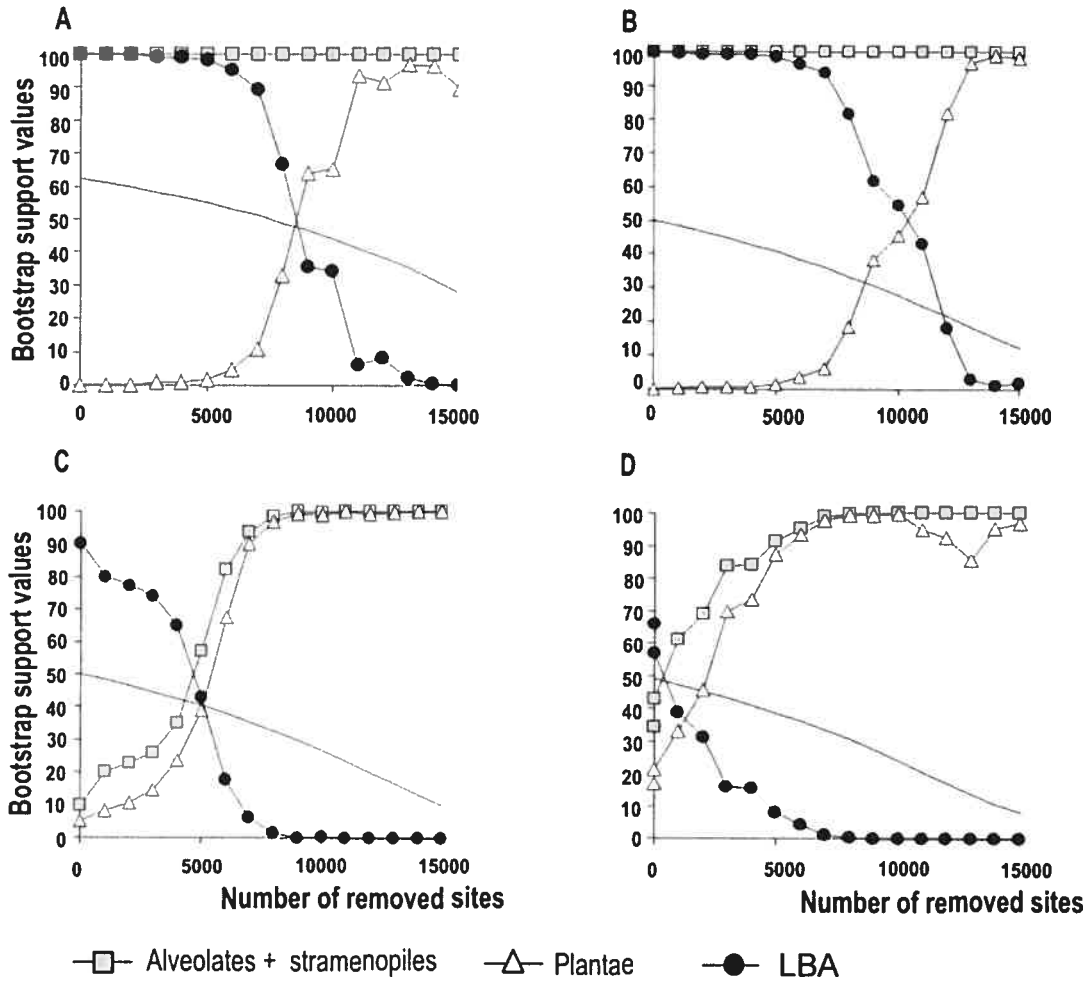

FIGURE S4: Progressive removal of the fastest evolving sites from the datasets of Figs. 3B, 4A, 4B and 4C. The rates of each site were calculated on the correct topology in which *Cyanidioschyzon* branches with the Plantae. Y and X axes refer to bootstrap values (in %) and number of amino acid positions removed, respectively. Plain line represents the number of parsimony informative positions.

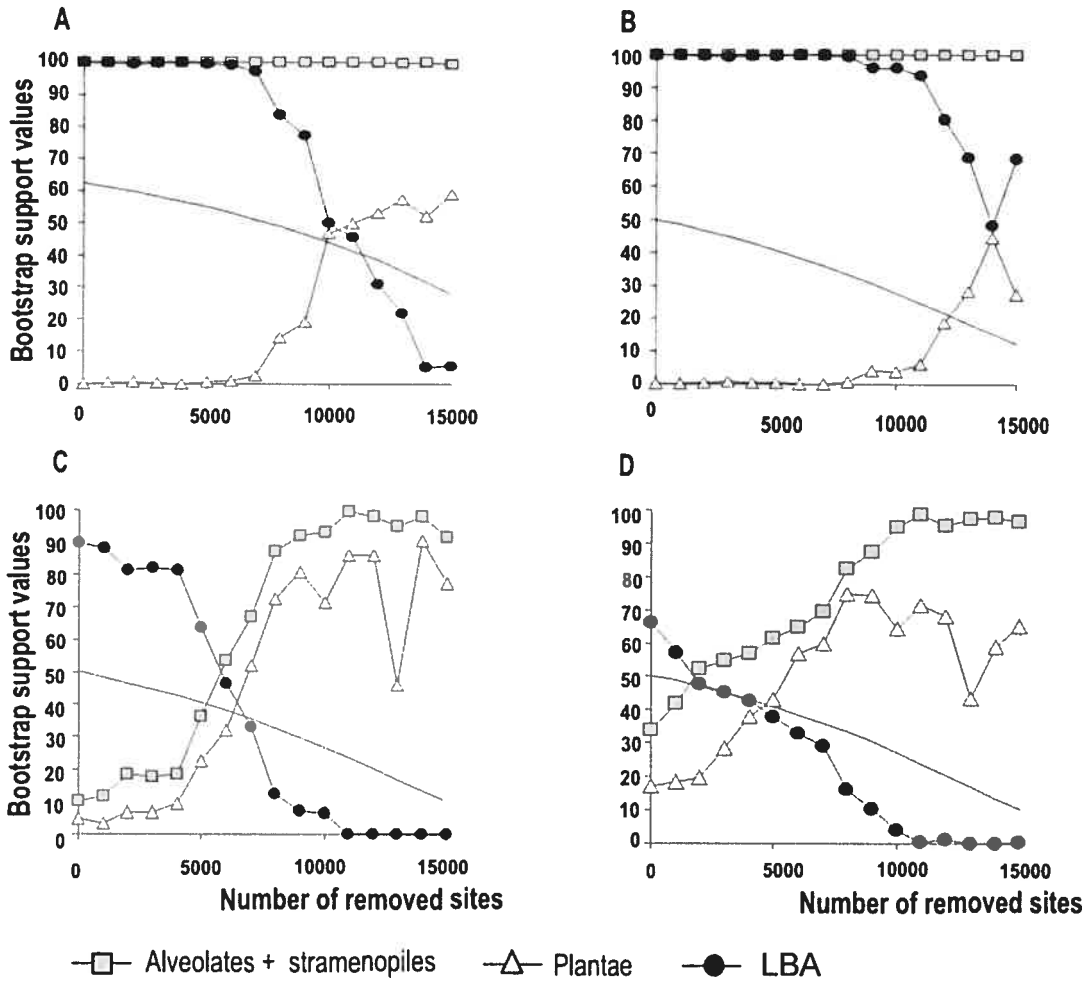

FIGURE S5: Progressive removal of the fastest evolving sites from the datasets of Figs. 3B, 4A, 4B and 4C. The rates of each site were calculated on all the 2,000 retained topologies and the mean of the rates for each site was computed. Y and X axes refer to bootstrap values (in %) and number of amino acid positions removed, respectively. Plain line represents the number of parsimony informative positions.

CHAPITRE IV : LE PLACEMENT DE *MESOSTIGMA* DANS LES STREPTOPHYTES

PUBLIÉ DANS **MOLECULAR BIOLOGY AND EVOLUTION** 2007; 24(3):723-31

PHYLOGENETIC ANALYSES OF NUCLEAR, MITOCHONDRIAL AND PLASTID DATASTETS SUPPORT THE PLACEMENT OF MESOSTIGMA IN THE STREPTOPHYTA

NAIARA RODRÍGUEZ-EZPELETA[1,*], HERVÉ PHILIPPE[1,*] HENNER BRINKMANN[1], BURKHARD BECKER[2] AND MICHAEL MELKONIAN[2]

[1] *Canadian Institute for Advanced Research, Centre Robert Cedergren, Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal, Québec, H3T 1J4, Canada.*

[2] *Botanisches Institut, Lehrstuhl I, Universität zu Köln, Gyrhofstr. 15, Köln, 50931, Germany.*

[*] *Contribution égale*

Phylogenetic Analyses of Nuclear, Mitochondrial, and Plastid Multigene Data Sets Support the Placement of *Mesostigma* in the Streptophyta

Naiara Rodríguez-Ezpeleta,[1] *Hervé Philippe,*[1] *Henner Brinkmann,* Burkhard Becker,† and Michael Melkonian†*

*Département de Biochimie, Centre Robert Cedergren, Canadian Institute of Advanced Research, Université de Montréal, Montréal, Québec, Canada; and †Botanisches Institut, Lehrstuhl I, Universität zu Köln, Köln, Germany

All extant green plants belong to 1 of 2 major lineages, commonly known as the Chlorophyta (most of the green algae) and the Streptophyta (land plants and their closest green algal relatives). The scaly green flagellate *Mesostigma viride* has an important place in the debate on the origin of green plants. However, there have been conflicting results from molecular systematics as to whether *Mesostigma* diverges before the Chlorophyta/Streptophyta split or is an early diverging flagellate member of the Streptophyta. Previous studies employed either a limited taxon sampling (plastid and mitochondrial genomes) or a small number of phylogenetically informative sites (single nuclear genes). Here, we use large data sets from the nuclear (125 proteins; 29,319 positions), mitochondrial (33 proteins; 6,622 positions), and plastid (50 proteins; 10,137 positions) genomes with an expanded taxon sampling (21, 13, and 28 species, respectively) to reevaluate the phylogenetic position of *Mesostigma*. Our study supports the placement of *Mesostigma* in the Streptophyta (as an early diverging lineage) and provides evidence that systematic biases have played a role in generating some of the previous conflicting results. Importantly, we demonstrate that using an increased taxon sampling as well as more realistic models of evolution allows increasing congruence among the nuclear, mitochondrial, and plastid data sets.

Introduction

Since its discovery in 1894 (Lauterborn 1894), the ubiquitous, but nonabundant, freshwater green flagellate *Mesostigma* has had an inconspicuous research history until Manton and Ettl (1965) described the ultrastructure of the spectacular scale covering of the cell and its 2 flagella, which allied the organism with the then newly recognized green algal class Prasinophyceae (Christensen 1962). Ultrastructural studies in the early 1970s renewed interest in the Prasinophyceae, when morphologically similar scales were found on flagellate reproductive cells of algae such as the stonewort *Chara corallina* (Moestrup 1970), that had been implicated in the evolution of embryophyte land plants based on ultrastructural features of mitosis/cytokinesis and the flagellar apparatus (Pickett-Heaps and Marchant 1972). The search for a flagellate member of the land plant lineage among the prasinophyte flagellates remained unsuccessful until a cruciate flagellar root system with 2 multilayered structures (MLSs) was eventually described in *Mesostigma viride* (Rogers et al. 1981; Melkonian 1983, 1989), which linked this organism to both major lineages of green plants, the Chlorophyta (with cruciate flagellar root systems) and the Streptophyta (with unilateral flagellar root systems and a MLS), the latter including the embryophyte land plants (Bremer et al. 1987). Because many structural features of *Mesostigma*, however, resembled those of other prasinophyte flagellates such as *Pyramimonas*, *Mesostigma* was often regarded as closely related to such prasinophytes (Moestrup and Throndsen 1988; Melkonian 1990; Moestrup 1991), and it was not until molecular phylogenetic analyses of nuclear-encoded genes (small subunit [SSU] rDNA and actin) revealed the affiliation of *Mesostigma* with the Streptophyta (Melkonian et al. 1995; Bhattacharya et al. 1998; Marin and Melkonian 1999) that this view changed (but for a traditional taxonomic treatment of *Mesostigma*, see Moestrup 2002). The initial molecular phylogenetic analyses renewed interest in this organism and sparked a flurry of further studies aiming to determine the "true" position of *Mesostigma* in the tree of life. Unfortunately, depending on the data set used, conflicting conclusions were reached (for reviews, see Lewis and McCourt 2004; McCourt et al. 2004): whereas phylogenetic analyses based on chloroplast and mitochondrial genes and genomes led to the conclusion that *Mesostigma* is the earliest green plant divergence branching before the split into the Chlorophyta and Streptophyta (Lemieux et al. 2000; Turmel, Ehara, et al. 2002; Turmel, Otis, and Lemieux 2002b), other studies using a single chloroplast gene (*rbcL*; Delwiche et al. 2002) or a combination of 4 genes from the chloroplast, mitochondrial, and nuclear genomes (Karol et al. 2001) concluded that *Mesostigma* represents the earliest divergence within the Streptophyta. In the meantime, molecular data accumulated that revealed a number of embryophyte traits in *Mesostigma*, thus refuting its close affiliation with other prasinophyte flagellates (Nedelcu et al. 2006; Petersen et al. 2006; Simon et al. 2006). The availability of EST data from *Mesostigma* (Simon et al. 2006) afforded the possibility to perform a large-scale multigene phylogenetic analysis to reevaluate its phylogenetic position. Here, we report multigene phylogenetic analyses of large nuclear, mitochondrial, and chloroplast data sets. The phylogenetic analyses based on the nuclear, mitochondrial, and plastid data sets are consistent with the conclusion that *Mesostigma* is a member of the Streptophyta. We discuss possible reasons for the incongruence between previous studies.

Materials and Methods

Construction of the Nuclear, Plastid, and Mitochondrial Data Sets

The nuclear data set is based on an available alignment (TREEBASE, accession number SN2312) to which EST and trace sequences downloaded from National Center

[1] These authors contributed equally to this work.

Key words: *Mesostigma*, *Chlorokybus*, phylogenomics, taxon sampling, systematic error, heterotachy.

E-mail: [redacted] de.

Mol. Biol. Evol. 24(3):723–731. 2007
doi:10.1093/molbev/msl200
Advance Access publication December 16, 2006

for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) from *Mesostigma viride*, *Selaginella moellendorffii*, *Volvox carteri*, *Scenedesmus obliquus*, *Prototheca wickerhamii*, *Helicosporidium sp.*, *Scherffelia dubia*, *Ostreococcus tauri*, *Galdieria sulphuraria*, and *Chondrus crispus* were added as described (Philippe et al. 2004). Among the Viridiplantae, the 15 species with the largest number of genes sequenced, that represent the major groups were selected. Four red algae and 2 glaucophytes were used to root the tree. Unambiguously aligned sequence blocks were extracted with Gblocks (Castresana 2000) and manually verified.

Potential paralogs were identified and removed as described (Philippe et al. 2004). When including all orthologous proteins that are available from at least 11 out of the 21 used species, the data set contains 125 proteins, totaling 29,319 amino acid positions. On average, 37% of the amino acids are missing.

The mitochondrial data set consists of 33 proteins, a total of 6,622 amino acid positions, from 13 species, including the jakobid *Reclinomonas* and 4 red algae that were used as an outgroup. The plastid data set consists of 50 plastid-encoded proteins, a total of 10,137 amino acid positions, from 28 species, including a glaucophyte and 8 red plastid-containing eukaryotes used as an outgroup. In all, 13% and 5% of the amino acid positions are missing in the mitochondrial and plastid data sets, respectively. Both data sets were aligned with ClustalW (Thompson et al. 1994), refined manually using MUST (Philippe 1993), and filtered from ambiguously aligned positions with Gblocks (Castresana 2000).

Phylogenetic Analyses

The concatenated data sets of nuclear, mitochondrial, and plastid sequences were analyzed by maximum parsimony (MP) and maximum likelihood (ML). MP analyses were performed using PAUP* 4.0 b10 (Swofford 2002) with the Tree Bisection-Reconnection search and 10 random addition of species. ML analyses were performed using PhyML 2.4 (Guindon and Gascuel 2003) and TREEFINDER (Jobb et al. 2004) with an evolutionary model consisting of the WAG matrix of amino acid substitution (Whelan and Goldman 2001), estimated amino acid frequencies, and a gamma distribution split into 4 categories to model the rate heterogeneity among sites (WAG + F + Γ). The reliability of each internal branch was evaluated based on 1,000 (MP) or 100 (ML) bootstrap replicates.

Because the probability of getting trapped on a local maximum using heuristic searches is high when large data sets are used (Salter 2001), we also performed exhaustive ML analyses on a set of topologies obtained by constraining several groups. In general, groups supported by 100% bootstrap values (BVs) in previous analyses were constrained. In some cases, because constraining all the groups supported at 100% BV resulted in a too small set of topologies, we let some of them free (see information about constrained groups in the figures). In all cases, the 3 alternative positions for *Mesostigma* were exhaustively tested. The constraints allowed us to reduce the number of topologies to be analyzed to 81, 405, and 675 for the nuclear, plastid, and mitochondrial data sets, respectively. For each data set, the likelihood of each topology was calculated using Tree-Puzzle (Schmidt et al. 2002) with the WAG + F + Γ model on the concatenated data set. Sitewise likelihood values were calculated by PAML (Yang 1997) and used to perform RELL bootstrap analyses (Kishino et al. 1990; Bapteste et al. 2002) with 10,000 replicates to assess the statistical support for each unconstrained branch.

Amino Acid Coding to Reduce Compositional Bias

To reduce the possible impact of compositional bias, we recoded the amino acids into 4 functional groups. To do that, we used the same coding as Hrdy et al. (2004) but combined the aromatic Phenylalanine, Tyrosine, and Tryptophane (FYW) and the hydrophobic Methionine, Valine, Isoleucine, and Leucine (MVIL) amino acids in a single category and coded the rare cysteine as missing data. This allowed the use of a general time reversible (GTR) matrix with 4 character states implemented in most programs. The parameters of the GTR matrix were estimated by PAUP* using a Neighbor-Joining tree. The 3 sets of topologies (from nuclear, plastid, and mitochondrial data sets) were exhaustively analyzed by Tree-Puzzle with a GTR + F + Γ model. RELL bootstrap analyses (10,000 replicates) were performed as described above.

Removal of the Fast-Evolving Sites to Reduce the Impact of Some Systematic Errors

To reduce the impact of systematic errors, we eliminated the fast-evolving sites (Philippe, Delsuc, et al. 2005). To do that, we calculated the sitewise rates by PAML on an alignment that does not contain *Mesostigma* using the ML topology from which *Mesostigma* was pruned. This strategy was used because the phylogenetic position of *Mesostigma* may influence the estimation of the sitewise evolutionary rates potentially biasing the site removal approach (Rodriguez-Ezpeleta et al. 2007). The sites were then sorted according to their evolutionary rates and progressive removals of the fastest sites (1,000 each time) were performed. RELL bootstrap analyses (1,000 replicates) were performed after each removal and plotted against the alignment size.

Separate Analysis to Reduce the Impact of Heterotachy

To reduce systematic errors due to rate heterogeneity among sites through time (heterotachy), we performed separate analyses on previously defined partitions of the data set. To do that, we proceeded as described above for the exhaustive analysis but allowed branch lengths, alpha parameter of the gamma distribution, and stationary amino acid frequencies to be estimated independently for each partition. In order to evaluate the performance of the separate versus the concatenated model, we used the second-order Akaike Information Criterion (Akaike 1973), AIC_c (Hurvich and Tsai 1989): $AIC_c = 2\log L + 2K + 2K(K+1)/n - K - 1$, K being the number of free parameters of the model and n the number of positions in the data set. The number of free parameters was calculated as 1 alpha parameter + $2s - 3$ branch lengths (s being the number of species) and 19 amino acid frequencies.

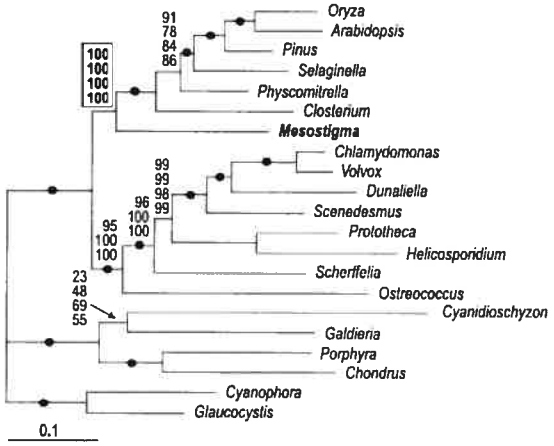

FIG. 1.—Phylogenetic analysis of 125 nuclear-encoded proteins. Optimal ML tree obtained by the analysis of 125 concatenated nuclear-encoded proteins (29,319 amino acid positions). Numbers represent (in order) support values obtained for 1,000 bootstrap replicates in MP analyses, 100 in unconstrained ML (TREEFINDER and PhyML), and 10,000 in the constrained ML analyses. No numbers indicate that the branch was supported by 100% BV with all methods. Black dots indicate the groups constrained in the exhaustive ML analysis. The scale bar denotes the estimated number of amino acid substitutions per site.

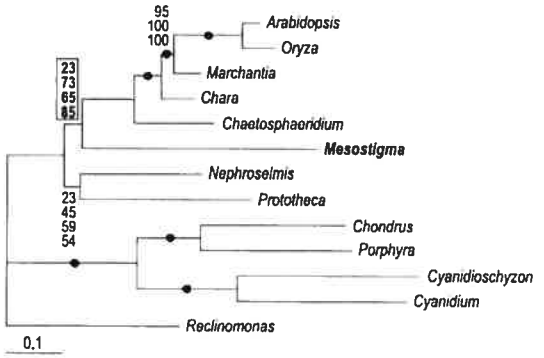

FIG. 2.—Phylogenetic analyses of 33 mitochondrial-encoded proteins. Optimal ML tree obtained by the analysis of 33 concatenated mitochondrial-encoded proteins (6,622 amino acid positions). See figure 1 for details.

Results

Phylogenetic Analysis Based on Nuclear Genes

We assembled a data set of 125 evolutionarily conserved orthologous proteins (29,319 amino acid positions) from 15 different taxa of Viridiplantae and 4 red algae and 2 glaucophytes as outgroups to gain more insight into the phylogenetic position of *M. viride*. The ML phylogeny inferred from the concatenation of the 125 proteins clearly placed *Mesostigma* in a sister-group position to the 6 other streptophyte taxa included in the analysis (fig. 1). This position of *Mesostigma* is supported by 100% BVs in all analyses (fig. 1). Almost all nodes in the tree are well resolved (except for the 2 long-branch thermophilic red algae), and the overall tree topology agrees well with phylogenies derived from SSU rDNA sequence comparisons (e.g., Marin and Melkonian 1999). The amino acid coding into 4 functional groups (to reduce the impact of compositional bias), the removal of fast-evolving sites, or the separate analysis of 2 partitions (ribosomal and nonribosomal proteins) did not change the results (data not shown).

Phylogenetic Analysis Based on Mitochondrial Genes

The data set consists of 33 mitochondrion-encoded orthologous proteins (6,622 amino acid positions) from 8 different taxa of Viridiplantae and 4 red algae and the jakobid flagellate *Reclinomonas* as outgroups. The ML phylogeny inferred from the concatenation of the 33 proteins again placed *Mesostigma* in a sister-group position to the 5 other streptophyte taxa included in this analysis (fig. 2). This position is, however, only weakly supported by BVs (no support in MP, 65–85% BV support in the ML [TREEFINDER and PhyML] and exhaustive analyses) (fig. 2). Again we note that the tree topology is not only largely congruent with that of the nuclear-encoded data set (considering the different taxon sampling) but overall is also not in conflict with several previous single-gene phylogenies. Our phylogeny, however, is in conflict with the phylogenetic analysis based on mitochondrial proteins presented by Turmel, Otis, and Lemieux (2002b), who showed that *Mesostigma* branched at the base of the Viridiplantae. To understand this discrepancy, we did the following additional experiments (summarized in the Supplementary Material online, supplementary figs. S1 and S2): 1) we reduced our data set to the same proteins used by Turmel, Otis, and Lemieux (2002b), that is, 4,842 amino acids; 2) we reduced the taxon sampling to the same species used by these authors (8 species); 3) we reduced both the number of proteins and the taxon sampling; and 4) finally, we analyzed all data sets with (supplementary fig. S1) and without (supplementary fig. S2) taking rate heterogeneity among sites into account. The results of Turmel, Otis, and Lemieux (2002b) could only be reproduced using their taxon sampling and without using a gamma distribution to model the rate heterogeneity among lineages. Applied to our original data set, the amino acid coding into 4 functional groups or the separate analysis of 2 partitions (ribosomal and nonribosomal proteins) does not change the results; however, the removal of fast-evolving sites improves the bootstrap support value for the placement of *Mesostigma* as sister group of streptophytes (97% when the 1,500 fastest evolving sites are eliminated; data not shown).

Phylogenetic Analysis Based on Plastid Genes

The data set consists of 50 plastid-encoded orthologous proteins (10,137 amino acid positions) from 19 taxa of Viridiplantae and 8 eukaryote taxa with red-algal type plastids and the glaucophyte *Cyanophora* as outgroups. The ML tree inferred from the concatenation of the 50 proteins places *Mesostigma* together with the streptophyte genus *Chlorokybus* (fig. 3*A*) with strong support (100% BV in all analyses). *Mesostigma* + *Chlorokybus* (MC) emerge at the base of the Viridiplantae with low bootstrap support in the ML analysis (64–68% BV) and no support in the MP analysis. If *Chlorokybus* was excluded from the analysis,

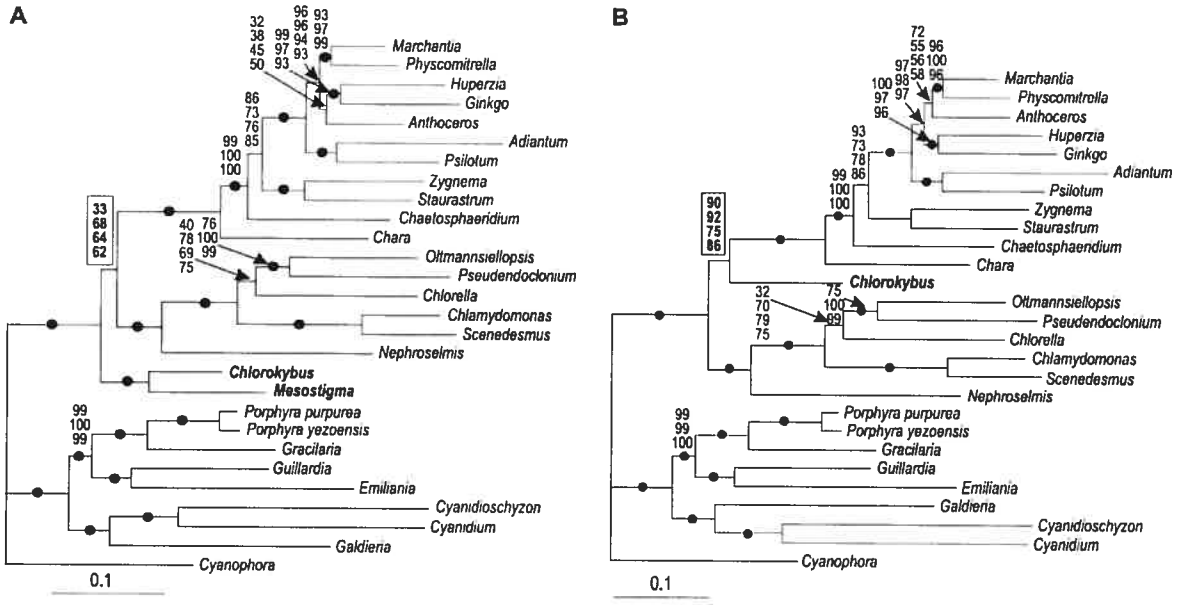

FIG. 3.—Phylogenetic analyses of 50 plastid-encoded proteins. Optimal ML tree obtained by the analysis of 50 concatenated plastid-encoded proteins (10,137 amino acid positions) with the complete data set (*A*) or when *Mesostigma* was excluded (*B*). See figure 1 for details.

the support for the basal position of *Mesostigma* increased significantly in the ML analyses (85–93% BV; summarized in the Supplementary Material online, supplementary fig. S3). If *Mesostigma*, however, was excluded from the analysis (fig. 3*B*), *Chlorokybus* grouped with the other Streptophyta in a basal position with BVs of 75–92%. In these 3 cases, the amino acid coding into 4 functional groups or the removal of fast-evolving sites did not change the results (not shown).

However, these 2 approaches do not overcome tree-reconstruction artifacts due to heterotachy, that is, rate variation across sites through time (Lopez et al. 2002), which are most likely present in plastid data sets (Lockhart et al. 2006). In an attempt to detect heterotachy, we divided the data set into 3 functional classes: translation (ribosomal proteins), RNA polymerase (A, B, and B′ subunits of the RNA polymerase), and photosynthesis (the remaining proteins, all directly implicated in photosynthesis except 2—acetylCoA carboxylase and cytochrome biogenesis protein). As shown in figure 4, the differences on the branch lengths inferred from the 3 data sets are extreme. In particular, streptophytes and chlorophytes evolve about 3 times faster with respect to the remaining species (including *Mesostigma* and *Chlorokybus*) in the RNA polymerase data set and not in the other 2. As expected, important differences in the BV for the position of MC are observed when analyzing these tree data sets independently (table 1). The translation data set (2,199 amino acid positions) supports the placement of MC in the Streptophyta with 96% BV, whereas the support for the same relationship is nonexistent with the RNA polymerase data set (1,498 amino acid positions). Albeit being the largest (6,449 amino acid positions), the photosynthesis data set does not discriminate among the 3 alternatives.

A way to reduce artifacts due to heterotachous behaviors is to use a separate model (Yang 1996; Kolaczkowski and Thornton 2004). However, a fully separate analysis considering independent parameters for each one of the 50 proteins only slightly decreases the support for the probably incorrect basal position of MC. In fact, there are too many free parameters (3,650) in this model, and the AIC (table 1) indicates that it is the one that worst fits the data. In contrast, when only 3 partitions (translation, RNA polymerase, and photosynthesis) are considered, only 146 additional parameters with respect to the concatenate model are needed, and the fit is better in this case (table 1). Interestingly, support for the sister group of MC and streptophytes increases from 32% to 57% with the partially separate model, indicating that heterotachy accounts at least in part for the misplacement of MC in the plastid data set.

Discussion

The incongruence between 2 single-gene phylogenies can be the result of 1) paralogy (generated by gene duplication), lateral gene transfer, or lineage sorting, 2) stochastic error, derived from the use of too few phylogenetically informative sites, and 3) systematic error, arising from the presence of nonphylogenetic signal in the data that is not accounted for in the tree-reconstruction models employed (Phillips et al. 2004). Nonphylogenetic signals mainly derive from variable evolutionary rates across lineages leading to the well-known long-branch attraction (LBA) artifact, heterogeneous nucleotide/amino acid compositions leading to artificial attraction of taxa with the same bias, and heterogeneity of the evolutionary rate of a given position through time, that is, heterotachy (Philippe, Delsuc, et al. 2005).

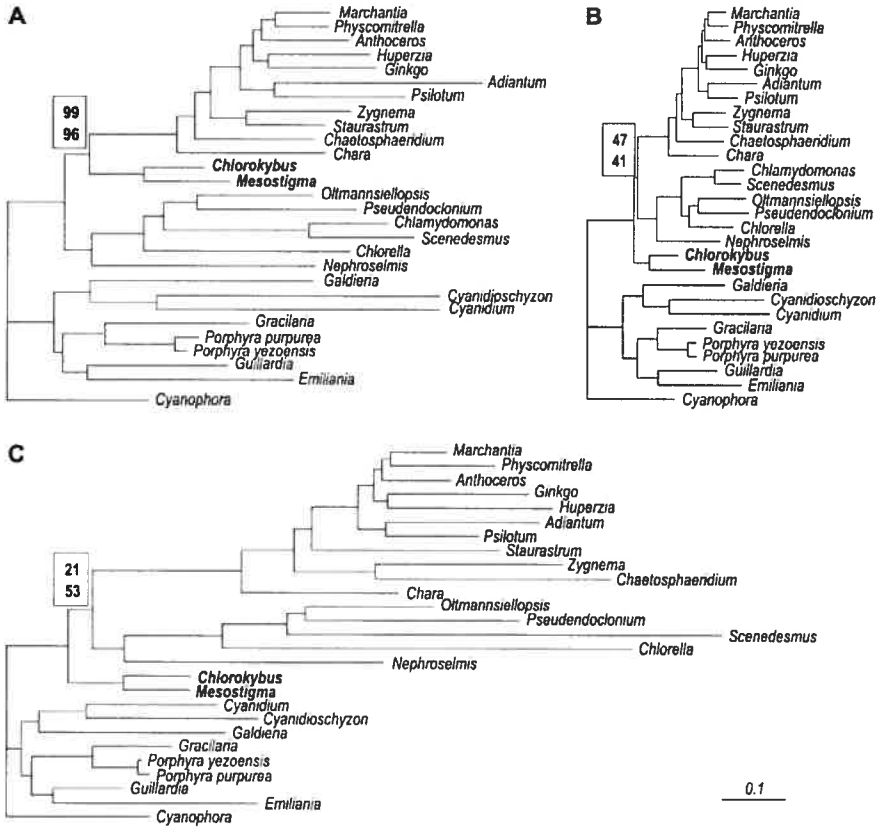

FIG. 4.—Phylogenetic trees obtained from 3 subsets of the plastid data set. Optimal ML trees obtained by the analysis of the translation (*A*), photosynthesis (*B*), and RNA polymerase data sets (*C*). Numbers represent support values for the position of *Mesostigma* and *Chlorokybus* obtained for 100 bootstrap replicates in unconstrained (TREEFINDER) (above) and 10,000 in the constrained ML analyses (below). The scale bar denotes the estimated number of amino acid substitutions per site.

The extent of taxon sampling may also account for incongruence between different phylogenies. Empirical evidence (e.g., Brinkmann et al. 2005; Philippe, Lartillot, et al. 2005) argues for a rich taxon sampling in phylogenetic analyses because this enables better detection of multiple substitutions and thus recognition of LBA artifacts (Felsenstein 1978; Hendy and Penny 1989). Sometimes, however, the number of extant lineages is extremely sparse and it may never be possible to attain a rich or balanced taxon sampling.

Previous single-gene/few-gene phylogenetic analyses trying to assess the phylogenetic position of *Mesostigma* may have suffered from some or all the above problems and yielded conflicting results. Given these restrictions, the conclusions from these analyses are often contradictory and somewhat limited. Phylogenies based on the nuclear-encoded SSU rDNA (Melkonian et al. 1995; Marin and Melkonian 1999) and actin-coding genes (Bhattacharya et al. 1998) suggested the divergence of *Mesostigma* at the base of the Streptophyta, although the basal divergences within the Streptophyta remained unresolved. Phylogenies derived from plastid-encoded genes are incongruent. The concatenation of the 2 plastid-encoded rRNA genes reported phylogenies that consistently placed *Mesostigma* at the base of the Chlorophyta and Streptophyta (Turmel, Ehara, et al. 2002), whereas phylogenies based on the *rbcL* gene or on 4 genes from the 3 genomes (*atpB* and *rbcL* from the plastid genome, *nad5* from the mitochondrial genome,

Table 1
Bootstrap Support Values Obtained with Different Data Sets for the 3 Alternative Positions of *Mesostigma* and *Chlorokybus*

	Number of positions	With Streptophyta	Basal	With Chlorophyta	AIC$_c$[a]
Complete CON	10,137	32	62	6	380,298.65
Complete SEP50	10,137	33	50	17	380,878.36
Complete SEP3	10,137	57	33	10	376,995.03
Translation CON	2,199	97	3	0	
Photosynthesis CON	6,449	22	41	37	
RNApol CON	1,489	0	59	41	

NOTE.—Bootstrap support values are based on 10,000 RELL replicates on the exhaustive analysis (see Materials and Methods). CON (concatenated), same branch lengths, stationary amino acid frequencies, and alpha parameter were used for the whole data set; SEP (separate), each protein (SEP50) or each partition (translation, RNA polymerase, and photosynthesis; SEP3) was allowed to have its own branch lengths, stationary amino acid frequencies, and alpha parameter.

[a] $AIC_c = -2\log L + 2K + 2K(K+1)/n - K - 1$, where K is the number of free parameters and n the number of positions.

and SSU rDNA from the nuclear genome) placed *Mesostigma* at the base of the Streptophyta (Karol et al. 2001; Delwiche et al. 2002).

Phylogenomics, the genome-scale approach to phylogenetic inference, is thought to overcome the limitations of single-gene phylogenies by combining many genes and ultimately complete genomes (Philippe, Delsuc, et al. 2005). The use of large data sets theoretically overcomes incongruence because such data sets reduce the impact of stochastic error when more and more genes are considered. Several empirical studies have confirmed these predictions (e.g., Qiu et al. 1999; Madsen et al. 2001; Bapteste et al. 2002); however, conflicting results have also emerged (such as the question of the monophyly of the Ecdysozoa, Lophotrochozoa, and Protostomia; see Philippe, Lartillot, et al. 2005). Systematic error, caused by the presence of nonphylogenetic signals in the data, is not expected to disappear with the addition of data because, unlike stochastic error, it does not average out over sites. If nonphylogenetic signal is strong enough, it will cause the tree-reconstruction method to be inconsistent and lead to an incorrect, but statistically strongly supported tree (Felsenstein 1978; Phillips et al. 2004).

Two multigene analyses have been carried out to date that specifically address the phylogenetic position of *Mesostigma* in the Viridiplantae, and both refer to organelle phylogenomics. Lemieux et al. (2000) sequenced the entire chloroplast DNA of *Mesostigma* (118,360 bp) and analyzed a subset (53) of the 135 proteins encoded on the plastome (10,629 amino acid positions) with a taxon sampling that contained 3 Embryophyta and 3 Chlorophyta (and *Cyanophora paradoxa* as the outgroup). The tree topology in which *Mesostigma* diverged before the Streptophyta and Chlorophyta was strongly favored over alternative topologies that placed *Mesostigma* at the base of either the Streptophyta or the Chlorophyta. It must be noted, however, that their taxon sampling was very limited and lacked, for example, other streptophyte algae, and only one outgroup taxon was used. Furthermore, the probabilistic methods used (ML analysis under the JTT-F model) for tree reconstruction assumed a uniform rate of substitution. Additional studies from this group increased taxon sampling by adding complete plastid genomes of both streptophyte algae and Chlorophyta (Turmel, Otis, and Lemieux 2002a; Pombert et al. 2005, 2006; Turmel et al. 2005, 2006) but did not address the phylogenetic position of *Mesostigma*. Other multigene analyses using chloroplast proteins gave mixed results concerning the placement of *Mesostigma* (Martin et al. 2002, 2005). In a second approach, Turmel, Otis, and Lemieux (2002b) sequenced the mitochondrial genome of *M. viride* (42,424 bp) and analyzed a subset (19) of the 65 proteins encoded on the chondriome (4,139 amino acid positions) with a taxon sampling that contained 2 embryophytes and 2 Chlorophyta (plus 3 red algae as outgroups). The tree topology in which *Mesostigma* diverged before the Streptophyta and Chlorophyta was supported by BV of 100% in PROTML, distance, and MP analyses assuming a uniform substitution rate across sites. However, when rate variation across sites (8 gamma categories) was taken into consideration, in ML analyses (JTT model), the BV dropped significantly and support for this topology was low (63% or 67%, excluding or including invariant sites, respectively).

How do the results obtained in the present study compare with those previously published? We have assembled a large nuclear data set (125 proteins, 29,319 positions) with a reasonable taxon sampling (15 Viridiplantae, among them 6 streptophytes and 8 chlorophytes + *Mesostigma*). Phylogenetic analyses, involving different methods of tree reconstruction and addressing likely causes of systematic error such as compositional bias and fast-evolving sites, lead us to conclude that *Mesostigma* is an early branching member of the Streptophyta. Because only one other streptophyte alga (*Closterium*) was included in the analysis, we cannot yet address the relationship between *Mesostigma* and other streptophytes such as *Chlorokybus*, *Klebsormidium*, or *Chaetosphaeridium*, which must await the generation of EST data from these organisms. Similarly, we note that early branching chlorophytes such as *Pyramimonas* (Nakayama et al. 1998) are still lacking. In general, the results obtained from multigene analyses of the nuclear data set corroborate earlier analyses of nuclear-encoded single genes (SSU rDNA, actin) that placed *Mesostigma* in the Streptophyta.

For the mitochondrial data set, our results are in accordance with the nuclear data set but are in conflict with the mitochondrial protein phylogeny of Turmel, Otis, and Lemieux (2002b). Additional analyses adjusting the data set (from 6,622 to 4,842 amino acid positions) and taxon sampling (from 13 to 8 taxa) to those used by Turmel, Otis, and Lemieux (2002b) revealed the likely reasons for the discrepancy: poor taxon sampling in the ingroup (lack of other streptophyte algae) as well as in the outgroup (only the long-branch red algae were chosen) appeared to be responsible and the number of positions used was less important as long as the rate heterogeneity among lineages was modeled (see Results). However, when no gamma distribution was used, bootstrap support for the placement of *Mesostigma* with the streptophytes was abolished. This is in accordance with the data of Turmel, Otis, and Lemieux (2002b) who showed that the bootstrap support for the position of *Mesostigma* at the base of the Streptophyta and Chlorophyta was lowered when the rate heterogeneity among lineages was modeled. We conclude that the phylogeny of mitochondrial proteins places *Mesostigma* in the Streptophyta when taxon sampling is improved and the rate heterogeneity among lineages is modeled.

The phylogeny derived from the plastid data set reveals that the inclusion of *Chlorokybus*, which strongly groups with *Mesostigma*, decreases the bootstrap support for the basal position of *Mesostigma* in the Viridiplantae (fig. 3) and that different subsets of the complete data set provide conflicting results (fig. 4 and table 1). When the translational proteins are used, significant support for the placement of *Mesostigma* and *Chlorokybus* (MC) in the Streptophyta is obtained (this is also true for the data sets with *Mesostigma* or *Chlorokybus* only), whereas the RNA polymerase data set significantly rejects this relationship (table 1; supplementary tables S1 and S2, Supplementary Material online). This can be explained by the disproportionately fast evolutionary rate of the streptophyte RNA polymerase, which attracts this group to either the

Chlorophyta or the outgroup (fig. 4*C*). Albeit its large size (6,449 positions) and no apparent evolutionary rate heterogeneities (fig. 4*B*), the photosynthesis data set does not support or reject any of the 3 alternatives (table 1; supplementary tables S1 and S2, Supplementary Material online). This lack of resolution is likely due to the slow evolutionary rate of these proteins.

The plastid data set illustrates tree-reconstruction artifacts due to heterotachy (Kolaczkowski and Thornton 2004). There exist highly heterotachous branch lengths for the different functional classes that cannot be acknowledged when a single set of branch lengths is used to analyze the concatenation of the 50 plastid proteins. As a result, an erroneous topology (MC at the base of green plants) is recovered. Although a separate model a priori defined by 3 function-based gene partitions may correct for heterotachy, the improvement with respect to concatenation is only marginal (from 32% to 57% BV for the sister group of MC and streptophytes). This suggests that even if the branch length differences are extreme between the 3 partitions, among-gene heterotachy is not the only cause of systematic error in this data set (Philippe, Zhou, et al. 2005). For instance, changes in the proportion of variable sites across the phylogeny within genes, which is the case for the RNA polymerase subunits (Lockhart et al. 2006), are not corrected by the type of separate model we used.

The separate model of the 50 proteins has a lower fit to the data because it implies much more free parameters, but only slightly improving the likelihood—note that previous studies (Bapteste et al. 2002; Philippe et al. 2004; Rodriguez-Ezpeleta et al. 2005) suggested, using an inadequate AIC approximation (see Posada and Buckley 2004), that the separate model was better. In any case, the difficultly of correctly locating *Mesostigma* with plastid proteins constitutes an interesting case study for testing new, more realistic, models of sequence evolution.

Apparently, *Mesostigma* attracts *Chlorokybus* to a position in the tree that does not conform to its typical streptophyte traits such as a subapical flagellar insertion and unilateral flagellar root system in the zoospores, which are morphological synapomorphies of the Streptophyta (Rogers et al. 1980; Lewis and McCourt 2004). Given that the evolutionary rate of *Mesostigma* is higher than that of *Chlorokybus*, this suggests that the position of *Mesostigma* in phylogenetic trees based on plastid data sets is affected by systematic errors. According to Jeffroy et al. (2006), phylogenies based on multiple genes can be biased by systematic errors and should be carefully scrutinized for possible tree-reconstruction errors. Supporting their conclusions, here, we have shown the importance of the use of 1) probabilistic methods that aim to capture real substitution patterns (Steel 2005)—in the mitochondrial data set and 2) an increased taxon sampling to corroborate the results—in the plastid and mitochondrial data sets.

In conclusion, the inclusion of *Mesostigma* in the Streptophyta, likely as an early branching lineage, is weakly supported by the mitochondrial and plastid data sets but significantly by the nuclear data set. This corroborates exciting recent findings about land plant–specific molecular/biochemical traits in *Mesostigma* such as the *GapA/B* gene duplication (Petersen et al. 2006; Simon et al. 2006), plant-type peroxisomal glycolate oxidase (Stabenau and Winkler 2005; Simon et al. 2006), the bud-induced (*BIP*) multigene family (Nedelcu et al. 2006), and F-box family proteins (Simon et al. 2006), suggesting that many typical embryophyte traits may have evolved at the level of the unicellular ancestor of the streptophytes before the transition to land, presumably when such a flagellate adapted to a freshwater/brackish habitat (Simon et al. 2006). This underpins the pivotal role that *Mesostigma* is likely to play in the coming years as a model to unravel the intricacies of the early steps in the evolution of streptophytes.

Supplementary Material

Figures S1–S3, tables S1 and S2, and the alignments used are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

H.P. acknowledges Génome Québec, the Canadian Research Chair and the Université de Montréal for financial support and the Réseau Québecois de Calcul de Haute Performance for computational resources. N.R.E. has been supported by 'Programa de Formación de Investigadores del Departamento de Educación, Universidades e Investigación' (Government of Basque Country). Part of this work was supported by grants from the Deutsche Forschungsgemeinschaft (Be 1779/7-1 and Be 1779/7-2).

Literature Cited

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Csaki, editor. Proceedings 2nd International Symposium on Information Theory. Budapest (Hungary): Akademia Kiado. p. 267–281.

Bapteste E, Brinkmann H, Lee JA, et al. (11 co-authors). 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc Natl Acad Sci USA. 99:1414–1419.

Bhattacharya D, Weber K, An SS, Berning-Koch W. 1998. Actin phylogeny identifies *Mesostigma viride* as a flagellate ancestor of the land plants. J Mol Evol. 47:544–550.

Bremer K, Humphries CJ, Mishler BD, Churchill SP. 1987. On cladistic relationships in green plants. Taxon. 36:339–349.

Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol. 54:743–757.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Christensen. T. 1962. Alger. In: Böcher TW, Lange M, and Sørensen T, editors. Botanik, Systematisk Botanik vol. 2, Copenhagen: Munksgaard p. 1–178.

Delwiche CF, Karol KG, Cimino MT, Sytsma KJ. 2002. Phylogeny of the genus *Coleochaete* (Coleochaetales, Charophyta) and related taxa inferred by analysis of the chloroplast gene *rbc*L. J Phycol. 38:394–403.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool. 27: 401–410.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. Syst Zool. 38:297–309.
Hrdy I, Hirt RP, Dolezal P, Bardonova L, Foster PG, Tachezy J, Embley TM. 2004. *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature. 432:618–622.
Hurvich CM, Tsai C-L. 1989. Regression and time series model selection in small samples. Biometrika. 76:297–307.
Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.
Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol. 4:18.
Karol KG, McCourt RM, Cimino MT, Delwiche CF. 2001. The closest living relatives of land plants. Science. 294: 2351–2353.
Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. J Mol Evol. 31:151–160.
Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 431:980–984.
Lauterborn R. 1894. Über die Winterfauna einiger Gewässer der Oberrheinebene. Biol Zbl. 14:390–398.
Lemieux C, Otis C, Turmel M. 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature. 403:649–652.
Lewis LA, McCourt RM. 2004. Green algae and the origin of land plants. Am J Bot. 91:1535–1556.
Lockhart P, Novis P, Milligan BG, Riden M, Rambaut A, Larkum T. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol. 23:40–45.
Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol. 19:1–7.
Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, Amrine HM, Stanhope MJ, de Jong WW, Springer MS. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature. 409:610–614.
Manton I, Ettl H. 1965. Observations on the fine structure of *Mesostigma viride* Lauterborn. J Linn Soc Lond Bot. 59:175–184.
Marin B, Melkonian M. 1999. Mesostigmatophyceae, a new class of streptophyte green algae revealed by SSU rRNA sequence comparisons. Protist. 150:399–417.
Martin W, Deusch O, Stawski N, Grünheit N, Goremykin V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10:203–209.
Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA. 99:12246–12251.
McCourt RM, Delwiche CF, Karol KG. 2004. Charophyte algae and land plant origins. Trends Ecol Evol. 19:661–666.
Melkonian M. 1983. *Mesostigma*, a key organism in the evolution of two major classes of green algae and related to the ancestry of land plants. Br Phycol J. 18:206.
Melkonian M. 1989. Flagellar apparatus ultrastructure in *Mesostigma viride* (Prasinophyceae). Plant Syst Evol. 164:93–122.
Melkonian M. 1990. Phylum Chlorophyta: class Prasinophyceae. In: Margulis L, Corliss JO, Melkonian M, Chapman DJ, editors. Handbook of protoctista. Boston (MA): Jones and Bartlett Publishers. p. 600–607.
Melkonian M, Marin B, Surek B. 1995. Phylogeny and evolution of the algae. In: Arai K, Kato M, Doi Y, editors. Biodiversity and evolution. Tokyo (Japan): The National Science Mueseum Foundation. p. 153–176.
Moestrup Ø. 1970. The fine structure of mature spermatozoids of *Chara corralina*, with special reference to microtubules and scales. Planta. 93:295–308.
Moestrup Ø. 1991. Further studies of presumably primitive green algae, including the description of Pedinophyceae class. Nov. and *Resultor* gen. nov. J Phycol. 27:119–133.
Moestrup Ø. 2002. Phylum Prasinophyta. In: John DM, Whitton BA, Brook AJ, editors. The freshwater algal flora of the British Isles. Cambridge (UK): Cambridge University Press. p. 281–286.
Moestrup Ø, Throndsen J. 1988. Light and electron microscopical studies on *Pseudoscourfieldia marina*, a primitive scaly green flagellate (Prasinophyceae) with posterior flagella. Can J Bot. 66:1415–1434.
Nakayama T, Marin B, Kranz HD, Surek B, Huss VAR, Inouye I, Melkonian M. 1998. The basal position of scaly green flagellates among the green algae (Chlorophyta) is revealed by analyses of nuclear-encoded SSU rRNA sequences. Protist. 149:367–380.
Nedelcu AM, Borza T, Lee RW. 2006. A land plant-specific multigene family in the unicellular *Mesostigma* argues for its close relationship to Streptophyta. Mol Biol Evol. 23:1011–1015.
Petersen J, Teich R, Becker B, Cerff R, Brinkmann H. 2006. The GapA/B gene duplication marks the origin of Streptophyta (charophytes and land plants). Mol Biol Evol. 23:1109–1118.
Philippe H. 1993. MUST, a computer package of management utilities for sequences and trees. Nucleic Acids Res. 21:5264–5272.
Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. Annu Rev Ecol Evol Syst. 36:541–562.
Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol. 22:1246–1253.
Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol. 21:1740–1752.
Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol. 5:50.
Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol. 21:1455–1458.
Pickett-Heaps JD, Marchant HJ. 1972. The phylogeny of the green algae: a new proposal. Cytobios. 6:255–264.
Pombert J-F, Lemieux C, Turmel M. 2006. The complete chloroplast DNA sequence of the green alga *Oltmannsiellopsis viridis* reveals a distinctive quadripartite architecture in the chloroplast genome of early diverging ulvophytes. BMC Biol. 4:3.
Pombert J-F, Otis C, Lemieux C, Turmel M. 2005. The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. Mol Biol Evol. 22:1903–1918.
Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol. 53:793–808.
Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature. 402:404–407.
Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, Bohnert H, Philippe H, Lang BF. 2005.

Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol. 15:1325–1330.

Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol. In press.

Rogers CE, Domozych DS, Stewart KD, Mattox KR. 1981. The flagellar apparatus of *Mesostigma viride* (Prasinophyceae): multilayered structures in a scaly green flagellate. Plant Syst Evol. 138:247–258.

Rogers CE, Mattox KR, Stewart KD. 1980. The zoospore of *Chlorokybus atmophyticus*, a charophyte with sarcinoid growth habit. Am J Bot. 67:774–783.

Salter LA. 2001. Complexity of the likelihood surface for a large DNA dataset. Syst Biol. 50:970–978.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18:502–504.

Simon A, Glöckner G, Felder M, Melkonian M, Becker B. 2006. EST analysis of the scaly green flagellate *Mesostigma viride* (Streptophyta): implications for the evolution of green plants (Viridiplantae). BMC Plant Biol. 6:2.

Stabenau H, Winkler U. 2005. Glycolate metabolism in green algae. Physiol Plant. 123:235–245.

Steel M. 2005. Should phylogenetic models be trying to 'fit an elephant'? Trends Genet. 21:307–309.

Swofford DL. 2002. PAUP* phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland: Sinauer Associates.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Turmel M, Ehara M, Otis C, Lemieux C. 2002. Phylogenetic relationships among streptophytes as inferred from chloroplast small and large subunit rRNA gene sequences. J Phycol. 38:364–375.

Turmel M, Otis C, Lemieux C. 2002a. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc Natl Acad Sci USA. 99:11275–11280.

Turmel M, Otis C, Lemieux C. 2002b. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. Mol Biol Evol. 19:24–38.

Turmel M, Otis C, Lemieux C. 2005. The complete chloroplast DNA sequences of the charophycean green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. BMC Biol. 3:22.

Turmel M, Otis C, Lemieux C. 2006. The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. Mol Biol Evol. 23: 1324–1338.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.

Yang Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol. 42:587–596.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13: 555–556.

Peter Lockhart, Associate Editor

Accepted December 12, 2006

SUPPLEMENTAL DATA

Supplementary tables

Table S1: Bootstrap support values obtained with different datasets for the three alternative positions of *Mesostigma* when *Chlorokybus* is excluded

	Number of positions	With Streptophyta	Basal	With Chlorophyta	AIC_c
Complete CON	10,137	5	90	5	373,644.57
Complete SEP50	10,137	24	64	12	373,941.05
Complete SEP3	10,137	26	62	12	370,309.09
Translation CON	2,199	82	18	0	
Photosynthesis CON	6,449	14	35	51	
RNA pol CON	1,489	0	60	40	

Bootstrap support values are based on 10,000 RELL replicates on the exhaustive analysis (see Materials and Methods). CON, same branch lengths, stationary amino acid frequencies and alpha parameter were used for the whole dataset; SEP, each protein (SEP50) or each partition, translation, RNA polymerase and photosynthesis, (SEP3) was allowed to have its own branch lengths, stationary amino acid frequencies and alpha parameter. The solution with the highest BV in each case is highlighted in grey.
[1] AIC_c= -2LogL + 2K + 2K(K+1)/n-K-1, where K is the number of free parameters and n, the number of positions.

Table S2: Bootstrap support values obtained with different datasets for the three alternative positions of *Chlorokybus* when *Mesostigma* is excluded

	Number of positons	With Streptophyta	Basal	With Chlorophyta	AIC_c
Complete CON	10,137	86	12	2	369,911.76
Complete SEP50	10,137	72	13	15	370,266.36
Complete SEP3	10,137	88	9	3	366,664.91
Translation CON	2,199	97	3	0	
Photosynthesis CON	6,449	43	31	26	
RNA pol CON	1,489	7	34	59	

Bootstrap support values are based on 10,000 RELL replicates on the exhaustive analysis (see Materials and Methods). CON, same branch lengths, stationary amino acid frequencies and alpha parameter were used for the whole dataset; SEP, each protein (SEP50) or each partition, translation, RNA polymerase and photosynthesis, (SEP3) was allowed to have its own branch lengths, stationary amino acid frequencies and alpha parameter. The solution with the highest BV in each case is highlighted in grey.
[1] AIC_c= -2LogL + 2K + 2K(K+1)/n-K-1, where K is the number of free parameters and n, the number of positions.

Supplementary figures

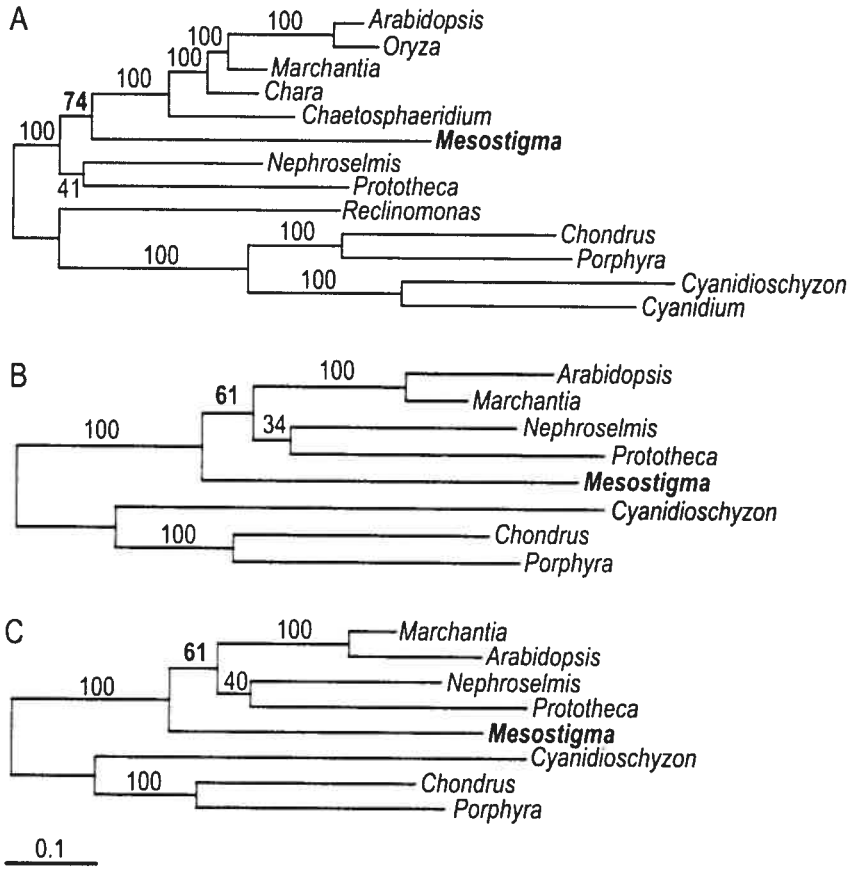

Figure S1: Phylogenetic analyses of different mitochondrial datasets taking the rate among sites variation into account

Maximum Likelihood tree (TreeFinder WAG+F+Γ) obtained by the analysis of (A) 19 concatenated mitochondrial-encoded proteins (the same as in Turmel et al. (2002a)) and 13 species, (B) 33 concatenated mitochondrial-encoded proteins and 8 species (the same as in Turmel et al. (2002a)), and (C) 19 mitochondrial encoded proteins and 8 species (the same proteins and species as in Turmel et al. (2002a)). Numbers represent support values obtained by performing 100 bootstrap replicates. The scale bar denotes the estimated number of amino acid substitutions per site.

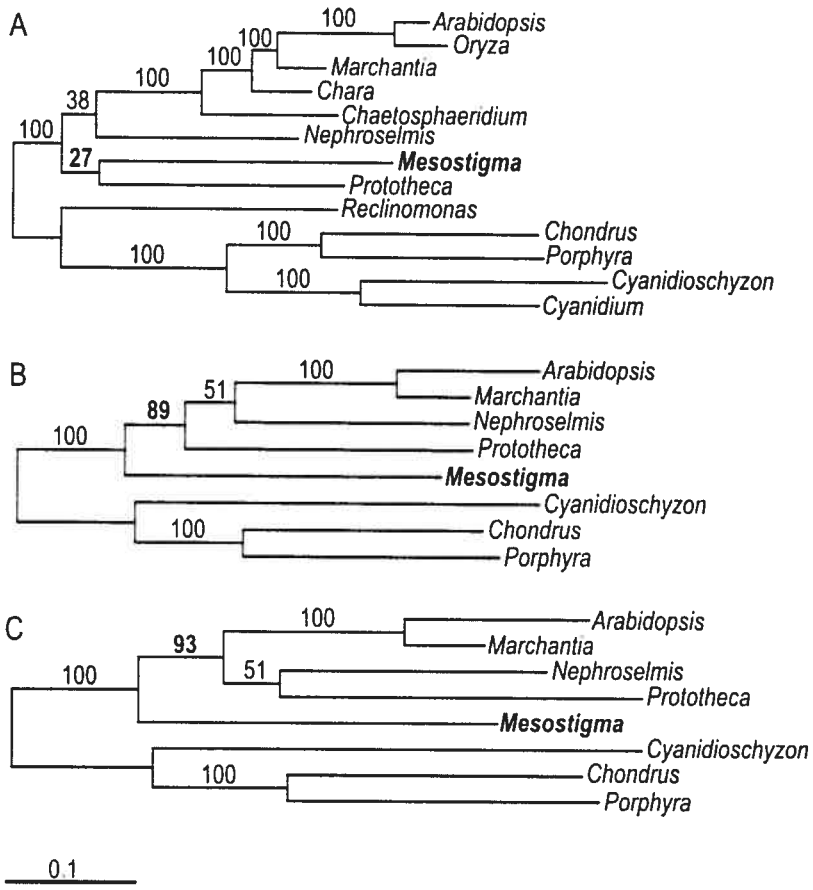

Figure S2: Phylogenetic analyses of different mitochondrial datasets without taking the rate among sites variation into account

The analyses were performed as in Figure S1 on the same datasets with the only difference being that the rate among site variation was not taken into account.

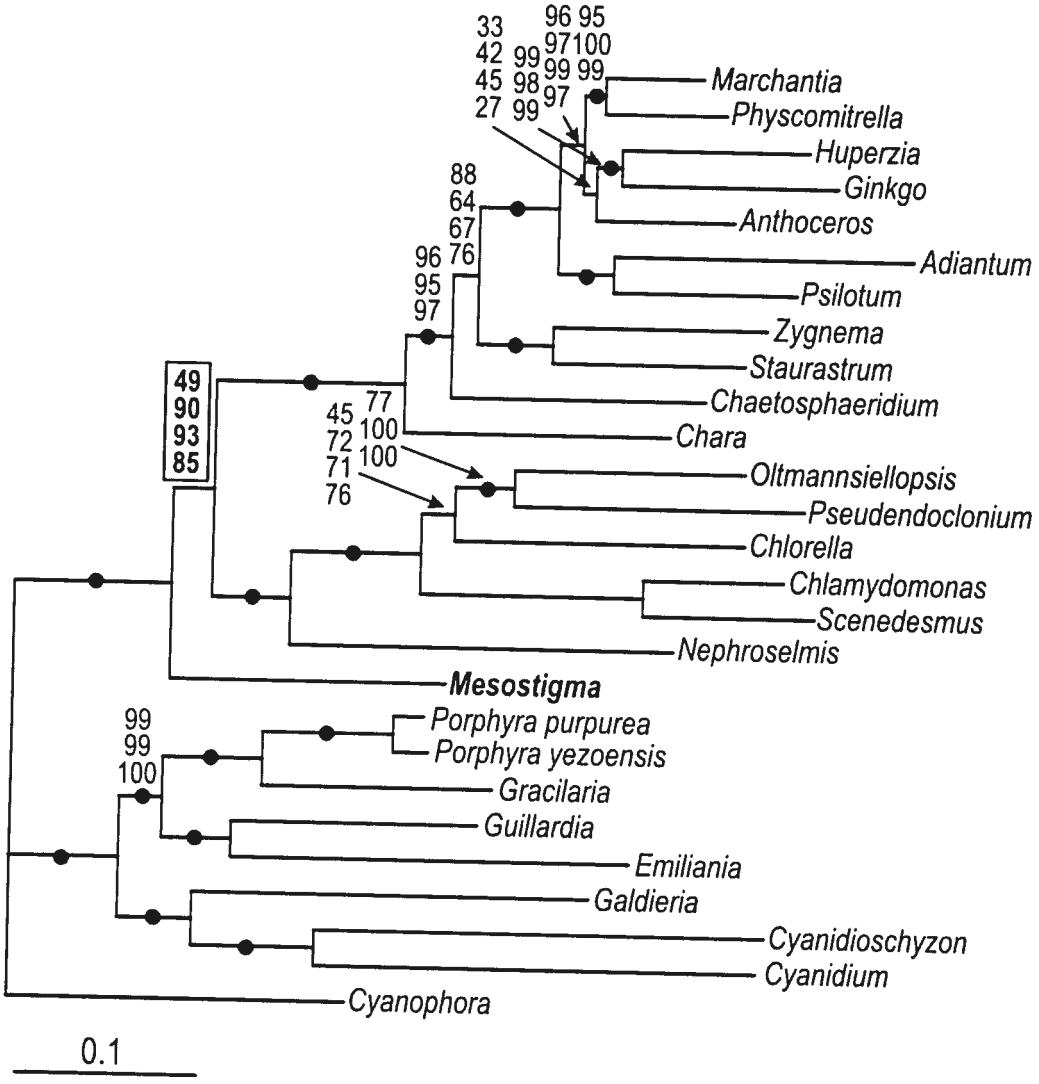

Figure S3: Phylogenetic analyses of 50 plastid encoded genes excluding *Chlorokybus*
Best Maximum Likelihood obtained by the analysis of 50 concatenated plastid-encoded proteins (10,137 amino acid positions). Numbers represent support values obtained by performing 1000 bootstrap replicates on the Maximum Parsimony, and 100 bootstrap replicates on the Maximum Likelihood (TreeFinder and PhyML) and exhaustive analyses. No numbers indicate that the branch was supported by 100% bootstrap value with all methods. Black dots indicate that the corresponding branches were constrained for the exhaustive Maximum Likelihood analysis. The scale bar denotes the estimated number of amino acid substitutions per site.

CHAPITRE V : LA POSITION PHYLOGÉNÉTIQUE DES JAKOBIDES ET MALAWIMONADINES

EN RÉVISION POUR **CURRENT BIOLOGY**

PHYLOGENOMIC EVIDENCE FOR THE SISTER-GROUP RELATIONSHIP OF JAKOBIDS AND EUGLENOZOA

NAIARA RODRÍGUEZ-EZPELETA[1], HENNER BRINKMANN[1], GERTRAUD BURGER[1], MICHAEL W. GRAY[2], HERVÉ PHILIPPE[1] AND B. FRANZ LANG[1]

[1] *Canadian Institute for Advanced Research, Centre Robert Cedergren, Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal, Québec, H3T 1J4, Canada.*

[2] *Department of Biochemistry and Molecular Biology, Dalhousie University, 5850 College Street, Halifax, Nova Scotia, B3H 1X5, Canada.*

Phylogenomic evidence for the sister-group relationship of jakobids and Euglenozoa

Naiara Rodríguez-Ezpeleta[*], Henner Brinkmann[*], Gertraud Burger[*], Michael W. Gray[†], Hervé Philippe[*] and B. Franz Lang[*]

[*]*Centre Robert Cedergren, Département de Biochimie, Université de Montréal, 2900 Boulevard Édouard-Montpetit, Montréal, Québec, H3T 1J4, Canada*

[†]*Department of Biochemistry and Molecular Biology, Dalhousie University, 5850 College Street, Halifax, Nova Scotia, B3H 1X5, Canada.*

Classification: Biological Sciences, Evolution.

Corresponding author: B. Franz Lang[redacted]

Keywords: Jakobids, malawimonads, Euglenozoa, excavates, phylogenomics

SUMMARY

Jakobids are a group of flagellate protists that possess the most complete mitochondrial genome described thus far. They share several ultrastructural features with malawimonads and a few anaerobic protists. These lineages together with Euglenozoa and heterolobosean amoebae have been called 'excavates'. Currently, molecular phylogenies provide no compelling support for the monophyly of this group, and their relationship to other eukaryotic lineages is rather contentious. In order to clarify the phylogenetic position of jakobids and malawimonads, we sequenced a total of ~30,000 ESTs from five and two members of these groups, respectively. Here, we report the first large-scale phylogenies (based on alignments of 170 nucleus-encoded proteins comprising 40,090 amino acid positions) that include jakobids and malawimonads. Our analyses show that jakobids robustly branch with Euglenozoa, a result drawing significant statistical support and independent of taxon sampling. We further observe a sister-group relationship between malawimonads and jakobids+Euglenozoa after removing from the data set fast-evolving species, which are known to destabilize otherwise robust phylogenies. We discuss the implications of our results for the concept of 'Excavata' and for early eukaryotic evolution in general.

INTRODUCTION

Jakobids and malawimonads are free-living, mitochondria-containing flagellates. They are heterotrophic, commonly observed in freshwater and marine habitats where they feed on bacteria. The two groups are characterized by distinctive cytoskeletal architecture with two flagella that emerge near the apical end of the cell [1, 2]. Four jakobid genera have been recognized (*Reclinomonas*, *Histiona*, *Jakoba* and *Seculamonas*), whereas malawimonads comprise a single genus (*Malawimonas*) [1-3]. Jakobids and malawimonads are central to understanding eukaryotic evolution. In particular, jakobids possess the most bacteria-like mitochondrial genome known [3, 4]; in contrast, the mitochondrial DNA (mtDNA) of malawimonads is relatively reduced, with only a few of the extra mitochondrion-encoded genes seen in jakobids [3, 5]. Jakobid mitochondrial genomes have protein-coding genes that exhibit Shine-Dalgarno-like motifs for translation initiation; they encode rRNA and RNase P RNA species having secondary structures that are highly similar to those of their bacterial homologs [4, 6]; and they contain a complement of genes larger than that of any other eukaryotic mtDNA. Additionally, they possess genes for bacteria-like tmRNAs [7], which have not been observed in other characterized mtDNAs. Most strikingly, jakobid mtDNAs encode four subunits (RpoA-D) of a bacterial-type $\alpha_2\beta\beta'\sigma$ RNA polymerase (except that *Jakoba libera* mtDNA retains only two *rpo* genes; Burger, Gray and Lang, unpublished). In all other eukaryotes studied to date, the mitochondrial RNA polymerase is a nucleus-encoded 'T3/T7 phage-type' enzyme [8]. There is little doubt that the jakobid mitochondrial *rpo* genes are derived from the bacterial ancestor of mitochondria [4], and that this bacterial-type RNA polymerase was replaced by the nucleus-encoded phage-type version in other eukaryotes. However, whether or not this substitution occurred early and only once in eukaryotic history remains uncertain.

Jakobids and malawimonads share several specific ultrastructural features with an assemblage of amitochondriate eukaryotes that includes retortamonads, *Trimastix* and *Carpediemonas* [1, 2]. Characters unifying these organisms are a ventral feeding groove, flagellar vanes and other elements of the flagellar apparatus and the cytoskeleton (for a review, see [9]). Based on these observations, a new supergroup, the 'excavates', has been proposed by adding diplomonads (*e.g., Giardia*) and heteroboloseans, both of which possess a feeding groove but lack flagellar vanes, as well as oxymonads, which lack a feeding groove but possess most of the other distinctive features of this grouping [9-12]. Finally, some investigators have expanded this supergroup to include euglenozoans (*e.g., Trypanosoma*) and parabasalids (*e.g., Trichomonas*), based on phylogenies that are, however, not strongly supported (*e.g.*, [9, 13, 14]). However, euglenozoans and parabasalids lack all of the morphological characters typical of excavates. This leaves us with a proposed supergroup, 'excavates', that is considered by some as ill-defined and inappropriately named, and by others as a window on eukaryotic origins [9, 15]. Until unequivocal phylogenetic analyses have resolved these uncertainties, we suggest that the term 'excavates' be discontinued as a formal organismal designation.

Notably, none of the numerous phylogenetic analyses based on nuclear rRNAs [9, 16, 17] or up to six nucleus-encoded proteins [13, 18, 19] recover significant phylogenetic affinities between jakobids and malawimonads. Similarly, the predicted relationships of excavate lineages to other main eukaryotic groups are inconsistent and remain unresolved [13, 16-22]. Several factors may be responsible for the lack of resolution in these molecular phylogenies, including the inadvertent inclusion of paralogous genes (in the case of nucleus-encoded proteins), insufficient sequence data and, finally, systematic error such as long-branch attraction (LBA) [23-25].

In order to determine the phylogenetic position of jakobids and malawimonads, we have generated EST data from seven of the eight recognized jakobid and malawimonad species. The assembled dataset includes 170 nucleus-encoded proteins (40,090 amino acid positions) from representatives of major eukaryotic groups. We present here the first genome-scale phylogenetic analysis that includes jakobids and malawimonads.

RESULTS

Phylogenetic analyses with the complete dataset

The ML tree inferred from the concatenation of 170 nucleus-encoded proteins (Fig. 1) shows 100% bootstrap support value (BV) for nearly all branches, and is congruent with global eukaryotic phylogenies published earlier. The tree, which was rooted according to a gene fusion [26, 27], confirms the monophyly of several previously proposed super-ensembles such as Plantae (red algae, green plants and glaucophytes) [28], and the sister-group relationship of alveolates (apicomplexans, dinoflagellates and ciliates) and stramenopiles [19, 28-30]. Notably, jakobids group together with Euglenozoa with 100% support. In addition, the Approximate Unbiased (AU) test [31] of alternative topologies significantly rejects any topology that does not support this relationship (Table S1).

The jakobid+Euglenozoa monophyly is also supported by insertions/deletions (indels) in two genes coding for large subunit ribosomal proteins. Insertions of 3 and 4-5 amino acids in Rpl22 and Rpl24A, respectively, are found exclusively in jakobids and Euglenozoa, but are absent in all other eukaryotes for which data are available (Fig. 2). The Rpl24A insertion is further absent in Archaea (an archaeal Rpl22 homolog has not been identified).

Despite the large number of amino acid positions used, the relative branching order of malawimonads, jakobids+Euglenozoa, alveolates+stramenopiles, Plantae and opisthokonts+Amoebozoa remains unresolved. The highest, albeit not significant, bootstrap support (57% BV) is observed for the association of jakobids+Euglenozoa and alveolates+stramenopiles, as well as a grouping of this ensemble with Plantae (at 58% BV). Interestingly, the most frequently occurring alternative topology in the bootstrap analyses unites jakobids+Euglenozoa with malawimonads (41% BV), placing them at the base of

Plantae/alveolates+stramenopiles (41% BV). The AU test does not discriminate between the two topologies (Table S1).

Phylogenetic analyses excluding fast-evolving species

In order to understand the basis of the conflicting results, we created two partial datasets, one that excludes the fast-evolving Euglenozoa and the other that excludes jakobids. Interestingly, removal of Euglenozoa (Fig S1A) unifies jakobids and malawimonads into a clade (100% BV) emerging at the base of Plantae/alveolates+stramenopiles (93% BV). In contrast, removal of jakobids (Fig S1B) attracts Euglenozoa to alveolates+stramenopiles (98% BV), while malawimonads remain at the base of Plantae/alveolates+stramenopiles (98% BV).

In order to create a dataset without fast-evolving species, but including representatives of all groups in Fig. 1, we removed all Euglenozoa except *Euglena*, and all alveolates except the slowest-evolving *Toxoplasma* and dinoflagellates (Fig. 3). With this taxon sampling, the monophyly of jakobids+Euglenozoa+malawimonads is supported at 91% BV, and the divergence of this grouping at the base of Plantae+alveolates+stramenopiles has 94% BV. On the other hand, alternative topologies that do not support the monophyly of jakobids+Euglenozoa+malawimonads are not rejected by the AU test (Table S2).

DISCUSSION

Implications for the monophyly of excavates

Excavates as defined initially by ultrastructural features include retortamonads, *Carpediemonas*, diplomonads, *Trimastix*, heteroloboseans, jakobids and malawimonads [9, 11, 12]. When this hypothesis was tested by molecular phylogenetics, the coherence of excavates seemed to vanish: some excavates considered typical (*e.g.*, malawimonads) did not cluster with the other members, whereas Euglenozoa, which lacks the distinctive excavate morphology, joined the group [16, 18-20]. Notably, however, most of these molecular phylogenies have non-significant support (BV < 95%). This lack of resolution can be explained by the use of limited sequence data (at most six genes), a number insufficient to resolve the deep and potentially rapid diversification of early eukaryotes [28, 29, 32, 33]. In addition, these phylogenies include data from parasitic species (*Trichomonas vaginalis* and *Giardia lamblia*) that have been shown to be extremely fast-evolving and thereby strongly affected by LBA [27, 34, 35].

Our phylogenetic analysis, which is based on 170 proteins including newly generated data from (non-parasitic) slowly-evolving taxa (five jakobids and two malawimonads), places jakobids as a sister group to Euglenozoa beyond any doubt. This relationship requires no more than 5,000 amino acid positions to reach >95% bootstrap support (Fig.

S2). Even so, the relationship of jakobids+Euglenozoa to other eukaryotes remains unresolved. Jakobids associate with malawimonads only when fast-evolving species are excluded (Fig. 3, S1A) and a robust malawimonads+jakobids clade forms only when all euglenozoan taxa have been removed from the dataset (Fig. S1A).

It is well known that fast-evolving species tend to mislead phylogenetic reconstruction by LBA [23, 24]. In particular, the fast-evolving Euglenozoa (specifically kinetoplastids) are attracted to other long branches in the dataset [25]. It also has been shown that elimination of long branches often improves phylogenetic accuracy (e.g., [25, 36]). We therefore interpret the inconsistent support for a jakobid+Euglenozoa / malawimonad relationship as the result of two conflicting signals in the dataset [25, 37]: a genuine signal, uniting jakobids+Euglenozoa and malawimonads, placing them at the base of Plantae+alveolates+stramenopiles; and an artifactual non-phylogenetic signal, attracting jakobids+Euglenozoa towards alveolates+stramenopiles.

Effective ways to overcome systematic error include the removal of fast-evolving sites, grouping of amino acids into functional categories, and use of more realistic models of evolution (e.g. the CAT model; [38])[25]. However, applied to our dataset, none of these techniques provided conclusive results (data not shown). Therefore, the sister-group relationship that we suggest here, i.e., jakobids+Euglenozoa with malawimonads, remains to be confirmed with broader taxon sampling.

Implications for mitochondrial evolution

Jakobids possess the most bacteria-like mtDNA currently known [4]. Through their phylogenetic association with Euglenozoa, jakobids are now allied with organisms that have some of the most eccentric mtDNAs yet discovered. For example, trypanosomatid mitochondrial genomes are composed of catenated networks of 'maxicircles' that encode 'cryptic' mitochondrial genes and 'minicircles' that specify guide RNAs; the latter are required for decryption of maxicircle-encoded genes by post-transcriptional mRNA editing. Similarly, the mitochondrial genome of *Euglena* appears to be organized as a heterogeneous collection of numerous DNA molecules [3, 39]. Moreover, recent analyses of another euglenozoan group, Diplonemea, have revealed a unique genome organization, in which small gene pieces are each encoded on separate circular chromosomes ([40]; Roy and Burger, unpublished). The sister-group relationship of Euglenozoa and jakobids demonstrated here implies that their common ancestor had a prototypical bacteria-like mtDNA, as in jakobids, with members of Euglenozoa subsequently inventing various types of multi-partite mitochondrial genomes, highly fragmented genes and RNA editing.

Rooting the eukaryotic tree

The phylogenetic trees presented here are unrooted. This is because the archaeal outgroup (for which most of the 170 genes used here are available) is too distant to permit us a reliable inference of the eukaryotic origin (results not shown), so that other evidence has to be sought. For instance, the bacteria-like, mitochondrion-encoded RNA polymerase of jakobids is substituted by a phage-type, nucleus-encoded enzyme in all other eukaryotes, a character that would place the root basal to jakobids (position A in Figs. 1 and 3). Another story is told by specific insertions in the Rpl24A proteins of jakobids and Euglenozoa, indels that are absent in all other eukaryotes and Archaea. Assuming that this insertion has been gained only once, the root of eukaryotes should be prior to the divergence of jakobids and Euglenozoa.

Yet other combinations of rare genetic characters refute the inference that jakobids are basal, instead placing the root between opisthokonts+Amoebozoa and the remaining eukaryotes (position B in Figs. 1 and 3). Supporting evidence includes: (i) fusion of the dihydrofolate reductase and thymidylate synthase genes (which are separate genes in prokaryotes) in all eukaryotes except opisthokonts [26, 27]; (ii) fusion of three of the six genes encoding enzymes in the pyrimidine synthesis pathway, a genomic arrangement seen only in Amoebozoa and opisthokonts [41, 42]; and (iii) an insertion in elongation factor 1 alpha (EF1a) found only in opisthokonts [43]. Under the assumption that the EF1a insertion and the gene fusions are primitively absent and have been gained only once, the root of eukaryotes would be placed between opisthokonts+Amoebozoa and all other eukaryotes.

These mutually exclusive scenarios for the placement of the eukaryotic root rely all on the (most parsimonious) interpretation of genetic changes that are (arbitrarily) declared to be singular. Yet, the evolutionary history of any of these traits might be more complex than has been assumed. For example, the bacteria-type mitochondrial RNA polymerase of the proto-mitochondrial genome might have been replaced several times independently during eukaryotic evolution. Co-existence of the two types of mitochondrial RNA polymerase genes over a prolonged period (as seen, e.g., in the chloroplasts of land plants [44, 45]) would set the stage for multiple independent losses in different lineages. For gene fusions and gene insertions and deletions, multiple independent occurrences are as conceivable.

In conclusion, the strong phylogenetic weight usually assigned to rare genomic changes is most likely misplaced. These characters are as prone to homoplasy as are sequence characters (e.g., via convergence or reversion)[46]. We argue that single characters such as gene absence/presence, indels and gene fusions are not suited to infer phylogenetic relationships, but might rather be used to add support to otherwise robust sequence-based analyses.

EXPERIMENTAL PROCEDURES

Construction of cDNA libraries and EST sequencing

cDNA libraries from five jakobids (*Reclinomonas americana, Jakoba libera, Jakoba bahamensis, Histiona aroides* and *Seculamonas ecuadoriensis*) and two malawimonads (*Malawimonas californiana* and *Malawimonas jakobiformis*) were generated as described [47]. Plasmids were purified using the QIAprep 96 Turbo Miniprep Kit (Qiagen), and sequencing reactions were performed with the ABI Prism BigDyeTM Terminators version 3.0/3.1 (Perkin-Elmer, Wellesley, MA, USA). The purified reaction products were separated and analyzed on an MJ BaseStation automatic sequencer. Trace files were imported into the TBestDB database (http://tbestdb.bcm.umontreal.ca/searches/login.php) for automated processing, including base calling by PHRED, sequence quality- and vector-trimming, and clustering by PHRAP [48, 49], followed by automated function annotation using AutoFact [50].

Dataset construction

Data from jakobids and malawimonads, and additional sequences from GenBank (http://www.ncbi.nlm.nih.gov/), were added to an existing multiple alignment of protein sequences [28] as described earlier [51]. Ambiguously aligned positions were eliminated with Gblocks [52] followed by manual verification using MUST [53]. Selection of species, genes and orthologous sequences was performed with SCaFoS [54]. Briefly, we chose (i) species that represent all major eukaryotic groups for which genomic data are available from more than one member; (ii) genes that are present in at least 20 of the selected species, and (iii) among these, the slowest-evolving orthologous sequences. The latter sequences were identified as described in [55].

Species evolving at accelerated rates are known to induce tree reconstruction artifacts. Therefore, taxa were excluded when more slowly-evolving relatives were available (*e.g.*, the red alga *Cyanidioschyzon merolae*; [25]). In the case of fungi, animals and embryophyte plants, only representative (preferentially slowly-evolving) members were used. Chimaeras were constructed in some instances either to increase the number of sequence positions, or to obtain the slowest-evolving proteins (see [54]): *Homo* (*H. sapiens, Mus musculus,Canis familiaris, Rattus norvegicus*), *Ciona* (*C. intestinalis, C. savignyi*), *Strongylocentrotus* (*S. purpuratus, Hemicentrotus pulcherrimus*), Ixodidae (*Boophilus microplus, Rhipicephalus appendiculatus, Amblyomma americanum, Amblyomma variegatum*), *Porphyra* (*P. yezoensis, P. haitanensis*), *Leishmania* (*L. major, L. donovani*), *Cryptosporidium* (*C. parvum, C. hominis*), *Toxoplasma* (*T. gondii, Sarcocystis neurona, Neospora caninum*), *Theileria* (*T. annulata, T. parva*), dinoflagellates (*Alexandrium tamarense, Amphidinium carterae, Lingulodinium polyedrum, Karenia brevis*) and *Phytophthora* (*P. sojae, P. ramorum, P. infestans*). The dataset contains 170 proteins (40,090 amino acid positions). Overall, 29% of the theoretical total number of amino acids in the alignment were unavailable (= missing data).

Phylogenetic analyses

Protein sequences were used for phylogenetic analyses. First, we performed preliminary Maximum Likelihood (ML) analyses using TreeFinder [56], with the WAG amino acid replacement matrix, stationary amino acid frequencies estimated from the dataset, and four categories of gamma-distributed rates across sites (WAG+F+Γ4 model). Statistical support was evaluated based on 100 bootstrap replicates.

Because the probability of becoming trapped in local minima is high in genome-scale datasets [57], exhaustive searches were performed as described earlier [28, 51]. Briefly, the topology was 'constrained' (fixed) for the following well-defined and well-supported eukaryotic groups: opisthokonts, alveolates, stramenopiles, Euglenozoa, Amoebozoa, malawimonads, jakobids and Plantae (except for two weakly supported internal branches, one each within jakobids and Plantae). The likelihood values of the resulting 93,555 possible topologies were calculated with PROTML, based on the JTT amino acid substitution matrix without rate-across-site variation to save computer time and memory (separate JTT+F model). From this analysis, the top 1,733 trees (according to their likelihood values) were selected, plus every 500th subsequent topology for a total of 2,000 topologies. The likelihood values of these 2,000 topologies were then determined using TREE-PUZZLE with the WAG+F+Γ4 model. In addition, different combinations of taxon sampling were analyzed using the same procedures. For more technical details see [25].

ACKNOWLEDGEMENTS

We wish to thank Ignacio G. Bravo for helpful comments on the manuscript, and Jean-François Bouffard, Lise Forget, Jung Hwa Seo, Shona Teijeiro, Zhang Wang and Yun Zhu for excellent technical assistance. This work has been supported by Genome Quebec/Canada, the Canadian Institute for Advanced Research, the Canada Research Chairs Program (BFL, MWG and HP) and the Canadian Institute of Health Research (MOP 15331; GB, 42475; BFL). NRE has been supported by the 'Programa de Formación de Investigadores del Departamento de Educación, Universidades e Investigación' (Government of Basque Country).

References

1. O'Kelly, C.J. (1993). The jakobid flagellates: structural features of Jakoba, Reclinomonas and Histiona and implications for the early diversifcation of eukaryotes. J Euk Microbiol *40*, 627-636.
2. O'Kelly, C.J., and Nerad, T.A. (1999). Malawimonas jakobiformis n. gen., n. sp. (Malawimonadidae n. fam): a Jakoba-like heterotrophic nanoflagellate with discoidal mitochondrial cristae. J Euk Microbiol *46*, 522-531.
3. Gray, M.W., Lang, B.F., and Burger, G. (2004). Mitochondria of protists. Annu Rev Genet *38*, 477-524.
4. Lang, B.F., Burger, G., O'Kelly, C.J., Cedergren, R., Golding, G.B., Lemieux, C., Sankoff, D., Turmel, M., and Gray, M.W. (1997). An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature *387*, 493-497.
5. Gray, M.W., Lang, B.F., Cedergren, R., Golding, G.B., Lemieux, C., Sankoff, D., Turmel, M., Brossard, N., Delage, E., Littlejohn, T.G., Plante, I., Rioux, P., Saint-Louis, D., Zhu, Y., and Burger, G. (1998). Genome structure and gene content in protist mitochondrial DNAs. Nucleic Acids Res *26*, 865-878.
6. Seif, E., Cadieux, A., and Lang, B.F. (2006). Hybrid E. coli--Mitochondrial ribonuclease P RNAs are catalytically active. Rna *12*, 1661-1670.
7. Jacob, Y., Seif, E., Paquet, P.O., and Lang, B.F. (2004). Loss of the mRNA-like region in mitochondrial tmRNAs of jakobids. Rna *10*, 605-614.
8. Cermakian, N., Ikeda, T.M., Cedergren, R., and Gray, M.W. (1996). Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage. Nucleic Acids Res *24*, 648-654.
9. Simpson, A.G. (2003). Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). Int J Syst Evol Microbiol *53*, 1759-1777.
10. Patterson, D.J. (1999). The Diversity of Eukaryotes. Am Nat *154*, S96-S124.
11. Simpson, A.G., and Patterson, D.J. (1999). The ultrastructure of Carpediemonas membranifera: (Eukaryota), with reference to the excavate hypothesis. European Journal of Protistology *35*.
12. Cavalier-Smith, T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. Int J Syst Evol Microbiol *52*, 297-354.
13. Edgcomb, V.P., Roger, A.J., Simpson, A.G., Kysela, D.T., and Sogin, M.L. (2001). Evolutionary relationships among "jakobid" flagellates as indicated by alpha- and beta-tubulin phylogenies. Mol Biol Evol *18*, 514-522.
14. Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. Science *290*, 972-977.
15. Cavalier-Smith, T. (1998). A revised six-kingdom system of life. Biol Rev Camb Philos Soc *73*, 203-266.
16. Simpson, A.G., Roger, A.J., Silberman, J.D., Leipe, D.D., Edgcomb, V.P., Jermiin, L.S., Patterson, D.J., and Sogin, M.L. (2002). Evolutionary history of "early-diverging" eukaryotes: the excavate taxon Carpediemonas is a close relative of Giardia. Mol Biol Evol *19*, 1782-1791.

17. Cavalier-Smith, T. (2004). Only six kingdoms of life. Proc Biol Sci *271*, 1251-1262.
18. Archibald, J.M., O'Kelly, C.J., and Doolittle, W.F. (2002). The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. Mol Biol Evol *19*, 422-431.
19. Simpson, A.G., Inagaki, Y., and Roger, A.J. (2006). Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. Mol Biol Evol *23*, 615-625.
20. Cavalier-Smith, T. (2003). The excavate protozoan phyla Metamonada Grasse emend. (Anaeromonadea, Parabasalia, Carpediemonas, Eopharyngia) and Loukozoa emend. (Jakobea, Malawimonas): their evolutionary affinities and new higher taxa. Int J Syst Evol Microbiol *53*, 1741-1758.
21. Burger, G., Saint-Louis, D., Gray, M.W., and Lang, B.F. (1999). Complete sequence of the mitochondrial DNA of the red alga Porphyra purpurea. Cyanobacterial introns and shared ancestry of red and green algae. Plant Cell *11*, 1675-1694.
22. Lang, B.F., O'Kelly, C., Nerad, T., Gray, M.W., and Burger, G. (2002). The closest unicellular relatives of animals. Curr Biol *12*, 1773-1778.
23. Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool *27*, 401-410.
24. Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., and Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol *54*, 743-757.
25. Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., and Philippe, H. (In press). Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol.
26. Stechmann, A., and Cavalier-Smith, T. (2002). Rooting the eukaryote tree by using a derived gene fusion. Science *297*, 89-91.
27. Philippe, H., Lopez, P., Brinkmann, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Muller, M., and Le Guyader, H. (2000). Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc R Soc Lond B Biol Sci *267*, 1213-1221.
28. Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S.C., Roure, B., Burger, G., Loffelhardt, W., Bohnert, H.J., Philippe, H., and Lang, B.F. (2005). Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol *15*, 1325-1330.
29. Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M., and Philippe, H. (2002). The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A *99*, 1414-1419.
30. Patron, N.J., Rogers, M.B., and Keeling, P.J. (2004). Gene replacement of fructose-1,6-bisphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates. Eukaryot Cell *3*, 1169-1175.

31. Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. Syst Biol *51*, 492-508.
32. Knoll, A.H. (1992). The early evolution of eukaryotes: a geological perspective. Science *256*, 622-627.
33. Philippe, H., and Adoutte, A. (1998). The molecular phylogeny of Eukaryota: solid facts and uncertainties. In Evolutionary relationships among Protozoa, G. Coombs, K. Vickerman, M. Sleigh and A. Warren, eds. (Dordrecht: Kluwer), pp. 25-56.
34. Philippe, H., and Germot, A. (2000). Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Mol Biol Evol *17*, 830-834.
35. Embley, T.M., and Hirt, R.P. (1998). Early branching eukaryotes? Curr Opin Genet Dev *8*, 624-629.
36. Aguinaldo, A.M., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., and Lake, J.A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. Nature *387*, 489-493.
37. Baurain, D., Brinkmann, H., and Philippe, H. (In press). Lack of resolution in the animal phylogeny: closely spaced cladeogeneses or undetected systematic errrors? Mol Biol Evol.
38. Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol *21*, 1095-1109.
39. Tessier, L.H., van der Speck, H., Gualberto, J.M., and Grienenberger, J.M. (1997). The cox1 gene from Euglena gracilis: a protist mitochondrial gene without introns and genetic code modifications. Curr Genet *31*, 208-213.
40. Marande, W., Lukes, J., and Burger, G. (2005). Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. Eukaryot Cell *4*, 1137-1146.
41. Nara, T., Hshimoto, T., and Aoki, T. (2000). Evolutionary implications of the mosaic pyrimidine-biosynthetic pathway in eukaryotes. Gene *257*, 209-222.
42. Stechmann, A., and Cavalier-Smith, T. (2003). The root of the eukaryote tree pinpointed. Curr Biol *13*, R665-666.
43. Baldauf, S.L., and Palmer, J.D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci U S A *90*, 11558-11562.
44. Gray, M.W., and Lang, B.F. (1998). Transcription in chloroplasts and mitochondria: a tale of two polymerases. Trends Microbiol *6*, 1-3.
45. Richter, U., Kiessling, J., Hedtke, B., Decker, E., Reski, R., Borner, T., and Weihe, A. (2002). Two RpoT genes of Physcomitrella patens encode phage-type RNA polymerases with dual targeting to mitochondria and plastids. Gene *290*, 95-105.
46. Bapteste, E., and Philippe, H. (2002). The potential value of indels as phylogenetic markers: position of trichomonads as a case study. Mol Biol Evol *19*, 972-977.
47. Rodríguez-Ezpeleta, N., Teijeiro, S., Forget, L., Burger, G., and Lang, B.F. (In press). Generation of cDNA libraries: protists and fungi In Methods in Molecular Biology: Methods in ESTs, J. Parkinson, ed. (Totowa, NJ: Humana Press).

48. Ewing, B., Green, P., Hillier, L., and Wendl, M.C. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res *8*, 186-194.
49. Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res *8*, 175-185.
50. Koski, L.B., Gray, M.W., Lang, B.F., and Burger, G. (2005). AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics *6*, 151.
51. Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W., and Casane, D. (2004). Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol *21*, 1740-1752.
52. Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol *17*, 540-552.
53. Philippe, H. (1993). MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res *21*, 5264-5272.
54. Roure, B., Rodríguez-Ezpeleta, N., and Philippe, H. (In press). SCaFoS: Selection, Concatenation and Fusion of Sequences for phylogenomics. BMC Evol Biol.
55. Philippe, H., Lartillot, N., and Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. Mol Biol Evol *22*, 1246-1253.
56. Jobb, G., von Haeseler, A., and Strimmer, K. (2004). TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol *4*, 18.
57. Salter, L.A. (2001). Complexity of the likelihood surface for a large DNA dataset. Syst Biol *50*, 970-978.

Fig. 1

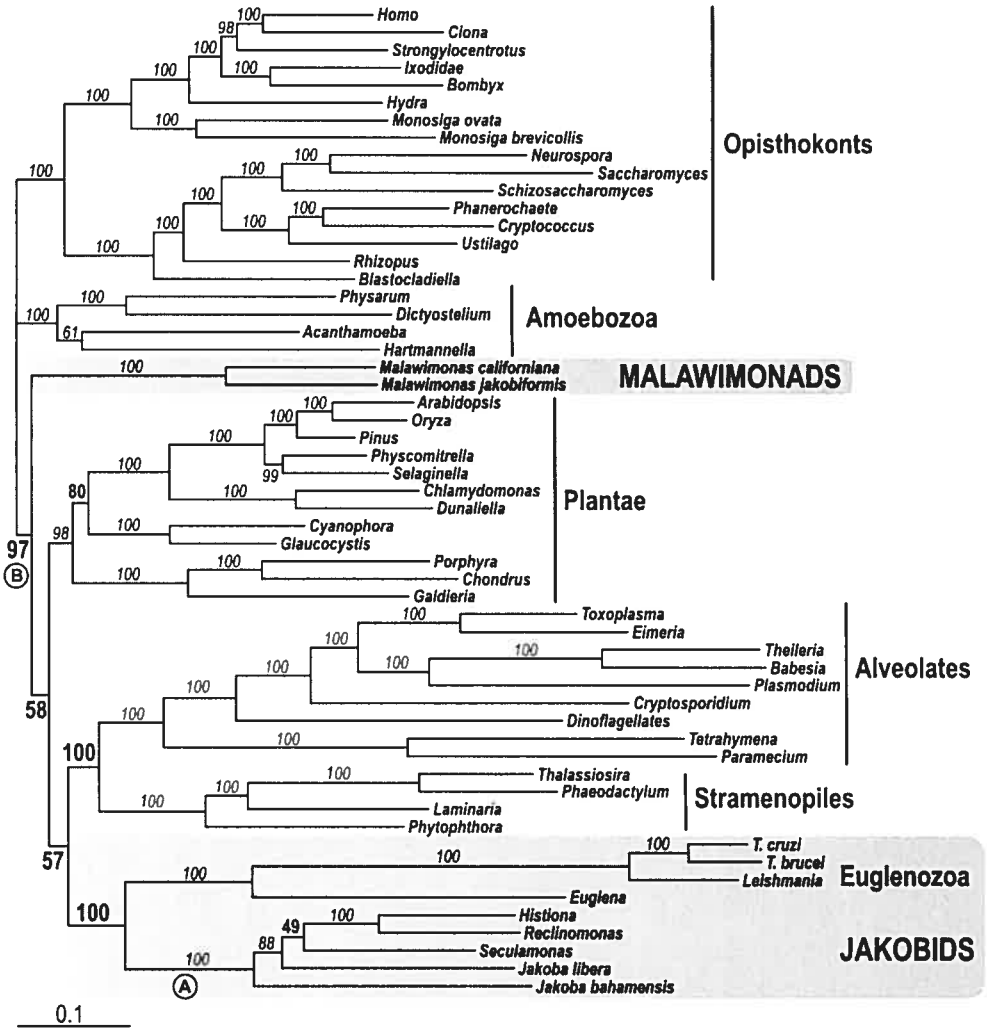

Fig 1. Maximum Likelihood tree obtained with 170 concatenated nucleus-encoded proteins (40,090 amino acid positions) from 56 eukaryotic species. Numbers in italics, support values of TreeFinder analysis (100 replicates) with the concatenated dataset and the WAG+F+Γ model. Numbers in bold, support values of exhaustive ML analysis (10,000 RELL replicates). The scale bar denotes the estimated number of amino acid substitutions per site. The tree was rooted between the opisthokonts and other eukaryotes (excluding Amoebozoa) according to a gene fusion [26, 27]. Encircled A and B indicate alternative positions of the root (see Discussion).

Fig. 2

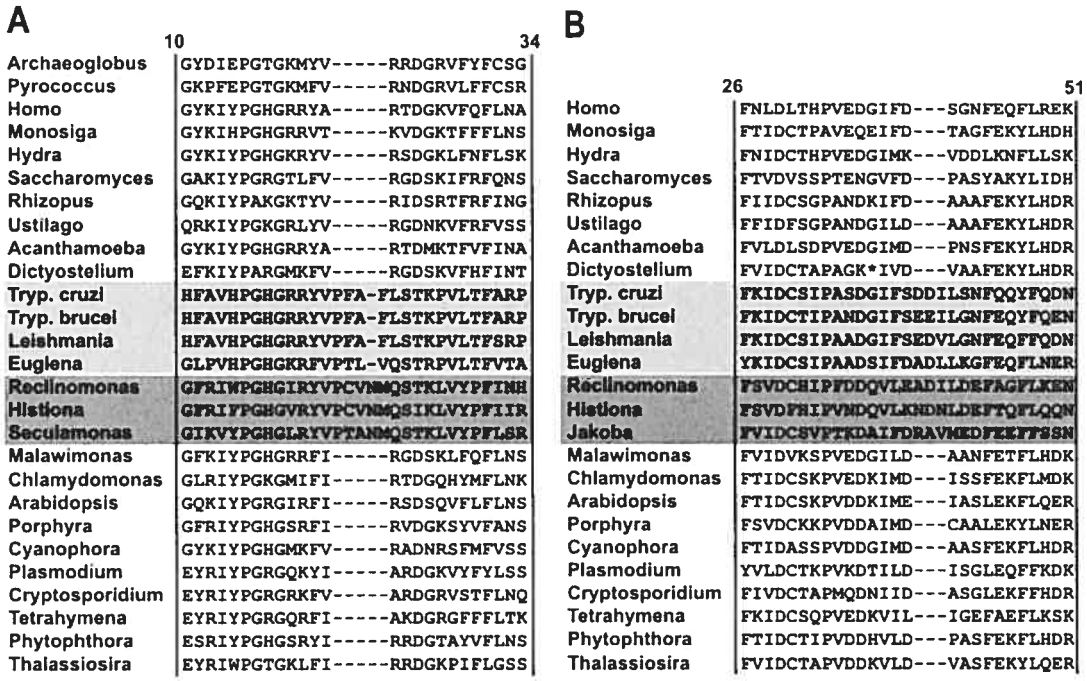

Fig. 2. Amino acid insertions specific to Euglenozoa and jakobids. Parts of the amino acid sequence alignments of Rpl24A (A) and Rpl22 (B) are shown. Numbers above the alignment indicate their position with respect to *Homo* as reference. Only one or two representative species per group are included. Euglenozoa and jakobids are highlighted in light and dark grey, respectively.

Fig. 3

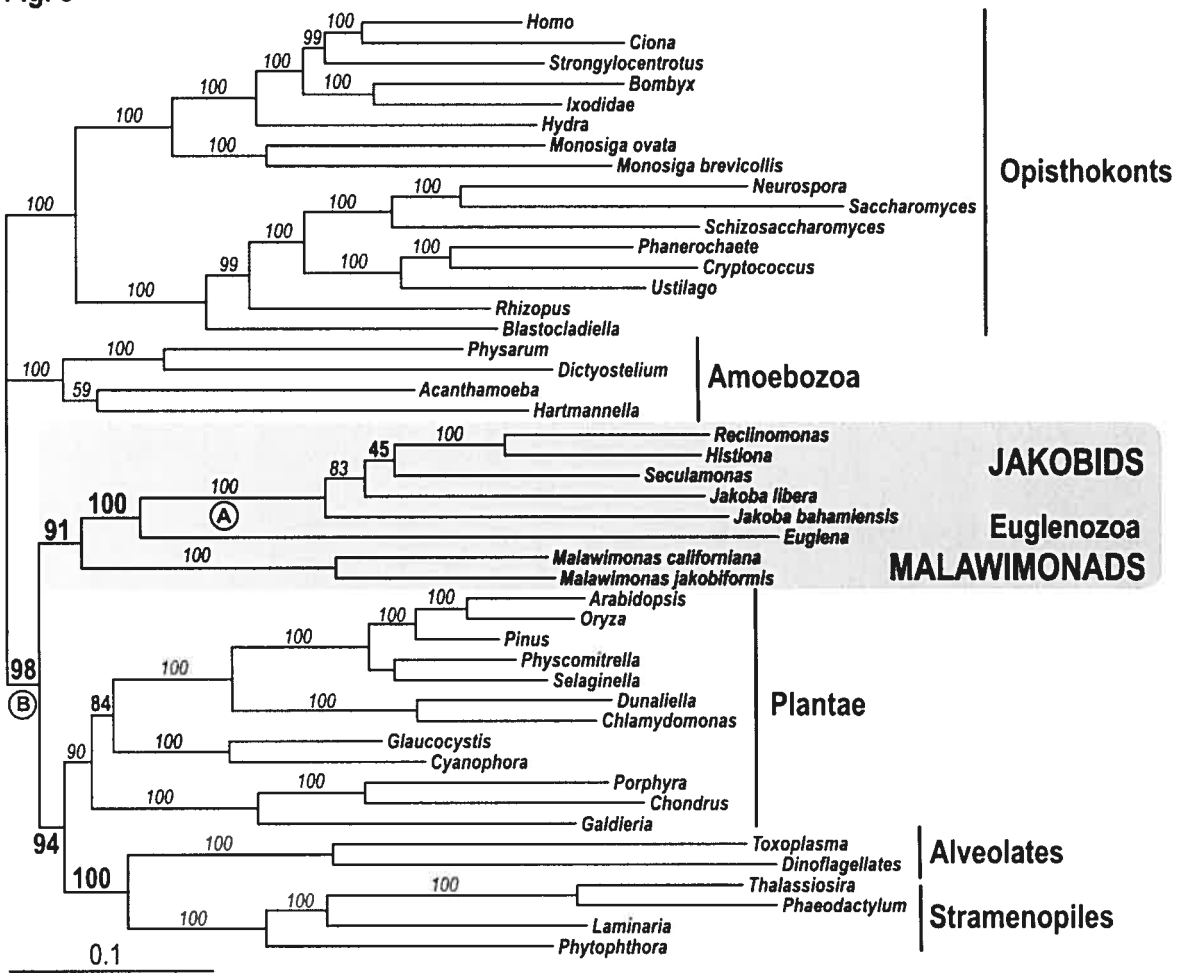

Fig. 3. Maximum Likelihood tree for slow-evolving species. The analysis was performed as in Fig. 1, except for restricting the taxon sampling to slow-evolving alveolates (*Toxoplasma* and dinoflagellates) and Euglenozoa (*Euglena*).

Supplementary Material

Table S1: Likelihood Tests of Alternative Tree Topologies

Rank	Tree topology	ΔlnL[a]	AU[b]
1	((Op,Am),(Ma,(Pl,((Ja,Eu),(St,Al))))); Fig. 1	-15.0	0.692
3	((Op,Am),((Ma,(Ja,Eu)),(Pl,(St,Al))))	15.0	0.471
14	((Op,Am),(Pl,((Ma,(Ja,Eu)),(St,Al))))	51.8	0.105
15	((Op,Am),(Ma,((Ja,Eu),(Pl,(St,Al)))))	46.5	0.055
87	((Op,Am),((Ja,(Ma,Eu)),(Pl,(St,Al))))	320.0	**6e-45**
104	((Op,Am),(((Ja,Ma),Eu),(Pl,(St,Al))))	329.1	**2e-55**
270	((Op,Am),(Ma,(Pl,(Ja,(Eu,(St,Al))))))	420.1	**4e-06**

Comparison of alternative topologies with CONSEL (Shimodaira and Hasegawa, 2001), inferred from maximum likelihood analyses of 170 proteins and 54 species. Topology 3 is the best topology where jakobids, Euglenozoa and malawimonads are monophyletic; topology 14 shows a different placement for the jakobids+Euglenozoa+malawimonads clade; topology 15 shows an alternative position for jakobids+Euglenozoa from that of topology 1; topology 87 is the best where jakobids and Euglenozoa are not monophyletic; topology 104 shows an alternative position for malawimonads from that in topology 87; topology 270 is the best topology where the monophyly of jakobids and Euglenozoa is disturbed by a group other than malawimonads. The following abbreviations are used: Op, opisthokonts; Am, Amoebozoa; Ma, malawimonads; Pl, Plantae; Ja, jakobids; Eu, Euglenozoa; St, stramenopiles; and Al, Alveolata. Significant values are shown in bold type.

[a]Log likelihood difference

[b]Approximate Unbiased test.

Table S2: Likelihood Tests of Alternative Tree Topologies Excluding Long Branches

Rank	Tree topology	ΔlnL[a]	AU[b]
1	((Op,Am),((Ma,(Ja,Eu)),(Pl,(St,Al)))); Fig 3	-52.0	0.975
10	((Op,Am),(Ma,((Ja,Eu),(Pl,(St,Al)))))	52.0	0.117
13	((Op,Am),(Ma,(Pl,((Ja,Eu),(St,Al)))))	76.5	0.094
16	((Op,Am),(((Ma,(Ja,Eu)),Pl),(St,Al)))	85.0	**0.012**
17	((Op,Am),(Pl,((Ma,(Ja,Eu)),(St,Al))))	84.7	**0.027**
29	(Op,(Ma,(Am,((Ja,Eu),(Pl,(St,Al))))))	95.6	**0.033**

Comparison of alternative topologies with CONSEL (Shimodaira and Hasegawa, 2001), inferred from maximum likelihood analyses of 170 proteins and 54 species. Topology 10 is the best topology where jakobids, Euglenozoa and malawimonads are not monophyletic; topology 13 shows a different placement for the jakobids+Euglenozoa from that in topology 10; topologies 16 and 17 show alternative positions for jakobids+Euglenozoa+malawimonads from that topology 1; topology 29 shows an alternative placement for malawimonads. The following abbreviations are used: Op, opisthokonts; Am, Amoebozoa; Ma, malawimonads; Pl, Plantae; Ja, jakobids; Eu, Euglenozoa; St, stramenopiles; and Al, Alveolata. Significant values are shown in bold type.

[a]Log likelihood difference

[b]Approximate Unbiased test

Fig. S1

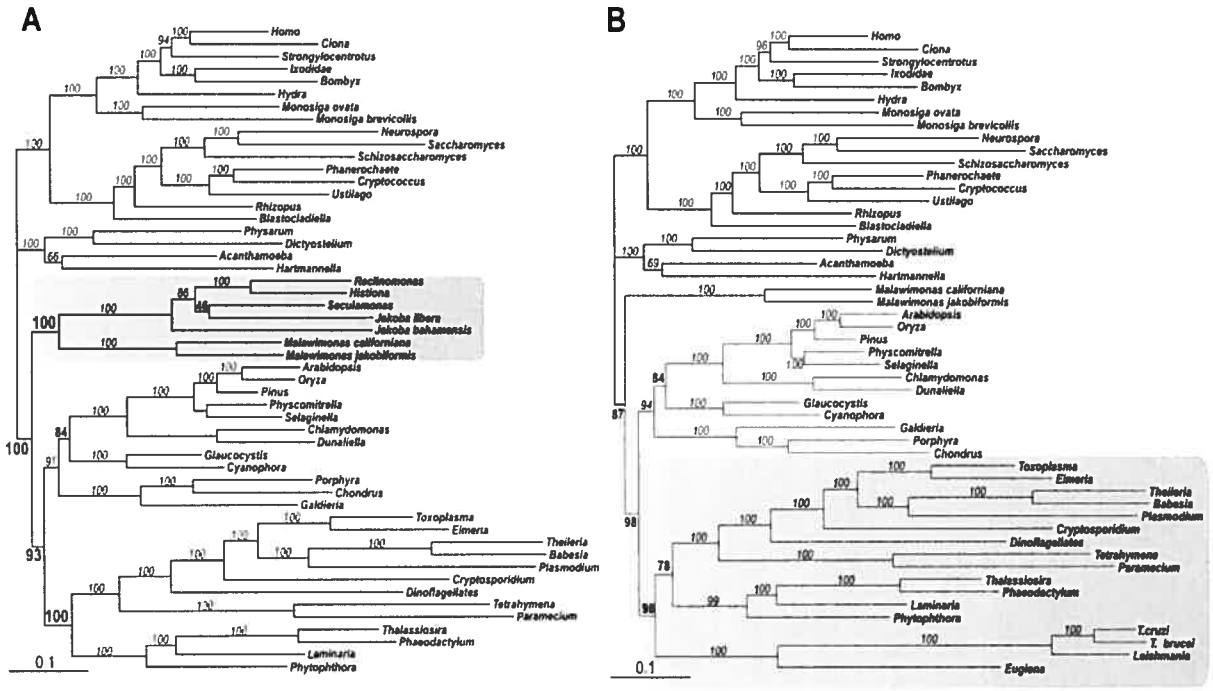

Fig S1. Maximum Likelihood tree obtained with 170 concatenated nucleus-encoded proteins but without Euglenozoa (A) or without jakobids (B). Numbers in italics, support values of TreeFinder analysis (100 replicates) with the concatenated dataset and the WAG+F+Γ model (PhyML provided similar results). Numbers in bold, support values of exhaustive ML analysis (10,000 RELL replicates). The scale bar denotes the estimated number of amino acid substitutions per site.

Fig. S2

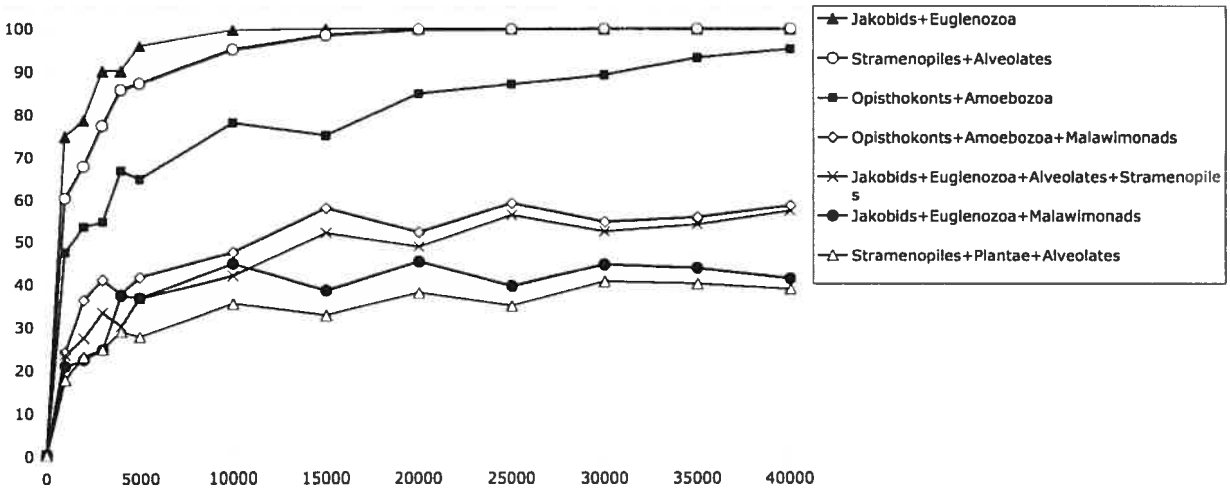

Fig. S2: Evolution of the bootstrap support values for the major groups, as a function of the number of amino acid positions.
Variable fractions of the complete dataset were randomly drawn 100 times. RELL bootstrap analyses were then performed on each of the 100 samples for each size fraction. The average of the bootstrap values for each size fraction was plotted against its size. Y and X branches refer to bootstrap values (in %) and number of amino acid positions, respectively.

References

1. Shimodaira, H., et M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246-7

DISCUSSION

1. L'approche EST

L'utilisation de jeux de données à l'échelle génomique avec un grand nombre d'espèces réduit les erreurs stochastiques et systématiques et s'avère indispensable pour résoudre la phylogénie des eucaryotes. Une telle approche demande un grand nombre de séquences génomiques de beaucoup d'espèces. Bien que les génomes complets de 45 eucaryotes ont été séquencés à de jour et que ceux de quelques centaines d'autres sont en cours[10], les groupes d'eucaryotes les plus obscurs ne sont pas représentés dans cet échantillonnage. En effet, étant donné le temps et les ressources requis, le séquençage des génomes complets de tous les eucaryotes est impensable. Il faut donc des alternatives pour l'obtention de données permettant de réaliser des analyses phylogénomiques incluant un grand nombre de taxons (Philippe et Telford, 2006).

Parmi ces alternatives, une approche standard et largement utilisée en systématique est l'amplification par PCR (pour *Polymerase Chain Reaction*) et le séquençage ciblé de quelques gènes universels. Par exemple, l'ARNr, les actines α et β, les tubulines α et β, les EF1α et 2 et les HSP70 et 90 ont été amplifiés pour un vaste éventail de groupes d'eucaryotes et utilisés pour étudier les relations entre eux. Cependant, il est maintenant connu qu'une dizaine de gènes ne sont pas suffisantes pour résoudre la phylogénie des eucaryotes. L'expansion de l'approche PCR à une vaste sélection de gènes est difficile car elle implique la définition *a priori* des gènes appropriés (p. ex., minimalement affectés par la paralogie ou le transfert horizontal) et le développement d'amorces dégénérées pour l'amplification de chaque gène, ce qui représente un coût considérable en temps et en argent. Une alternative réalisable et à mi-chemin entre le séquençage de génomes complets et l'approche par PCR est l'échantillonnage de séquences codantes de plusieurs génomes par le séquençage d'ESTs (Rudd, 2003). Dû aux différents niveaux d'expression des gènes, le séquençage d'une fraction relativement petite de transcrits permet d'obtenir un sous-ensemble de gènes chevauchant entre plusieurs espèces et qui sont, en plus, appropriés pour la reconstruction phylogénétique (Hughes *et al.*, 2006).

Pour augmenter le nombre de séquences génomiques disponibles pour les groupes d'eucaryotes considérés dans nos études, nous avons généré des banques d'ADNc d'un

[10] Selon la base de données GOLD (http://www.genomesonline.org/gold.cgi) actualisée le 4-02-2007

glaucophyte, de cinq jakobides et de deux malawimonadines. Les séquences obtenues ont été assemblées en « clusters » et déposées dans la base de données publique TBestDB (O'Brien *et al.*, 2007), qui contient des ESTs d'un grand échantillonnage de protistes et de champignons unicellulaires. Le Tableau VI montre le nombre d'ESTs et de clusters obtenus pour chaque espèce ainsi que le pourcentage de protéines retrouvées parmi les 140 les plus fréquentes du jeu de données de 56 eucaryotes utilisé dans le Chapitre V. Tel qu'illustré par les chiffres obtenus, il existe une variabilité importante entre les différents organismes au niveau de la redondance de la banque (nombre de ESTs par cluster). Par exemple, le nombre de clusters obtenus pour *Glaucocystis*, pour *Seculamonas* et pour *M. californiana* est presque le même, alors que le nombre de ESTs de *Glaucocystis* séquencés est nettement supérieur, la majorité d'entre eux provenant, en plus, de banques normalisées. Quoi qu'il en soit, le séquençage de quelques milliers d'ESTs a été suffisant pour pouvoir inclure des séquences de ces espèces dans des analyses phylogénomiques.

Tableau VI : Nombre de EST et de clusters obtenus pour chaque organisme

Espèce	Groupe	Nbr. de ESTs[1]	Nbr. de clusters[2]	Pourcentage du jeu test[3]
Glaucocystys nostochinearum	Glaucophyte	8 745 (67%)	2 831	65%
Reclinomonas americana	Jakobide	19 211 (39%)	6 797	83%
Histiona aroides	Jakobide	4 168 (0%)	1 763	51%
Jakoba libera	Jakobide	5 752 (3%)	2 565	49%
Jakoba bahamensis	Jakobide	4 812 (0%)	2 286	49%
Seculamonas ecuadoriensis	Jakobide	5 418 (0%)	2 217	67%
Malawimonas californiana	Malawimonadine	4 746 (3%)	2 314	64%
Malawimonas jakobiformis	Malawimonadine	11 001 (15%)	4 505	77%

[1] Nombre d'ESTs de plus que 60 nucléotides une fois que les régions de mauvaise qualité et que la séquence du vecteur aient été éliminées. Le pourcentage des ESTs provenant des banques normalisées est indiqué entre parenthèses.
[2] Nombre de clusters obtenus à partir de l'assemblage d'ESTs tel que décrit dans O'Brien *et al.* (2007).
[3] Pourcentage des 140 protéines les plus exprimées du jeu de données de 56 eucaryotes présenté dans le Chapitre V.

2. Implications biologiques sur l'évolution des eucaryotes

2.1. Une origine unique des plastes

Le test ultime pour confirmer une unique endosymbiose primaire à l'origine des plastes est la congruence entre les phylogénies basées sur des gènes du plaste et celles basées sur des gènes de l'hôte (nucléaires et mitochondriaux). Celles-ci doivent supporter respectivement, la monophylie des plastes par rapport aux cyanobactéries et la monophylie des trois lignées de photosynthétiques primaires par rapport aux autres eucaryotes.

2.1.1. Support pour la monophylie des plastes

Nos analyses phylogénétiques basées sur la concaténation de 50 protéines (10 334 positions d'acides aminés) de 16 plastes et de 15 cyanobactéries supportent la monophylie des plastes (VB = 100%). Cette étude est la première à inclure un aussi grand nombre de gènes d'un vaste échantillonnage de plastes et de cyanobactéries. Pour placer les plastes au sein de la phylogénie des cyanobactéries, nous avons aussi analysé un sous-ensemble de 26 protéines incluant 13 bactéries additionnelles utilisées pour raciner l'arbre. La phylogénie obtenue supporte la position basale de *Gloeobacter* (VB = 100%) et le placement des plastes comme groupe frère du reste des cyanobactéries (VB = 91%). Cependant, sachant que la présence d'un groupe extérieur a généralement un effet négatif ou, au plus, nul dans l'inférence phylogénétique (Hirt *et al.*, 1999; Holland, Penny et Hendy, 2003), nous avançons que la hausse dans la résolution des relations entre les cyanobactéries en présence d'un groupe externe est probablement due à un artefact de reconstruction phylogénétique.

Des artefacts de reconstruction phylogénétique comme l'attraction des longues branches, l'attraction des espèces avec composition nucléotidique similaire ou l'hétérotachie ont aussi été considérés responsables d'entraîner artificiellement la monophylie des plastes (Lockhart *et al.*, 1992; Lockhart *et al.*, 1998). Pour tester ceci, nous avons réalisé des analyses (i) incluant seulement les espèces à taux d'évolution lent pour éviter l'attraction des longues branches, (ii) avec la méthode LogDet pour traiter le biais

compositionnel et (iii) avec un modèle covarion pour traiter l'hétérotachie. La monophylie des plastes est supportée avec VB = 100% dans tous les cas. Le fait que ce résultat soit robuste à tous ces tests indique qu'il n'est pas dû à des artefacts d'inférence phylogénétique. De plus, la monophylie des plastes est aussi supportée par d'autres évidences non phylogénétiques, comme la présence de gènes codant pour une protéine de liaison à la chlorophylle de triple hélice et pour les protéines TIC110 et TOC34 (voir section 2.1.2.2 de l'introduction). Il existe des évidences pour la présence de TIC110 chez les glaucophytes (Steiner *et al.*, 2005); malheureusement, aucun des deux autres gènes n'a été trouvé dans les ESTs récemment générés. Ceci rend impossible de déterminer si ces protéines furent inventées par l'ancêtre des algues rouges et des algues vertes ou par celui des trois lignées de photosynthétiques primaires.

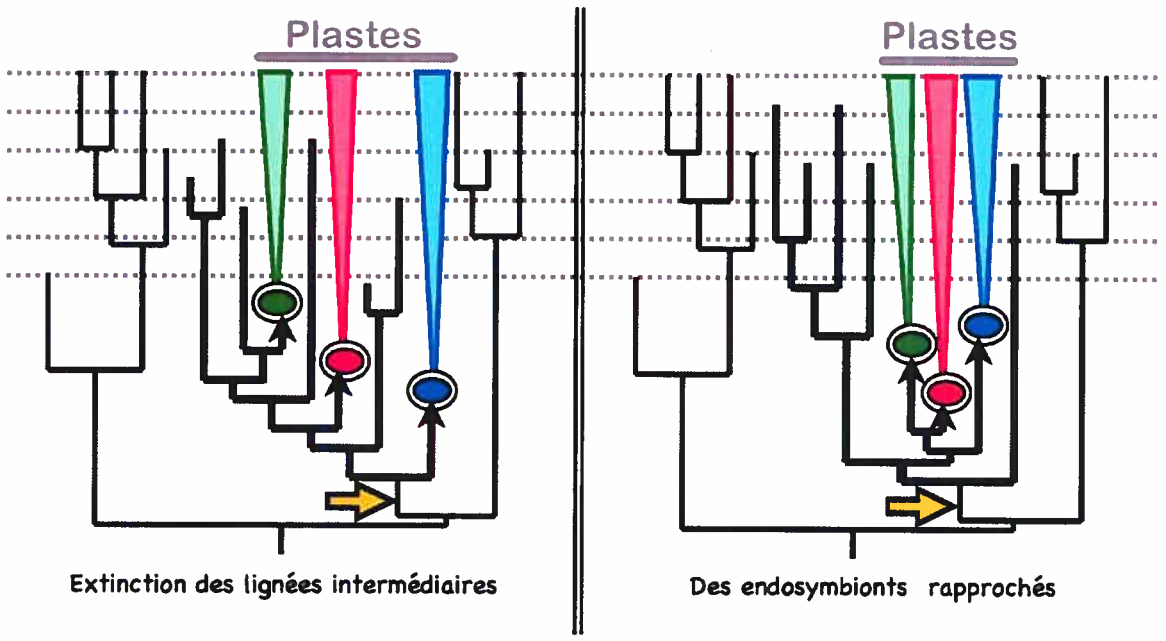

Figure 37 : La monophylie des plastes et des endosymbioses primaires multiples.
Deux scénarios qui concilient la monophylie des plastes avec des endosymbioses multiples sont présentés. Les arbres illustrent l'évolution hypothétique des cyanobactéries, incluant les endosymbioses menant aux trois lignées de plastes (en rouge, vert et bleue). Les lignes discontinues indiquent des événements d'extinction. Les flèches jaunes indiquent les branches où l'origine des caractéristiques uniques des plastes a pu avoir lieu.
Basé sur la Figure 5 de Stiller, Reel et Johnson (2003)

Bien que la monophylie des plastes semble être confirmée par les analyses phylogénétiques et par d'autres caractères indépendants, elle a seulement été testée par rapport aux cyanobactéries connues à ce jour. D'ailleurs, la monophylie des plastes est tout à fait compatible avec des endosymbioses primaires multiples de cyanobactéries différentes suivies de l'extinction des lignées intermédiaires (Figure 37A). Dans un tel cas, les caractères uniques aux plastes auraient été inventés chez l'ancêtre de la lignée cyanobactérienne dont tous les descendants sauf les plastes seraient disparus. Également, la monophylie des plastes peut aussi s'expliquer par des endosymbioses primaires multiples de cyanobactéries phylogénétiquement rapprochées (Figure 37B) (Stiller, Reel et Johnson, 2003). Étant dans l'impossibilité de tester ces deux hypothèses avec des analyses phylogénétiques sur des données du plaste, l'hypothèse d'une endosymbiose primaire unique doit être confirmée avec l'obtention de la monophylie des cellules hôtes à partir des données nucléaires et/ou mitochondriales.

2.1.2. Support pour la monophylie des hôtes

Nos analyses basées sur la concaténation de 143 protéines (30,113 positions d'acides aminés) codées dans le noyau de 34 espèces eucaryotes supportent la monophylie des trois groupes d'eucaryotes photosynthétiques primaires par rapport aux opisthocontes, aux Amoebozoa, aux alvéolés et aux straménopiles (BV = 97-98%). De plus, ce résultat est supporté par l'analyse d'un autre jeu de données (170 protéines de 56 groupes eucaryotes), qui inclut aussi les euglénozoaires, les jakobides et les malawimonadines. Cependant, il reste encore à confirmer si les Plantae sont monophylétiques en présence d'autres groupes eucaryotes pour lesquels des données génomiques n'étaient pas disponibles au moment de l'analyse : par exemple, les haptophytes, les cryptophytes, les Rhizaria et d'autres membres des excavés.

Les analyses basées sur des données du plaste et celles basées sur des données du noyau donnent donc des résultats congruents et supportent une unique endosymbiose primaire à l'origine des plastes des algues rouges, des plantes vertes et des glaucophytes. Pour tester si le troisième compartiment cellulaire, soit la mitochondrie, raconte la même histoire que les deux autres, nous avons réalisé des analyses phylogénétiques basées sur 11

protéines mitochondriales (analyses non publiées[11]). Dans l'arbre obtenu (Figure 38), les plantes vertes, les algues rouges et les glaucophytes forment un groupe monophylétique, mais son support n'est pas statistiquement significatif (VB = 50-62%). Étant donné le grand nombre de positions nécessaire pour retrouver la monophylie des Plantae avec les données nucléaires (Figure 3 du Chapitre II), ceci n'est pas surprenant. En effet, avec un nombre de positions équivalent à celui utilisé pour l'analyse mitochondriale, le support statistique pour les Plantae dans l'analyse nucléaire est d'environ 30%.

La manque de résolution obtenue dans les analyses adressant la monophylie des Plantae réalisées auparavant s'explique aussi par le faible nombre de positions utilisé. Par exemple, l'analyse la plus complète qui inclut les trois lignées d'eucaryotes photosynthétiques primaires qui a été réalisée jusqu'à date se base sur six gènes nucléaires, soit 1 938 positions d'acides aminés (Moreira, Le Guyader et Philippe, 2000). Dans cette étude, la monophylie des Plantae n'est pas statistiquement supportée. Nos analyses sont donc les premières à supporter la monophylie des trois groupes d'eucaryotes photosynthétiques primaires, ce qui, conjointement avec la monophylie des plastes, supporte un origine unique de ces derniers.

Si nos analyses avaient supporté la non-monophylie des plastes, on aurait conclu que des endosymbioses primaires multiples sont à l'origine des plastes. Cependant, cette même conclusion n'aurait pas pu être avancée si les analyses basées sur des protéines nucléaires ou mitochondriales avaient supporté la non-monophylie des Plantae. En effet, la non-monophylie des organismes portant des plastes primaires pourrait s'expliquer par une endosymbiose primaire unique suivie de la perte du plaste chez certaines lignées. D'ailleurs, un tel scénario fut proposé par Nozaki *et al.* (Nozaki, 2005; Nozaki *et al.*, 2003) suite au réanalyse du jeu de données de Moreira *et al.* (2000), dans lequel ils montraient que l'exclusion du EF2 était suffisante pour que la monophylie des Plantae ne soit plus supportée. Dans leur modèle (Figure 39), ils proposaient une seule endosymbiose primaire à la base des bicontes suivie de la perte du plaste chez les lignées dépourvues de plaste primaire. Bien évidemment, nos analyses, avec les espèces qu'elles comportent, ne supportent pas un tel scénario. Cependant, une telle possibilité doit être tenue en compte pour des analyses futures incluant d'autres groupes eucaryotes.

[11] Ces analyses incluent des séquences non publiques du génome mitochondrial de *Cyanophora paradoxa*,

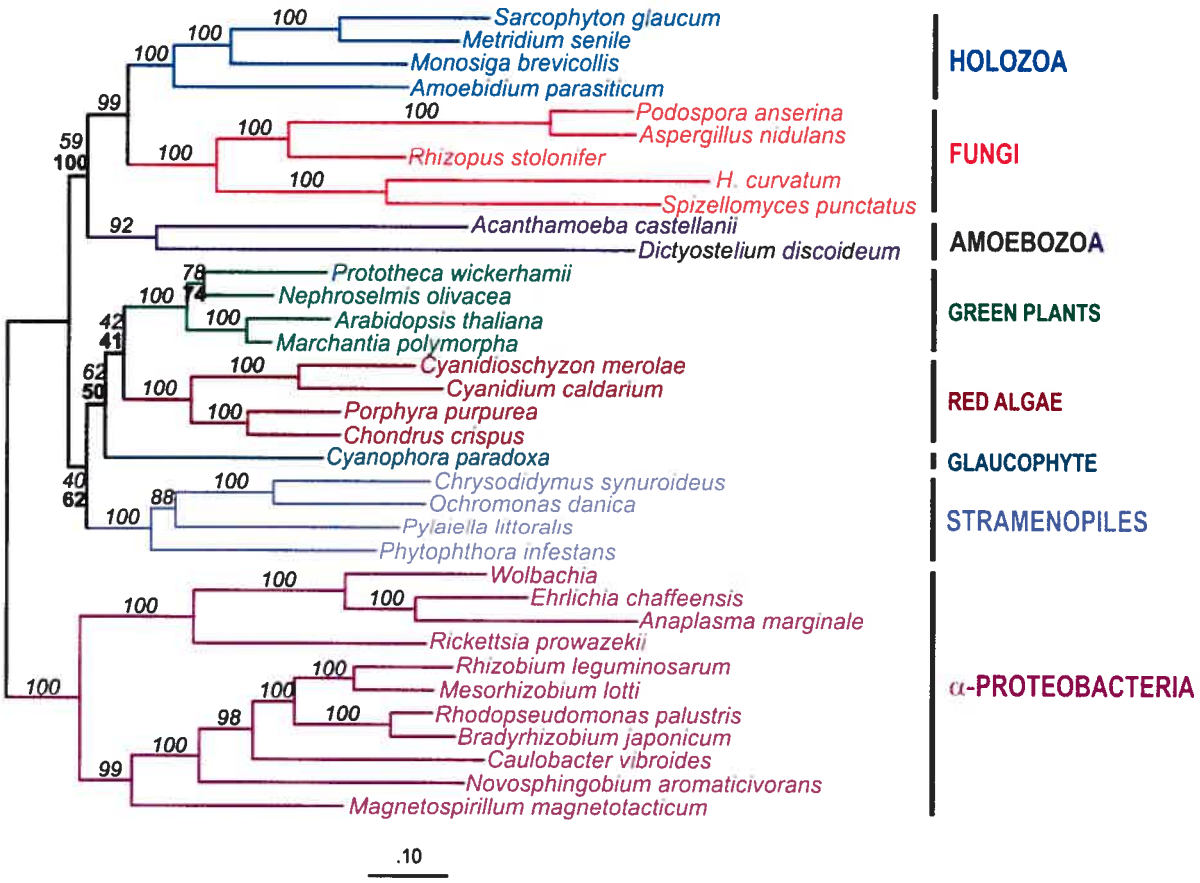

Figure 38 : Phylogénie basée sur la concaténation de 11 protéines mitochondriales

Arbre de vraisemblance maximale obtenu à partir de l'analyse de la concaténation de 11 protéines mitochondriales (2 298 positions d'acides aminés). Les nombres en italique représentent des valeurs de bootstrap obtenues par l'analyse de 100 pseudo-répliqués avec PhyML (Guindon et Gascuel, 2003) et le modèle WAG+F+Γ. Les nombres en gras représentent des valeurs de bootstrap obtenues par l'analyse de 10 000 pseudo-répliqués RELL par la méthode séparée. Le présence d'une seule valeur indique que la branche correspondante fut contrainte pour l'analyse séparée (voir « *Experimental Procedures* » dans le Chapitre II pour plus de détails).

mises à ma disposition par Franz Lang, Gertraud Burger et Michael W. Gray.

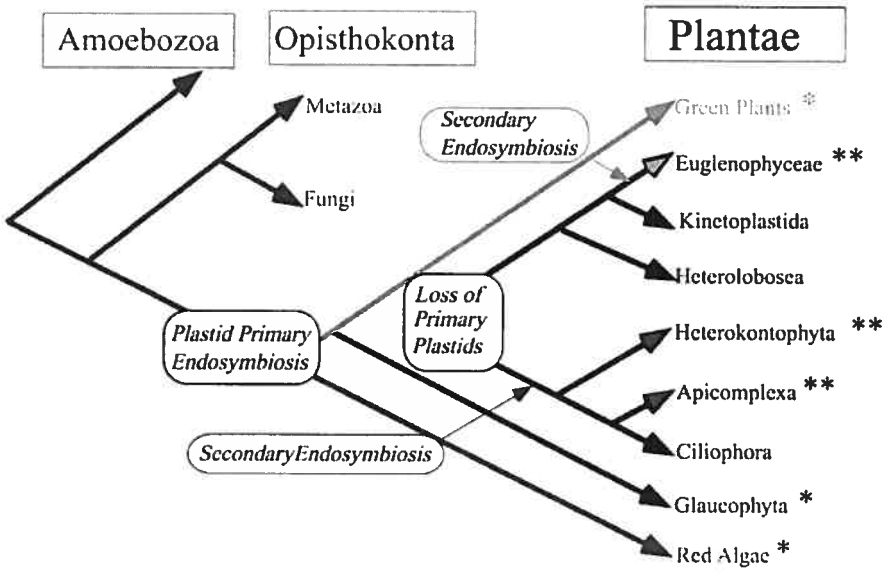

Figure 39 : Hypothèse alternative pour l'évolution des plastes

Hypothèse sur l'origine et l'évolution des plastes basée sur l'analyse de quatre gènes nucléaires de Nozaki *et al.* (2003). Un astérisque indique la présence de plaste primaire et deux, de plaste secondaire.

D'après Nozaki (2005)

2.1.3. Relations entre les trois groupes de Plantae

Malgré la grande quantité de données utilisée, les relations entre les trois groupes d'eucaryotes photosynthétiques primaires n'ont pas pu être résolues avec les analyses basées sur des gènes du plaste, ni avec celles basées sur des gènes nucléaires. Les premières favorisent la position basale des glaucophytes (VB = 78%), tandis que les deuxièmes favorisent la relation glaucophytes + plantes vertes (VB = 74%). Cependant, le support pour cette relation diminue (BV = 40%) quand l'algue rouge *Cyanidioschyzon* est exclue de l'analyse, suggérant que la position basale des algues rouges était provoquée par un artefact d'attraction des longues branches. Selon notre étude, aucune des trois alternatives pour les relations entre les Plantae ne peut être rejetée. Autrement dit, la pensé commune que les glaucophytes sont basales dû à leurs caractères primitifs, comme le peptidoglycane et les carboxysomes, reste encore à confirmer.

2.2. *Mesostigma* est un streptophyte

Les plantes vertes (ou Viridiplantae) se divisent en deux groupes : les streptophytes, qui contiennent les algues charophytes et les plantes terrestres, et les chlorophytes, qui contiennent le restant des algues vertes. L'algue flagellée d'eau douce *Mesostigma viridae* partage des caractères uniques à chacun de ces deux ensembles et, en conséquence, trois placements phylogénétiques pour cette espèce ont été suggérés : avec les streptophytes, avec les chlorophytes et à la base des deux groupes. Au lieu de résoudre la question, les analyses phylogénétiques n'ont fait qu'aviver la controverse : celles basées sur des gènes nucléaires supportaient le placement de *Mesostigma* dans les streptophytes, tandis que celles basées sur des gènes mitochondriaux et chloroplastiques supportaient le placement basal de cette espèce (Lewis et McCourt, 2004). Notre étude basée sur 125 protéines codées dans le noyau de 15 plantes vertes, quatre algues rouges et deux glaucophytes supporte fortement le placement de *Mesostigma* avec les streptophytes. De plus, nous avons montré que les analyses basées sur 50 protéines du plaste et sur 33 protéines mitochondriales sont congruentes avec ce résultat, en plus de trouver des explications aux contradictions observées dans les études précédentes.

Nos conclusions ont été récemment confirmées par Lemieux, Otis et Turmel (2007), le même groupe qui publia en 2000 deux analyses phylogénétiques qui supportaient la position basale de *Mesostigma* : l'une basée sur des gènes mitochondriaux (Turmel, Otis et Lemieux, 2002) et l'autre sur des gènes du plaste (Lemieux, Otis et Turmel, 2000). Cette fois-ci, ce même groupe de chercheurs a réanalysé le jeu de données du plaste en appliquant des méthodes pour extraire le signal phylogénétique, comme le retrait des sites rapides. Tout comme nous, ils concluent au placement de *Mesostigma* chez les streptophytes et aux artefacts de reconstruction phylogénétique pour expliquer leurs résultats précédents (Lemieux, Otis et Turmel, 2007).

Le placement de *Mesostigma* dans les streptophytes est congruent avec le fait qu'elle partage plus d'ESTs avec les plantes terrestres qu'avec le chlorophyte *Chamydomonas reinhardtii* (Simon *et al.*, 2006) et avec la présence d'une famille multigene et de la duplication du gène *gapA/B*, tous les deux spécifiques à *Mesostigma* et aux streptophytes (Nedelcu, Borza et Lee, 2006; Petersen *et al.*, 2006; Simon *et al.*, 2006).

2.3. Les excavés sont-ils monophylétiques?

Grâce au séquençage d'ESTs de cinq jakobides et de deux malawimonadines, nous avons pu réaliser des analyses visant à déterminer la position phylogénétique de ces deux groupes de protistes. Les arbres obtenus supportent fortement la relation entre les jakobides et les Euglenozoa et suggèrent que les malawimonadines sont apparentés à cet ensemble. Ces résultats ont des implications sur l'origine de la mitochondrie et sur une des hypothèses les plus controversées de la phylogénie des eucaryotes : l'hypothèse Excavata. Évidemment, nos conclusions doivent être confirmées par de nouvelles études où sont présents d'autres groupes d'excavés, les haptophytes, les cryptophytes et le seul grand groupe eucaryote non inclut, les Rhizaria.

2.3.1. La relation entre les jakobides et les Euglenozoa : la place des Heterolobosea

Étant donné que nos analyses supportent fortement la relation entre les Euglenozoa et les jakobides, il est particulièrement nécessaire de tester ce résultat en incluant des membres des Heterolobosea, un groupe d'excavés pour lequel de séquences génomiques n'étaient pas disponibles au moment de notre étude. En effet, la relation entre les Euglenozoa et les Heterolobosea est discutée depuis quelque temps à cause du type inhabituel de crêtes mitochondriales de forme discoïde partagé par ces deux groupes (Patterson, 1988). De plus, cette relation est aussi suggérée par plusieurs analyses phylogénétiques avec plus ou moins de support statistique (Baldauf *et al.*, 2000; Cavalier-Smith, 1998; Nikolaev *et al.*, 2004; Simpson, 2003). L'ensemble Euglenozoa + Heterolobosea est connu sous le nom Discristata (Baldauf *et al.*, 2000; Cavalier-Smith, 1998).

Des analyses plus récentes basées sur quelques gènes nucléaires suggèrent, mais avec un support statistique modéré (VB < 85%), que les jakobides pourraient interrompre l'ensemble Discristata et être apparentés soit aux Euglenozoa, soit aux Heterolobosea, ou émerger à la base des deux (Edgcomb *et al.*, 2001; Simpson, Inagaki et Roger, 2006; Simpson et Roger, 2004a; Simpson *et al.*, 2002). La combinaison des analyses précédentes (Baldauf *et al.*, 2000; Cavalier-Smith, 1998; Simpson, Inagaki et Roger, 2006) et les nôtres

suggère que les trois, Heterolobosea, Euglenozoa et jakobides, forment un groupe monophylétique. Maintenant que des données génomiques des Heterolobosea sont disponibles (http://genome.jgi-psf.org/Naegr1/Naegr1.home.html), des études visant à résoudre les relations entre ces trois groupes s'imposent.

En supposant que les jakobides, les Heterolobosea et les Euglenozoa forment un groupe monophylétique, des évidences pour deux des trois branchements possibles entre ces groupes existent. Premièrement, le fait que les jakobides ne possèdent pas de crêtes mitochondriales discoïdes, mais tubulaires –*Histiona* et *Reclinomonas*, ou aplaties – *Jakoba* (O'Kelly, 1993), suggère que les Heterolobosea et les Euglenozoa sont plus proches entre eux que chacun l'est des jakobides. Deuxièmement, la relation Heterolobosea + jakobides, avec les Euglenozoa émergeant à la base, est suggérée par des analyses phylogénétiques (VB = 69-87%) (Simpson, Inagaki et Roger, 2006; Simpson et Roger, 2004a).

Dans le dessein de trouver plus d'indices quant aux relations entre ces trois groupes, nous avons vérifié si les insertions communes aux jakobides et aux Euglenozoa trouvées dans les protéines *RPL22* et *24A* étaient présentes ou absentes chez l'Heterolobosea *Naegleria gruberi*, pour lequel les séquences génomiques sont maintenant disponibles. Cette espèce présente au même endroit que les jakobides et que les Euglenozoa une insertion dans la protéine *RPL24A*, supportant davantage la monophylie des ces trois groupes, mais n'apportant aucune information au sujet de leurs relations de parenté. Curieusement, la protéine *RPL22* de *Naegleria* ne contient pas l'insertion commune trouvée dans ses homologues chez jakobides et chez Euglenozoa, suggérant une troisième alternative, soit la position basale des Heterolobosea. En résumé, les indices trouvés jusqu'à date sont ambigus et ne permettent pas d'émettre de conclusion concernant les relations entre les Euglenozoa, les Heterolobosea et les jakobides.

2.3.2. La position phylogénétique des malawimonadines

Les malawimonadines possèdent, tout comme les Euglenozoa et les Heterolobosea, des crêtes mitochondriales discoïdes, ce qui a mené à penser que ces trois groupes étaient monophylétiques (O'Kelly, 1993). Cependant, cette relation n'est jamais trouvée dans les

phylogénies moléculaires. Curieusement, jusqu'à date, aucune relation phylogénétique entre les malawimonadines et un autre groupe d'excavés a été supportée (Archibald, O'Kelly et Doolittle, 2002; Simpson, Inagaki et Roger, 2006; Simpson et Roger, 2004a). Nos analyses sont donc les premières à suggérer une position phylogénétique pour les malawimonadines, soit ensemble avec les deux autres groupes d'excavés présents, les jakobides et les Euglenozoa.

2.3.3. Le future des excavés

Les trois groupes d'excavés inclus dans nos analyses (jakobides, Euglenozoa et malawimonadines) semblent former un groupe monophylétique. De plus, tel que mentionné dans la section 2.3.1, il est fort possible qu'un quatrième groupe, les Heterolobosea, soit inclus dans cet assemblage. Il existe deux autres groupes d'excavés pour lesquels des données génomiques sont aussi disponibles, soit le parabasalien *Trichomonas* et le diplomonadine *Giardia*. Ces espèces ne sont pas incluses dans les arbres présentés dans le chapitre V à cause de leur fort taux d'évolution, mais nous avons quand même fait des analyses additionnelles les incluant. La Figure 40 montre l'arbre basé sur le même jeu de données que celui présenté dans la Figure 3 du Chapitre V auquel des séquences de *Giardia* et de *Trichomonas* ont été ajoutées. Selon cette analyse, la monophylie des six groupes d'excavés présents est supportée avec VB = 73%. Évidemment, étant donné les longues branches de *Giardia* et de *Trichomonas*, ce résultat doit être pris avec précaution. Cependant, il faut prendre en compte que, l'ensemble Euglenozoa + jakobides + malawimonadines n'évoluant pas à un taux spécialement rapide, l'attraction entre ce dernier et l'ensemble et *Giardia* + *Trichomonas* pourrait être due à du vrai signal phylogénétique. Si ceci se confirmait, la monophylie d'au moins six groupes d'excavés, parmi les dix identifiés jusqu'à date, pourrait être conclue.

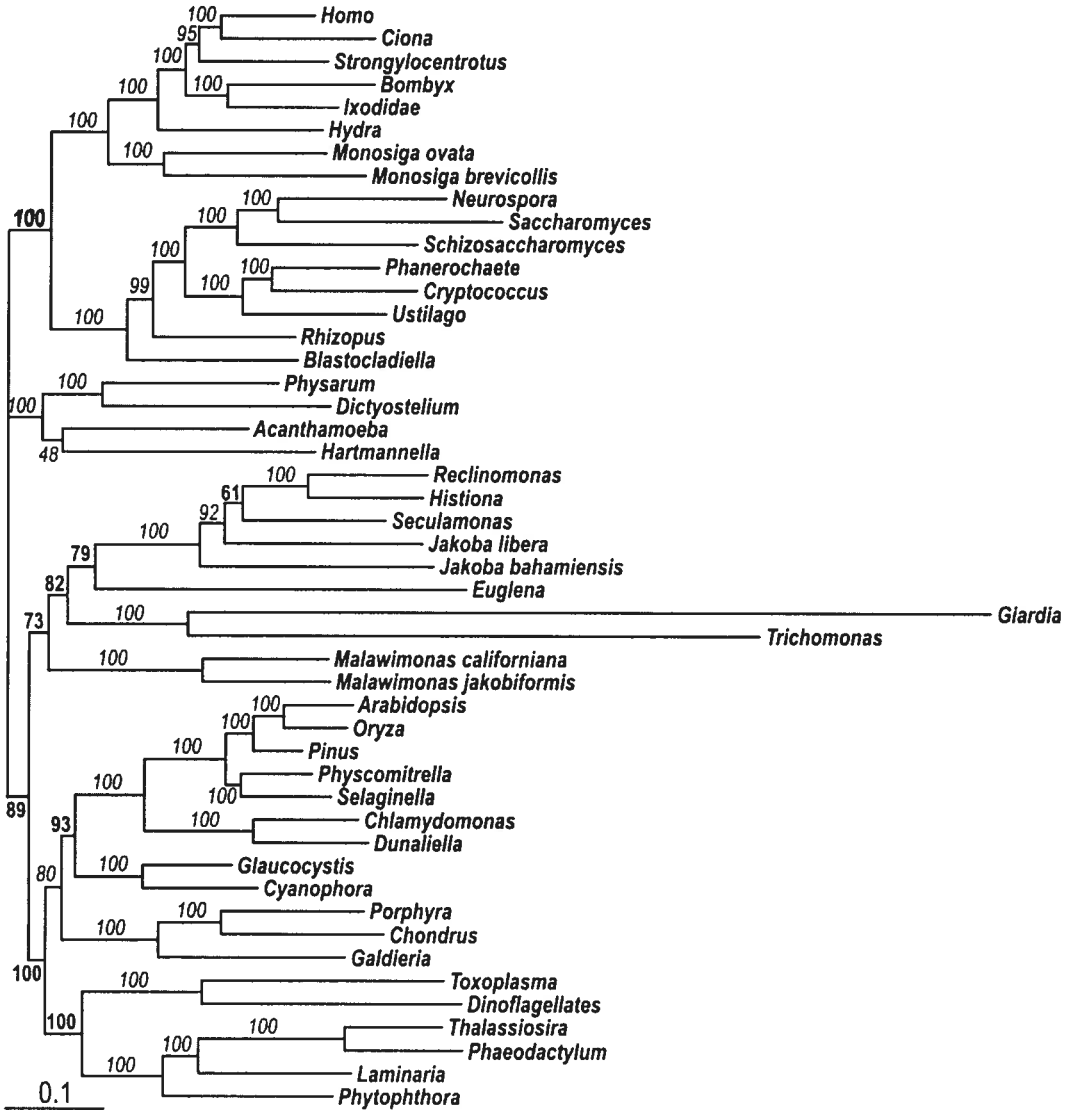

Figure 40 : Phylogénie incluant les parabasaliens et les diplomonadines

Arbre de vraisemblance maximale obtenu à partir de l'analyse de la concaténation de 170 protéines codées dans le noyau. Les nombres en italique représentent des valeurs de bootstrap obtenues par l'analyse de 100 pseudo-répliqués avec TreeFinder (Jobb, von Haeseler et Strimmer, 2004) et le modèle WAG+F+Γ. Pour l'analyse exhaustive, un jeu de 150 topologies fut créé en retenant les 10 meilleures de l'analyse sans *Giardia* + *Trichomonas* et en ajoutant ce groupe partout sauf à l'intérieur des straménopiles, des alvéolés, des Plantae, des champignons, des animaux, des Amoebozoa, des malawimonadines et des jakobides. Les nombres en gras représentent des valeurs de bootstrap obtenues par l'analyse de 10 000 pseudo-répliqués RELL.

Si nos suppositions sont correctes, les dix groupes dits excavés pourraient être monophylétiques, confirmant ainsi l'hypothèse Excavata. Comment expliquer alors le manque de résolution obtenue dans les analyses phylogénétiques visant à résoudre les relations parmi les membres du groupe? Pour quoi la monophylie des excavés serait-elle si difficile à confirmer? Nos résultats concernant la position phylogénétique des malawimonadines et des jakobides + Euglenozoa par rapport au reste des eucaryotes ont en partie la réponse à ces questions. Premièrement, la présence de la longe branche des Euglenozoa est suffisante pour introduire un artéfact d'inférence phylogénétique qui fausse les résultats (voir Figure S1 du Chapitre V). Pourtant, il existe d'autres groupes d'excavés dont le taux d'évolution est bien plus haut que celui des Euglenozoa et des artéfacts encore plus sévères sont attendus. Deuxièmement, même en absence des longues branches (Figure 3 du Chapitre V), la relation entre le groupe Euglenozoa + jakobides et les malawimonadines est difficile à retrouver (environ 35 000 positions d'acides aminés sont nécessaires pour que la VB atteigne 90%). Ceci indique que le signal phylogénétique pour cette relation est très faible, comme c'est probablement le cas pour d'autres relations entre les excavés.

3. Implications méthodologiques sur l'inférence de la phylogénie des eucaryotes

3.1. Construction de jeux de données phylogénomiques

Une des difficultés majeures de l'assemblage de jeux de données pour inférer la phylogénie des organismes est la paralogie et la xénologie. Les deux impliquent généralement plusieurs séquences par espèce pour un même gène, certains d'entre elles ne reflétant pas la vraie phylogénie. Comment choisir, parmi les multiples séquences, celles qui sont orthologues? Étant donné les limitations des méthodes automatisées de détection d'orthologues (voir section 1.1.2 de l'introduction), des approches alternatives sont nécessaires. La comparaison de séquences primaires et la sélection de celles qui ont le plus bas taux évolutif n'est pas un critère valide (Pearson et Sierk, 2005). Tout de même, cette méthode est couramment utilisée (p. ex., Simpson, Inagaki et Roger, 2006). La Figure 41 illustre un cas dans lequel le choix de la séquence la moins divergente mène à des résultats

erronés. Il est d'ailleurs possible que les résultats de Simpson *et al.* (2006) concernant les relations entre les excavés soient biaisés dû à une mauvaise sélection d'orthologues, d'autant plus que les gènes utilisés ont plusieurs copies paralogues pour certains groupes d'eucaryotes.

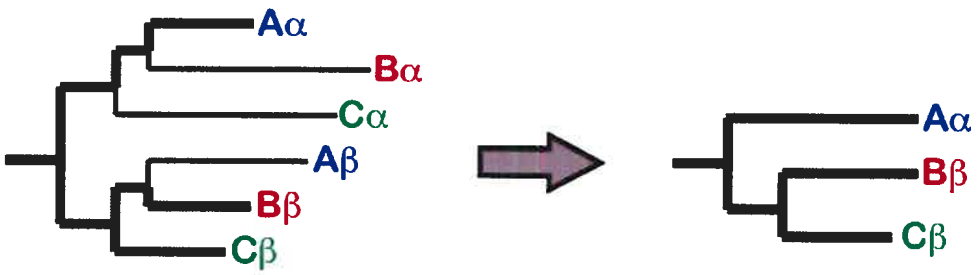

Figure 41 : Sélection des séquences orthologues selon la distance évolutive
Représentation de l'histoire évolutive de deux groupes de paralogues (α et β) pour les espèces A, B et C (à gauche). Dans ce cas, le choix de la copie la moins divergente pour chaque espèce donnerait lieu à un arbre erroné (à droite).

Poussés par la gravité de l'inclusion de séquences paralogues dans les études visant à résoudre l'arbre des eucaryotes, nous avons réalisé des analyses phylogénétiques préliminaires pour chaque gène et sélectionné les séquences manuellement. Nos critères pour cette sélection ont été les suivants : (i) quand une duplication est basale à tous les eucaryotes, les deux copies résultantes sont traitées indépendamment (p. ex., les chaperonnes CCT α, β, δ, ε, γ, η, θ et ζ); (ii) quand une duplication est unique à un groupe déterminé (p. ex., les plantes vertes), le groupe de séquences de la copie la moins divergente et avec le plus d'espèces est choisie; et (iii) quand l'identification d' orthologues n'est pas possible, le gène est exclu de l'analyse. Grâce à cette approche fastidieuse, mais efficace, nous avons réduit considérablement la présence de séquences paralogues dans nos alignements.

Un autre point crucial lors de la construction de jeux de données phylogénomiques est la présence de données manquantes, qui ne peut pas être évitée lorsqu'un grand nombre de gènes de beaucoup d'espèces est utilisé. Les données manquantes sont souvent

considérées un obstacle dans la reconstruction phylogénétique (Wiens, 1998), mais des études récentes basées sur des données simulées et réelles ont montré le contraire (Brinkmann *et al.*, 2005; Philippe *et al.*, 2004; Wiens, 2005). Particulièrement, il a été montré que (i) l'inclusion d'une espèce incomplète pour briser une longue branche réduit l'artefact d'attraction des longues branches (Wiens, 2005) et (ii) l'exclusion des positions rapides de quelques espèces (allant jusqu'à 90% de données manquantes pour certaines) améliore la qualité de l'inférence phylogénétique (Brinkmann *et al.*, 2005). Les analyses présentés dans le chapitre III supportent davantage ces conclusions car l'utilisation d'une espèce incomplète (avec 36% de données manquantes; *Porphyra*) pour représenter les algues rouges donne des résultats plus exactes que l'utilisation d'une espèce pour laquelle le génome complet est disponible (*Cyanidioschyzon*).

Bien que le « netoyage » des alignements, incluant le choix des espèces, des gènes et des séquences, ait été fait manuellement en partie, nous nous sommes aidés du logiciel SCaFoS (Roure, Rodríguez-Ezpeleta et Philippe, 2007), qui a facilité plusieurs des étapes du processus et qui permet aussi de les reproduire automatiquement. Étant donné que SCaFos a été développé au même temps que la construction des jeux de données présentés ici, ce logiciel est très complet et inclut des solutions efficaces à de nombreux problèmes rencontrés lors de la création de jeux de données phylogénomiques. Enfin, SCaFoS est un logiciel qui s'avère spécialement utile pour la création de futurs jeux de données à échelle génomique.

3.2. Nécessité d'une approche phylogénomique pour contrer l'erreur stochastique

De façon générale, nos analyses montrent que le signal phylogénétique des relations entre les groupes eucaryotes est faible. Ceci implique qu'une grande quantité de positions sont nécessaires pour contrer l'erreur stochastique. Évidemment, le nombre de positions nécessaires pour trouver un support statistiquement significatif n'est pas le même pour tous les groupes eucaryotes. En effet, ce nombre est moins élevé si le nombre de substitutions communes au groupe (longueur de la branche interne) est grand par rapport au nombre de substitutions accumulées par les lignées individuelles (longueur des

branches externes), et plus élevé, dans le cas contraire. Autrement dit, le nombre de positions requis pour trouver un support significatif pour un groupe donné corrèle avec le ratio branche externe / branche interne du groupe en question (Figure 42). Il faut quand même noter que d'autres facteurs comme le nombre d'espèces, les biais systématiques et la quantité de données manquantes peuvent également avoir un effet sur le nombre de positions nécessaires pour retrouver un groupe donné.

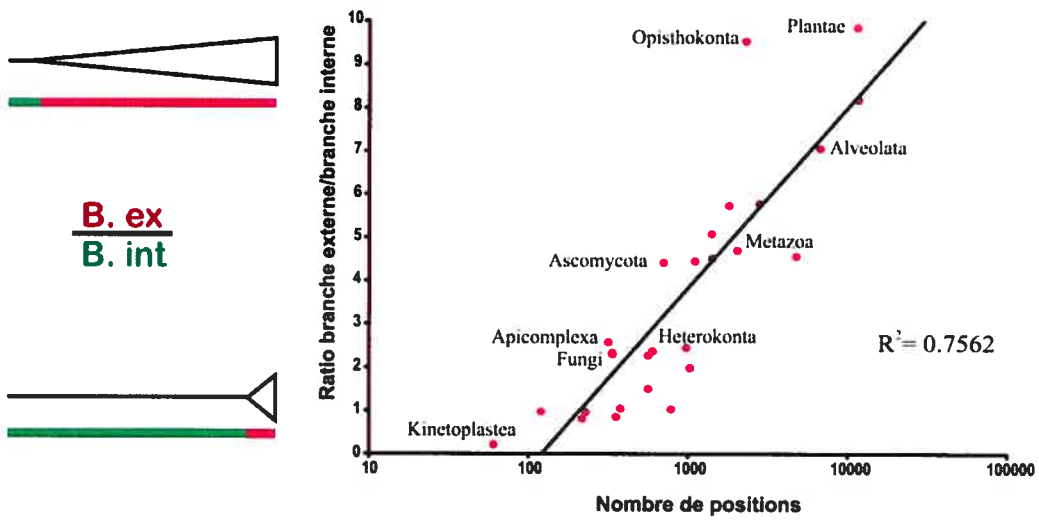

Figure 42 : Ratio branche externe / branche interne et nombre de positions
Pour chaque groupe monophylétique, le ratio entre la moyenne de la longueur des branches externes et de la branche interne est calculé et comparé au nombre de positions requis pour retrouver ce groupe avec une VB = 95%. Les points ont été ajustés par régression linéaire avec la formule VB = 100 $(1\text{-}e^{\text{-}bx})$ (Lecointre *et al.*, 1994), où x est le nombre de positions requis pour que la VB = 95%. Seulement les points correspondant aux groupes les plus connus sont nommés[12].

Ces observations expliquent que les analyses basées sur des gènes uniques ne résolvent pas les relations entre les grands groupes eucaryotes et que la combinaison d'une multitude de gènes soit nécessaire. Pourquoi la concaténation de plusieurs gènes apporte-elle un support significatif pour une topologie alors que celle-ci n'est pas supportée

[12] Ces analyses se basent sur un jeux de données de 141 protéines très similaire à celui présenté dans le Chapitre II

significativement par chacun des gènes analysés individuellement? La réponse à cette question est illustrée dans la Figure 43. Dans ce graphe, si le point rouge est plus haut que la barre horizontale, le support pour la topologie de la concaténation est supérieur au support moyen de l'ensemble des meilleures topologies pour au moins un des gènes. En plus, si le point rouge est plus haut que la ligne verticale grise, cette différence est significative. Selon la distribution des points rouges dans la Figure 43, on peut conclure que, pour la majorité des gènes, la topologie obtenue avec la concaténation est mieux supportée que la moyenne du reste des topologies, mais que ce support est rarement significatif. In existe donc un support faible mais cohérent dans la majorité des gènes pour la topologie obtenue avec la concaténation.

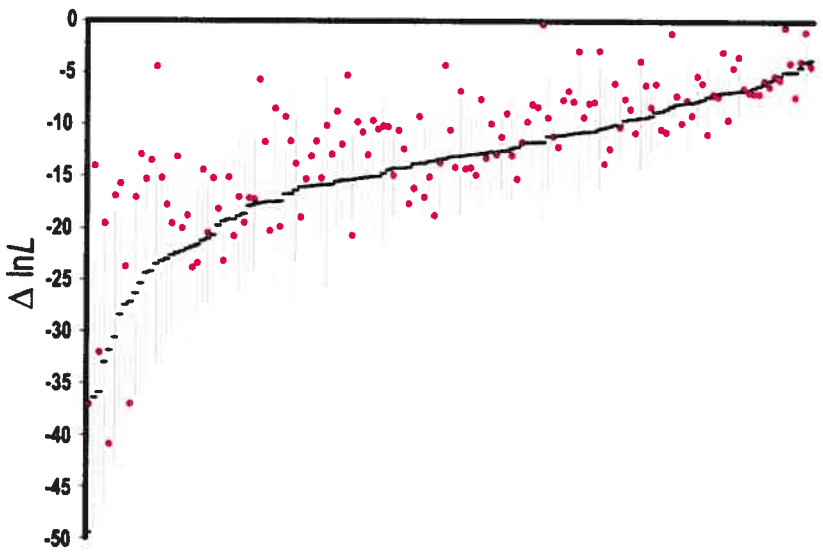

Figure 43 : Support faible mais cohérent pour la meilleure topologie
Pour chaque gène (axe des X), la moyenne (trait horizontal noir) et l'écart type (trait vertical gris) des différences de vraisemblance (Δ ln*L*) entre les 126 topologies supportées par au moins un des gènes et la meilleure topologie pour ce gène en particulier sont calculés. Ces valeurs sont comparées avec la différence de vraisemblance entre la meilleure topologie globale et la meilleure topologie pour chaque gène (points rouges). L'axe des X représente chacun des 141 gènes, rangés selon leur pouvoir discriminatif (plus élevé à gauche)[13].

[13] Voir note 12

3.3. La phylogénomique augmente le risque d'erreur systématique

L'erreur stochastique étant surmontée par l'utilisation de jeux de données génomiques, le problème majeur des futures analyses visant à inférer la phylogénie des eucaryotes est maintenant l'erreur systématique. Plusieurs études ont illustré la gravité de ce type d'erreur dans les analyses phylogénomiques. Par exemple, les analyses basées sur la concaténation de 106 gènes nucléaires de 8 et de 14 levures de Phillips, Delsuc et Penny (2004) et de Jeffroy *et al.* (2006) montrent une forte attraction artificielle entre les espèces qui partagent une composition en nucléotides similaire, conduisant à des arbres significativement supportés mais incorrects. Similairement, l'étude de Brinkmann *et al.* (2005), basée sur 133 protéines de 37 eucaryotes et 6 archées, supporte la position basale du microsporidien *Encephalitozoon* (VB = 96%), alors qu'il est bien connu que cette position est erronée et que cette espèce est un champignon. Ces deux exemples représentent des cas extrêmes. Premièrement, dans le jeu de données des levures, la solution erronée est supportée seulement si des méthodes non probabilistes sont utilisées ou si des nucléotides (où le biais compositionnel est plus évident) au lieu des acides aminés sont utilisés (Jeffroy *et al.*, 2006; Phillips, Delsuc et Penny, 2004). Deuxièmement, dans le jeu de données des eucaryotes, l'hétérogénéité du taux d'évolution entre les espèces est énorme, la branche d'*Encephalitozoon* étant de loin la plus longue (Brinkmann *et al.*, 2005).

Pour construire les phylogénies présentées dans cette thèse, des méthodes probabilistes avec les meilleurs modèles d'inférence phylogénétique disponibles à ce jour ont été utilisées. De plus, les espèces incluses ont des taux d'évolution modérés et les espèces avec des taux extrêmes, comme *Giardia* et *Trichomonas*, ont été généralement exclues. Cependant, des résultats statistiquement supportés (VB = 100%) mais erronés ont quand même été obtenus. Ceci s'explique par le faible signal phylogénétique contenu dans les séquences pour certaines relations, requérant, même en absence de longues branches, plus de 20 000 positions d'acides aminés pour être significativement supportées. En d'autres termes, même si l'erreur systématique induite par la présence d'espèces comme *Cyanidioschyzon* ou les Euglenozoa est faible, elle est suffisante pour contrer l'aussi faible signal phylogénétique. Une élément de comparaison est le taux d'évolution du champignon

Saccharomyces, qui est plus élevé que celui de *Cyanidioschyzon* (Figure 1 du Chapitre III). Comme le signal phylogénétique pour la monophylie des champignons est très fort (illustré par la longue branche interne du groupe), celle-ci n'est pas disturbée en présence de *Saccharomyces*, comparativement à la monophylie des Plantae en présence de *Cyanidioschyzon*. En effet, pour disturber la monophylie des champignons une erreur systématique plus forte (p. ex., celle induite par la présence d'*Encephalitozoon*) est nécessaire (Brinkmann *et al.*, 2005).

Les valeurs de bootstrap ne peuvent pas être utilisées pour détecter les erreurs systématiques lors des analyse phylogénétiques, car le fait qu'une relation soit supportée avec VB = 100% ne signifie pas qu'elle est exacte. Des méthodes alternatives sont donc nécessaires. Par exemple, une approche efficace utilisée dans nos études est de tester la robustesse des résultats à l'échantillonnage taxonomique, au retrait des sites rapides, au codage d'acides aminés et à l'utilisation de différents modèles d'évolution. En plus de leur pouvoir pour détecter les erreurs systématiques, ces approches sont aussi utiles pour les surmonter. Par exemple, l'utilisation de *Porphyra* au lieu de *Cyanidioschyzon* pour représenter les algues rouges est suffisante pour augmenter la VB de la monophylie des Plantae de 0% à 99%. Les méthodes pour surmonter les erreurs systématiques testées ici ne sont pas toutes pareillement efficaces pour tous les jeux de données. En effet, chaque analyse est différemment affectée par les artefacts d'inférence phylogénétique, qui dépendent du type et de la magnitude des biais accumulés pendant le processus évolutif ayant donné lieu aux séquences. De plus, les jeux de données sont normalement affectés par plusieurs types de biais en même temps.

Selon les résultats présentés dans le Chapitre III, le retrait des sites rapides semble être la méthode la plus efficace parmi celles qui ont été testées. En effet, les sites rapides sont les plus saturés et donc ceux qui accumulent le plus de biais. Différentes approches pour identifier les sites rapides ont été utilisées auparavant. La première, nommée SF (Brinkmann et Philippe, 1999), consiste à trier les positions selon le nombre de substitutions à l'intérieur des groupes monophylétiques prédéfinis et à construire de matrices contenant les positions avec au plus 0, 1, 2, etc. substitutions. Le désavantage de cette méthode est qu'elle requiert l'identification *a priori* de groupes monophylétiques et que ceux ci doivent contenir au moins trois espèces pour que le comptage du nombre de substitutions soit efficace. De plus, étant la manière de trier les sites si draconienne , les

matrices avec les positions les plus lentes n'ont pas assez de signal phylogénétique et les plus rapides contiennent déjà trop de signal non-phylogénétique. La deuxième approche fut définie par Brinkmann *et al.* (2005) et consiste à éliminer les séquences évoluant le plus rapidement pour une espèce donnée. Comme le tri des séquences se base sur la distance entre les deux groupes qui s'attirent, la connaissance *a priori* de ceux-ci est nécessaire. Cette approche est donc seulement utile dans les cas où les espèces problématiques sont connues d'avance. Finalement, la méthode utilisée par Ruiz-Trillo *et al.* (1999) est similaire à celle que nous avons utilisée. Une des différences réside dans le fait que, dans leur cas, l'estimation des taux se fait selon la catégorie de la distribution gamma discrète à laquelle les sites appartiennent, ce qui donne une estimation moins précise que le calcul d'un taux pour chaque site. La deuxième différence (et la plus importante) est que Ruiz-Trillo *et al.* (1999) n'utilisent pas de taux calculés sur des topologies « neutres » ou sur la moyenne d'un ensemble de topologies. En fait, ils utilisent une seule topologie qui peut être correcte ou non, tout dépendant de la méthode de reconstruction utilisée. Comme c'est montré dans notre étude, le résultat obtenu suite au retrait des sites rapides dépend très fortement de la topologie utilisée pour calculer les taux. Notre méthode (i) ne requiert pas un nombre minimale d'espèces par groupe monophylétique, (ii) n'implique pas la définition *a priori* des espèces problématiques et (iii) n'est pas biaisée par la topologie de départ.

Enfin, le retrait des sites n'est pas la panacée. D'ailleurs, cette méthode n'a pas donnée de résultats satisfaisants dans certains cas. Pour la majorité d'entre eux, les raisons nous sont inconnues : c'est possible qu'il reste trop peu de signal phylogénétique même après le retrait des sites rapides ou que d'autres biais non surmontables par cette méthode existent dans le jeux de données. Ceci semble être le cas pour les analyses de protéines du plaste visant à résoudre la position de *Mesostigma*. Ici, la violation du modèle dominante est l'hétérotachie crée de manière artificielle en concaténant de gènes avec de longueurs de branches relatives différentes. Cependant, le cas de *Mesostigma* demeure une énigme car il n'est toujours pas compris pourquoi la présence de cette espèce ou celle de *Chlorokybus* donnent des résultats si différents. Ce jeu de données représente donc un exemple clair de la complexité de l'évolution réelle des séquences et de la difficulté des modèles actuellement disponibles pour la capturer.

CONCLUSION

Nos analyses phylogénétiques basées sur un grand nombre de gènes ont permis de résoudre certaines questions longuement débattues comme la monophylie des eucaryotes photosynthétiques primaires, le placement de *Mesostigma* à l'intérieur des plantes vertes et la position des jakobides et des malawimonadines dans l'arbre des eucaryotes. De plus, des études approfondies de chacun des jeux de données utilisés ont permis de comprendre les difficultés rencontrées par d'autres analyses visant à éclaircir ces questions.

En général, la résolution de l'arbre des eucaryotes requiert l'analyse de jeux de données à échelle génomique afin de contrer le faible signal phylogénétique contenu dans les séquences. Cependant, utilisée aveuglément, cette approche peut s'avérer hasardeuse. En effet, l'obtention d'un arbre parfaitement résolu n' implique pas que celui-ci représente correctement l'évolution des séquences. Tel que montré par nos études, les artefacts d'inférence phylogénétique sont aggravés lors de l'utilisation d'un grand nombre de positions, et peuvent ainsi annuler le signal phylogénétique.

La solution ultime pour contrer les artefacts d'inférence phylogénétique est le développement de modèles d'évolution qui interprètent correctement les substitutions multiples. Toutefois, cette tâche n'est pas facile et des approches alternatives sont nécessaires. Ici, nous avons montré que l'élimination des données les plus saturées (espèces ou positions) est une méthode efficace pour augmenter le signal phylogénétique et donc le support pour la solution correcte. Puisqu'elle requiert l'élimination d'un nombre considérable de données, cette approche est uniquement possible avec des jeux de données phylogénomiques.

En dernier mot, les études présentées ici ont des implications sur la manière dont la résolution des branches de l'arbre des eucaryotes devra être adressée.

BIBLIOGRAPHIE

Abascal, F., R. Zardoya, et D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-5.

Adachi, J., et M. Hasegawa. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J Mol Evol 40:622-8.

Adachi, J., et M. Hasegawa. 1996a. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42:459-468.

Adachi, J., et M. Hasegawa. 1996b. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput Sci Monogr 28:1-150.

Adachi, J., P. J. Waddell, W. Martin, et M. Hasegawa. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol 50:348-58.

Adl, S. M., A. G. Simpson, M. A. Farmer, R. A. Andersen, O. R. Anderson, J. R. Barta, S. S. Bowser, G. Brugerolle, R. A. Fensome, S. Fredericq, *et al.* 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 52:399-451.

Adoutte, A., A. Germot, H. Le Guyader, et H. Philippe. 1996. Que savons-nous de l'histoire évolutive des Eucaryotes? De la diversification des protistes à la radiation des multicellulaires. Méd Sci 12:I-XVII.

Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, et J. A. Lake. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489-93.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267-281 *dans* Proceedings 2nd International Symposium on Information Theory (Petrov, et Csaki, eds.). Akademia Kiado, Budapest.

Allen, B. L., et M. Steel. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. Annals of combinatorics 5:1-15.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, et D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol 215:403-10.

Archibald, J. M. 2005. Jumping genes and shrinking genomes--probing the evolution of eukaryotic photosynthesis with genomics. IUBMB Life 57:539-47.

Archibald, J. M., et P. J. Keeling. 2002. Recycled plastids: a 'green movement' in eukaryotic evolution. Trends Genet 18:577-84.

Archibald, J. M., et P. J. Keeling. 2004. The evolutionary history of plastids: a molecular phylogenetic perspective. Pages 55-74 *dans* Organelles, genomes and eukaryotic phylogeny (R. P. Hirt, et D. S. Horner, eds.). CRC Press, New York.

Archibald, J. M., D. Longet, J. Pawlowski, et P. J. Keeling. 2003. A novel polyubiquitin structure in Cercozoa and Foraminifera: evidence for a new eukaryotic supergroup. Mol Biol Evol 20:62-6.

Archibald, J. M., C. J. O'Kelly, et W. F. Doolittle. 2002. The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. Mol Biol Evol 19:422-31.

Baele, G., J. Raes, Y. Van de Peer, et S. Vansteelandt. 2006. An improved statistical method for detecting heterotachy in nucleotide sequences. Mol Biol Evol 23:1397-405.

Baldauf, S. L. 1999. A Search for the Origins of Animals and Fungi: Comparing and Combining Molecular Data. Am Nat 154:S178-S188.

Baldauf, S. L., et W. F. Doolittle. 1997. Origin and evolution of the slime molds (Mycetozoa). Proc Natl Acad Sci U S A 94:12007-12.

Baldauf, S. L., et J. D. Palmer. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci U S A 90:11558-62.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, et W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972-7.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, *et al.* 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A 99:1414-9.

Bapteste, E., et H. Philippe. 2002. The potential value of indels as phylogenetic markers: position of trichomonads as a case study. Mol Biol Evol 19:972-7.

Bhattacharya, D., T. Helmchen, C. Bibeau, et M. Melkonian. 1995. Comparisons of nuclear-encoded small-subunit ribosomal RNAs reveal the evolutionary position of the Glaucocystophyta. Mol Biol Evol 12:415-20.

Bhattacharya, D., et L. Medlin. 1995. The phylogeny of plastids: a review based on comparisons of small-subunit ribosomal RNA coding regions. Journal of Phycology 31:489–498.

Bhattacharya, D., et K. Weber. 1997. The actin gene of the glaucocystophyte Cyanophora paradoxa: analysis of the coding region and introns, and an actin phylogeny of eukaryotes. Curr Genet 31:439-46.

Bower, F. O. 1908. The origin of a land flora. Macmillan, London.

Boxma, B., R. M. de Graaf, G. W. van der Staay, T. A. van Alen, G. Ricard, T. Gabaldon, A. H. van Hoek, S. Y. Moon-van der Staay, W. J. Koopman, J. J. van Hellemond, *et al.* 2005. An anaerobic mitochondrion that produces hydrogen. Nature 434:74-9.

Brinkmann, H., et H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817-25.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, et H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743-57.

Brochier, C., et H. Philippe. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature 417:244.

Bui, E. T., P. J. Bradley, et P. J. Johnson. 1996. A common evolutionary origin for mitochondria and hydrogenosomes. Proc Natl Acad Sci U S A 93:9651-6.

Burger, G., M. W. Gray, et B. F. Lang. 2003. Mitochondrial genomes: anything goes. Trends Genet 19:709-16.

Burger, G., D. Saint-Louis, M. W. Gray, et B. F. Lang. 1999. Complete sequence of the mitochondrial DNA of the red alga *Porphyra purpurea*. Cyanobacterial introns and shared ancestry of red and green algae. Plant Cell 11:1675-94.

Burki, F., et J. Pawlowski. 2006. Monophyly of Rhizaria and Multigene Phylogeny of Unicellular Bikonts. Mol Biol Evol.

Cai, X., A. L. Fuller, L. R. McDougald, et G. Zhu. 2003. Apicoplast genome of the coccidian *Eimeria tenella*. Gene 321:39-46.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540-52.

Cavalier-Smith, T. 1981. Eukaryote kingdoms: seven or nine? Biosystems 14:461-81.

Cavalier-Smith, T. 1983. A 6-kingdom classification and a unified phylogeny. Pages 1027-1034 *dans* Endocyobiology II: intracellular space as an oligogenetic ecosystem (W. Schwemmler, et E. A. Schenk, eds.). De Gruyter, Berlin.

Cavalier-Smith, T. 1987a. The origin of fungi and pseudofungi. Pages 339-353 *dans* Evolutionary biology of the fungi (A. D. M. Rayer, C. M. Brasier, et D. Moore, eds.). Cambridge University Press, Cambridge.

Cavalier-Smith, T. 1987b. The simultaneous symbiotic origin of mitochondria, chloroplasts and microbodies. Ann NY Acad Sci 503:55-71.

Cavalier-Smith, T. 1992. Percolozoa and the symbiotic origin of the metakaryote cell. Pages 399-406 *dans* Endocytobiology V (S. Sato, M. Ishida, et H. Ishikawa, eds.). Tübingen University Press, Tübingen.

Cavalier-Smith, T. 1993. Kingdom Protozoa and its 18 phyla. Microbiol Rev 57:953-94.

Cavalier-Smith, T. 1998. A revised six-kingdom system of life. Biol Rev Camb Philos Soc 73:203-66.

Cavalier-Smith, T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. Int J Syst Evol Microbiol 52:297-354.

Cavalier-Smith, T. 2003. The excavate protozoan phyla Metamonada Grasse emend. (Anaeromonadea, Parabasalia, Carpediemonas, Eopharyngia) and Loukozoa emend. (Jakobea, Malawimonas): their evolutionary affinities and new higher taxa. Int J Syst Evol Microbiol 53:1741-58.

Cavalier-Smith, T. 2004. Chromalveolate diversity and cell megaevolution: interplay of membranes, genomes and cytoskeleton. Pages 75-108 *dans* Organelles, genomes and eukaryotic phylogeny (R. P. Hirt, et D. S. Horner, eds.). CRC Press, New York.

Cavalier-Smith, T., et E. E. Chao. 1995. The Opalozoan Apusomonas is Related to the Common Ancestor of Animals, Fungi, and Choanoflagellates. Proc Biol Sci 261:1-6.

Cermakian, N., T. M. Ikeda, R. Cedergren, et M. W. Gray. 1996. Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage. Nucleic Acids Res 24:648-54.

Christensen, T. 1989. The Chromophyta, pas and present. Pages 1-12 *dans* The Chromophyte algae: problems and perspectives (J. C. Green, B. S. C. Leadbeater, et W. C. Diver, eds.). Oxford University Press.

Clark, C. G., et A. J. Roger. 1995. Direct evidence for secondary loss of mitochondria in Entamoeba histolytica. Proc Natl Acad Sci U S A 92:6518-21.

Conant, G. C., et P. O. Lewis. 2001. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. Mol Biol Evol 18:1024-33.

Corliss, J. O. 1984. The kingdom Protista and its 45 phyla. BioSystems 17:87-126.

Cunningham, C. W., H. Zhu, et D. M. Hillis. 1998. Best-fit maximum likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution 52.

Dacks, J. B., J. D. Silberman, A. G. Simpson, S. Moriya, T. Kudo, M. Ohkuma, et R. J. Redfield. 2001. Oxymonads are closely related to the excavate taxon Trimastix. Mol Biol Evol 18:1034-44.

Delsuc, F., H. Brinkmann, et H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361-75.

Delsuc, F., M. J. Phillips, et D. Penny. 2003. Comment on "Hexapod origins: monophyletic or paraphyletic?" Science 301:1482; author reply 1482.

Delwiche, C. F. 1999. Tracing the thread of plastid diversity through the tapestry of life. Am Nat 154:164-177.

Delwiche, C. F., M. Kuhsel, et J. D. Palmer. 1995. Phylogenetic analysis of *tufA* sequences indicates a cyanobacterial origin of all plastids. Mol Phylogenet Evol 4:110-28.

Delwiche, C. F., et J. D. Palmer. 1996. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. Mol Biol Evol 13:873-82.

Doolittle, W. F. 1998. A paradigm gets shifty. Nature 392:15-6.

Douglas, S. E., et C. A. Murphy. 1994. Structural, transcriptional and phylogenetic analyses of the atpB gene cluster from the plastid of Cryptomonas F (Cryptophyceae). Journal of Phycology 30:329-340.

Driskell, A. C., C. Ane, J. G. Burleigh, M. M. McMahon, C. O'Meara B, et M. J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172-4.

Durnford, D. G., J. A. Deane, S. Tan, G. I. McFadden, E. Gantt, et B. R. Green. 1999. A phylogenetic assessment of the eukaryotic light-harvesting antenna proteins, with implications for plastid evolution. J Mol Evol 48:59-68.

Eck, R. V., et M. O. Dayhoff. 1966. Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Springs, MD.

Edgcomb, V. P., A. J. Roger, A. G. Simpson, D. T. Kysela, et M. L. Sogin. 2001. Evolutionary relationships among "jakobid" flagellates as indicated by alpha- and beta-tubulin phylogenies. Mol Biol Evol 18:514-22.

Edwards, A. W. F., et L. L. Cavalli-Sforza. 1963. The reconstruction of evolution. Annals of Human Genetics 27:105-106.

Edwards, A. W. F., et L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. Pages 67-76 *dans* Phenetic and Phylogenetic classification (V. H. H. a. J. McNeill, ed.) Systematics Association Publ. No. 6, London.

Eisen, J. A., et C. M. Fraser. 2003. Phylogenomics: intersection of evolution and genomics. Science 300:1706-7.

Embley, T. M., et R. P. Hirt. 1998. Early branching eukaryotes? Curr Opin Genet Dev 8:624-9.

Embley, T. M., et W. Martin. 2006. Eukaryotic evolution, changes and challenges. Nature 440:623-30.

Embley, T. M., M. van der Giezen, D. S. Horner, P. L. Dyal, et P. Foster. 2003. Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. Philos Trans R Soc Lond B Biol Sci 358:191-201; discussion 201-2.

Fast, N. M., L. Xue, S. Bingham, et P. J. Keeling. 2002. Re-examining alveolate evolution using multiple protein molecular phylogenies. J Eukaryot Microbiol 49:30-7.

Felsenstein, J. 1978a. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401-410.

Felsenstein, J. 1978b. The number of evolutionary trees. Syst Zool 27:27-33.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17:368-376.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:738-791.

Felsenstein, J. 2001. PHYLIP (PHYLogeny Inference Package). Department of Genome Sciences, University of Washington.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Fisher, R. A. 1912. On an absolute criterion for fitting frequency curves. Messenger of Mathematics 41:155-160.

Fisher, R. A. 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1:3-32.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London 222:309-368.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Syst Zool 19:99-113.

Fitch, W. M. 1971a. The nonidentity of invariable positions in the cytochromes c of different species. Biochem Genet 5:231–241.

Fitch, W. M. 1971b. Toward defining the course of evolution: Minimum change for a specified tree topology. Syst Zool 20:406-416.

Fitch, W. M. 1977. On the problem of discovering the most parsimonious tree. Am Nat 111:223-257.

Fitch, W. M. 2000. Homology a personal view on some of the problems. Trends Genet 16:227-31.

Fitch, W. M., et E. Margoliash. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem Genet 1:65-71.

Fitch, W. M., et E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579-93.

Forster, M. R. 2002. Predictive Accuracy as an Achievable Goal of Science. Phil Sci 69:S124-S134.

Forster, M. R., et E. Sober. 2004. Why likelihood? in Likelihood and Evidence. University of Chicago Press, Chicago.

Foster, P. G. 2004. Modeling compositional heterogeneity. Syst Biol 53:485-95.

Foster, P. G., et D. A. Hickey. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol 48:284-90.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866-73.

Galtier, N., et M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci U S A 92:11317-21.

Galtier, N., et M. Gouy. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol 15:871-9.

Gee, H. 2003. Evolution: ending incongruence. Nature 425:782.

Germot, A., H. Philippe, et H. Le Guyader. 1996. Presence of a mitochondrial-type 70-kDa heat shock protein in Trichomonas vaginalis suggests a very early mitochondrial endosymbiosis in eukaryotes. Proc Natl Acad Sci U S A 93:14614-7.

Germot, A., H. Philippe, et H. Le Guyader. 1997. Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in Nosema locustae. Mol Biochem Parasitol 87:159-68.

Gilks, W. R., S. Richardson, et S. D. J. 1996. Markov Chain Monte Carlo in Practice London.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J Mol Evol 36:182-198.

Graham, L. E., C. F. Delwiche, et B. D. Mishler. 1991. Phylogenetic connections between the "green algae" and the "bryophytes". Advances in Bryology 4:213-244.

Gray, G. S., et W. M. Fitch. 1983. Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. Mol Biol Evol 1:57-66.

Gray, M. W. 1989. The evolutionary origins of organelles. Trends Genet 5:294-9.

Gray, M. W. 1992. The endosymbiont hypothesis revisited. Int Rev Cytol 141:233-357.

Gray, M. W. 1993. Origin and evolution of organelle genomes. Curr Opin Genet Dev 3:884-90.

Gray, M. W., G. Burger, et B. F. Lang. 1999. Mitochondrial evolution. Science 283:1476-81.

Gray, M. W., et B. F. Lang. 1998. Transcription in chloroplasts and mitochondria: a tale of two polymerases. Trends Microbiol 6:1-3.

Gray, M. W., B. F. Lang, et G. Burger. 2004. Mitochondria of protists. Annu Rev Genet 38:477-524.

Guindon, S., et O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696-704.

Haeckel, E. 1866. Generelle Morphologie der Organismen. Allgemeine Grundziige der organischen Formen-Wissenschaft, mechanisch begriindet durch die von Charles Darwin reformirte Descendenz-Theorie. Vol. II. Berlin: Georg Reimer.

Harper, J. T., et P. J. Keeling. 2003. Nucleus-encoded, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) indicates a single origin for chromalveolate plastids. Mol Biol Evol 20:1730-5 Epub 2003 Jul 28.

Harper, J. T., E. Waanders, et P. J. Keeling. 2005. On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. Int J Syst Evol Microbiol 55:487-96.

Hartigan, J. A. 1973. Minumum evolution fits to a given tree. Biometrics 29:53-65.

Hasegawa, M., H. Kishino, et T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-174.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

Hendy, M. D., et D. Penny. 1982. Branch and bound algorithms to determine minimal evolutionary trees. Math Biosci 59:277-290.

Hendy, M. D., et D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst Zool 38:297-309.

Henning, W. 1966. Phylogenetic systematics. University of Illinois Press, Urbana.

Hillis, D. M. 1996. Inferring complex phylogenies. Nature 383:130-1.

Hirt, R. P., J. M. Logsdon, Jr., B. Healy, M. W. Dorey, W. F. Doolittle, et T. M. Embley. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc Natl Acad Sci U S A 96:580-5.

Hogg, J. 1860. On the Distinctions of a Plant and an Animal, and on a Fourth Kingdom of Nature. . Edinburgh New Phil J, ns 12:216-225.

Holland, B. R., D. Penny, et M. D. Hendy. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock--a simulation study. Syst Biol 52:229-38.

Horner, D. S., R. P. Hirt, S. Kilvington, D. Lloyd, et T. M. Embley. 1996. Molecular data suggest an early acquisition of the mitochondrion endosymbiont. Proc Biol Sci 263:1053-9.

Hrdy, I., R. P. Hirt, P. Dolezal, L. Bardonova, P. G. Foster, J. Tachezy, et T. M. Embley. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432:618-22.

Huelsenbeck, J. P. 1995. Performance of phylogenetic methods in simulation. Syst Biol 44:17-48.

Huelsenbeck, J. P. 2002. Testing a covariotide model of DNA substitution. Mol Biol Evol 19:698-707.

Huelsenbeck, J. P., et K. A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Ann Rev Ecol Syst 28:437-466.

Huelsenbeck, J. P., et D. M. Hillis. 1993. Succes of phylogenetic methods in hte four-taxon case. Syst Biol 42:247-264.

Huelsenbeck, J. P., et F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-5.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, et J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-4.

Hughes, J., S. J. Longhorn, A. Papadopoulou, K. Theodorides, A. de Riva, M. Mejia-Chang, P. G. Foster, et A. P. Vogler. 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). Mol Biol Evol 23:268-78.

Hurvich, C. M., et C.-L. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76:297-307.

Ishida, K., Y. Cao, M. Hasegawa, N. Okada, et Y. Hara. 1997. The origin of chlorarachniophyte plastids, as inferred from phylogenetic comparisons of amino acid sequences of EF-Tu. J Mol Evol 45:682-7.

Jacob, Y., E. Seif, P. O. Paquet, et B. F. Lang. 2004. Loss of the mRNA-like region in mitochondrial tmRNAs of jakobids. Rna 10:605-14.

James-Clark, H. 1866. Note on the infusoria flagellata and the spongiae ciliatae. Am J Sci 1:113-114.

Jeffroy, O., H. Brinkmann, F. Delsuc, et H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends Genet 22:225-231.

Jobb, G., A. von Haeseler, et K. Strimmer. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol 4:18.

John, P., et F. R. Whatley. 1975. Paracoccus denitrificans and the evolutionary origin of the mitochondrion. Nature 254:495-498.

Jones, D. T., W. R. Taylor, et J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comp Appl Biosci 8:275-282.

Jukes, T. H., et C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-132 *dans* Mammalian protein metabolism (H. N. Munro, ed.) Academic Press, New York, NY.

Kaplan, A., et L. Reinhold. 1999. Co2 Concentrating Mechanisms in Photosynthetic Microorganisms. Annu Rev Plant Physiol Plant Mol Biol 50:539-570.

Keeling, P. J. 1998. A kingdom's progress: Archezoa and the origin of eukaryotes. BioEssays 20:87-95.

Keeling, P. J. 2001. Foraminifera and Cercozoa are related in actin phylogeny: two orphans find a home? Mol Biol Evol 18:1551-7.

Keeling, P. J., J. A. Deane, C. Hink-Schauer, S. E. Douglas, U. G. Maier, et G. I. McFadden. 1999. The secondary endosymbiont of the cryptomonad Guillardia theta contains alpha-, beta-, and gamma-tubulin genes. Mol Biol Evol 16:1308-13.

Keeling, P. J., et W. F. Doolittle. 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol Biol Evol 13:1297-305.

Keeling, P. J., et G. I. McFadden. 1998. Origins of microsporidia. Trends Microbiol 6:19-23.

Kelsey, C. R., K. A. Crandall, et A. F. Voevodin. 1999. Different models, different trees: the geographic origin of PTLV-I. Mol Phylogenet Evol 13:336-47.

Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120.

King, N., et S. B. Carroll. 2001. A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. Proc Natl Acad Sci U S A 98:15032-7.

Kishino, H., et M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol 29:170-9.

Kishino, H., T. Miyata, et M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. J Mol Evol 31:151-160.

Kolaczkowski, B., et J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980-4.

Koski, L. B., et G. G. Golding. 2001. The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52:540-542.

Kuma, K., N. Nikoh, N. Iwabe, et T. Miyata. 1995. Phylogenetic position of Dictyostelium inferred from multiple protein data sets. J Mol Evol 41:238-46.

Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci U S A 91:1455-9.

Land, A. H., et A. G. Doig. 1960. An Automatic Method for Solving Discrete Programming Problems. Econometrica 28:497-520.

Lang, B. F., G. Burger, C. J. O'Kelly, R. Cedergren, G. B. Golding, C. Lemieux, D. Sankoff, M. Turmel, et M. W. Gray. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387:493-7.

Lang, B. F., M. W. Gray, et G. Burger. 1999. Mitochondrial genome evolution and the origin of eukaryotes. Annu Rev Genet 33:351-97.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, et G. Burger. 2002. The closest unicellular relatives of animals. Curr Biol 12:1773-8.

Larget, B., et D. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution. Mol Biol Evol 16:750-759.

Lartillot, N., et H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-109.

Lecointre, G., H. Philippe, H. L. Van Le, et H. Le Guyader. 1994. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. Mol Phylogenet Evol 3:292-309.

Lemieux, C., C. Otis, et M. Turmel. 2000. Ancestral chloroplast genome in Mesostigma viride reveals an early branch of green plant evolution. Nature 403:649-52.

Lemieux, C., C. Otis, et M. Turmel. 2007. A clade uniting the green algae Mesostigma viride and Chlorokybus atmophyticus represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. BMC Biol 5:2.

Lewis, L. A., et R. M. McCourt. 2004. Green algae and the origin of land plants. Am J Bot 91:1535-1556.

Lewis, P. O. 2001. Phylogenetic systematics turns over a new leaf. Trends in Ecology and Evolution 16:30-37.

Lindmark, D. G., et M. Muller. 1973. Hydrogenosome, a cytoplasmic organelle of the anaerobic flagellate Tritrichomonas foetus, and its role in pyruvate metabolism. J Biol Chem 248:7724-8.

Linnaeus, C. 1758. Systema Naturae per regna tria naturæ secundum classes, ordines, genera, species, cum characteribus, differentilis, synonymis, locis. Editio decima, reformata ; Tom I Halmiae : Laurentiae Salvii

Lio, P., et N. Goldman. 1998. Models of molecular evolution and phylogeny. Genome Res 8:1233-44.

Liu, F. G., M. M. Miyamoto, N. P. Freire, P. Q. Ong, M. R. Tennant, T. S. Young, et K. F. Gugel. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. Science 291:1786-9.

Lockhart, P., P. Novis, B. G. Milligan, J. Riden, A. Rambaut, et T. Larkum. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol 23:40-5.

Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, et A. W. Larkum. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol 34:153-62.

Lockhart, P. J., A. W. Larkum, M. Steel, P. J. Waddell, et D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc Natl Acad Sci U S A 93:1930-4.

Lockhart, P. J., M. A. Steel, A. C. Barbrook, D. H. Huson, M. A. Charleston, et C. J. Howe. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol 15:1183-8.

Lockhart, P. J., M. A. Steel, M. D. Hendy, et D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605-612.

Löffelhardt, W., et H. J. Bohnert. 1994. Structure and function of the cyanelle genome. International Review of Cytology 151:29-65.

Longet, D., J. M. Archibald, P. J. Keeling, et J. Pawlowski. 2003. Foraminifera and Cercozoa share a common origin according to RNA polymerase II phylogenies. Int J Syst Evol Microbiol 53:1735-9.

Lopez, P., P. Forterre, et H. Philippe. 1999. The root of the tree of life in the light of the covarion model. J Mol Evol 49:496-508.

Lühe, M. 1913. Erstes Urreich der Tiere*dans* Handbuch der morphologie der wirbellosen Tiere (A. Lang, ed.) G. Fischer, Tiere Jena.

Maddison, W. P., et D. R. Maddison. 1992. MacClade: Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, Massachusetts.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, et M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610-4.

Margulis, L. 1970. Origin of eukaryotic cells. Yale Univ. Pess., New Haven, CT.

Martin, W., B. Stoebe, V. Goremykin, S. Hansmann, M. Hasegawa, et K. V. Kowallik. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393:162-165.

Mau, B., M. A. Newton, et B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55:1-12.

McFadden, G. I. 2001. Primary and secondary endosymbiosis and the origin of plastids. Journal of Phycology 37:951-959.

McFadden, G. I., et G. G. van Dooren. 2004. Evolution: red algal genome affirms a common origin of all plastids. Curr Biol 14:R514-6.

Mereschkowsky, C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biologisches Centralblatt 25:593-604.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, et E. Teller. 1953. Equations of state calculations by fast computing machines. J Chemical Physics 21:1087-1091.

Mooers, A. O., et E. C. Holmes. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol Evol 15:365-369.

Moreira, D., H. Le Guyader, et H. Philippe. 2000. The origin of red algae and the evolution of chloroplasts. Nature 405:69-72.

Moreira, D., et H. Philippe. 2001. Sure facts and open questions about the origin and evolution of photosynthetic plastids. Res Microbiol 152:771-80.

Moreira, D., S. von der Heyden, D. Bass, P. Lopez-Garcia, E. Chao, et T. Cavalier-Smith. 2006. Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. Mol Phylogenet Evol.

Mueller, L. D., et F. J. Ayala. 1982. Estimation and interpretation of genetic distance in empirical studies. Genetical Research 40:127-137.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, *et al.* 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348-51.

Nara, T., T. Hshimoto, et T. Aoki. 2000. Evolutionary implications of the mosaic pyrimidine-biosynthetic pathway in eukaryotes. Gene 257:209-22.

Nedelcu, A. M., T. Borza, et R. W. Lee. 2006. A land plant-specific multigene family in the unicellular Mesostigma argues for its close relationship to Streptophyta. Mol Biol Evol 23:1011-5.

Nei, M., et T. Gojobori. 1986. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418-426.

Nei, M., et S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Neyman, J. 1971. Molecular studies of evolution: A source of novel statistical problems. Pages 1-27 *dans* Statistical decision theory and related topics (S. S. G. a. J. Yackel, ed.) Academic Press, New York, NY.

Nikoh, N., N. Hayase, N. Iwabe, K. Kuma, et T. Miyata. 1994. Phylogenetic relationship of the kingdoms Animalia, Plantae, and Fungi, inferred from 23 different protein species. Mol Biol Evol 11:762-8.

Nikolaev, S. I., C. Berney, J. F. Fahrni, I. Bolivar, S. Polet, A. P. Mylnikov, V. V. Aleshin, N. B. Petrov, et J. Pawlowski. 2004. The twilight of Heliozoa and rise of Rhizaria, an emerging supergroup of amoeboid eukaryotes. Proc Natl Acad Sci U S A 101:8066-71.

Nozaki, H. 2005. A new scenario of plastid evolution: plastid primary endosymbiosis before the divergence of the "Plantae," emended. J Plant Res 118:247-55.

Nozaki, H., M. Matsuzaki, M. Takahara, O. Misumi, H. Kuroiwa, M. Hasegawa, T. Shin-i, Y. Kohara, N. Ogasawara, et T. Kuroiwa. 2003. The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids. J Mol Evol 56:485-97.

O'Brien, E. A., L. B. Koski, Y. Zhang, L. Yang, E. Wang, M. W. Gray, G. Burger, et B. F. Lang. 2007. TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). Nucleic Acids Res 35:D445-51.

O'Kelly, C. J. 1993. The jakobid flagellates: structural features of Jakoba, Reclinomonas and Histiona and implications for the early diversifcation of eukaryotes. J Euk Microbiol 40:627-636.

O'Kelly, C. J., et T. A. Nerad. 1999. Malawimonas jakobiformis n. gen., n. sp. (Malawimonadidae n. fam): a Jakoba-like heterotrophic nanoflagellate with discoidal mitochondrial cristae. J Euk Microbiol 46:522-531.

Olsen, G. J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harb Symp Quant Biol 52:825-37.

Owen, R. 1843. Lectures on the comparative anatomy and physiology of the invertebrate animals. Longman, Brown, Green & Longmans.

Owen, R. 1859. Palaeontology. The Encyclopaedia Britannica, 8th ed 17:91-176.

Page, R. D., et M. A. Charleston. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol Phylogenet Evol 7:231-40.

Palmer, J. D. 2003. The symbiotic birth and spread of plastids: how many times and whodunit? J Phycol 39:4-11.

Patron, N. J., M. B. Rogers, et P. J. Keeling. 2004. Gene replacement of fructose-1,6-bisphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates. Eukaryot Cell 3:1169-75.

Patterson, D. J. 1988. The evolution of protozoa. Mem Inst Oswaldo Cruz 83:580-600.

Patterson, D. J. 1994. Protozoa: evolution and systematics. Pages 1-14 *dans* Progress in Protistology (K. Hausmann, et N. Hülsmann, eds.). Gustav Fischer Verlag, Sttugart-Jena.

Pearson, W. R., et M. L. Sierk. 2005. The limits of protein sequence comparison? Curr Opin Struct Biol 15:254-60.

Penny, D., et M. D. Hendy. 1985. Testing methods of evolutionary tree constrution. Cladistics 1:266-278.

Penny, D., et M. D. Hendy. 1986. Estimating the reliability of evolutionary trees. Mol Biol Evol 3:403-417.

Petersen, J., R. Teich, B. Becker, R. Cerff, et H. Brinkmann. 2006. The GapA/B gene duplication marks the origin of Streptophyta (charophytes and land plants). Mol Biol Evol 23:1109-18.

Philippe, H., et A. Adoutte. 1998. The molecular phylogeny of protozoa: solid facts and uncertainties. Pages 25-56 *dans* Evolutionary relationships among protozoa (G. H. Coombs, K. Vickerman, M. A. Sleigh, et A. Warren, eds.). Kluwer Academic Publishers, Dordrecht, Netherlands.

Philippe, H., F. Delsuc, H. Brinkmann, et N. Lartillot. 2005a. Phylogenomics. Ann Rev Ecol Syst 36:541-562.

Philippe, H., et A. Germot. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Mol Biol Evol 17:830-4.

Philippe, H., A. Germot, et D. Moreira. 2000. The new phylogeny of eukaryotes. Curr Opin Genet Dev 10:596-601.

Philippe, H., N. Lartillot, et H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. Mol Biol Evol 22:1246-1253.

Philippe, H., et P. Lopez. 2001. On the conservation of protein sequences in evolution. Trends Biochem Sci 26:414-6.

Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller, et H. Le Guyader. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc R Soc Lond B Biol Sci 267:1213-21.

Philippe, H., E. A. Snell, E. Bapteste, P. Lopez, P. W. Holland, et D. Casane. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21:1740-52.

Philippe, H., et M. J. Telford. 2006. Large-scale sequencing and the new animal phylogeny. Trends Ecol Evol 21:614-20.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, et F. Delsuc. 2005b. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5:50.

Phillips, M. J., F. Delsuc, et D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455-8.

Phillips, M. J., et D. Penny. 2002. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol 28:171-185.

Posada, D., et K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-8.

Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, et M. W. Chase. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402:404-7.

Rambaut, A., et A. J. Drummond. 2003. Tracer, version 1.2. (http://evolve.zoo.ox.ac.uk).

Rannala, B., et Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43:304-11.

Reith, M., et J. Munholland. 1993. A high-resolution gene map of the chloroplast genome of the red alga *Porphyra purpurea*. The Plant Cell 5:465-475.

Richards, T. A., et T. Cavalier-Smith. 2005. Myosin domain evolution and the primary divergence of eukaryotes. Nature 436:1113-8.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman, et J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20:1692-704.

Rodrigue, N., N. Lartillot, D. Bryant, et H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347:207-17.

Roger, A. J. 1999. Reconstructing Early Events in Eukaryotic Evolution. Am Nat 154:S146-S163.

Roger, A. J., C. G. Clark, et W. F. Doolittle. 1996. A possible mitochondrial gene in the early-branching amitochondriate protist Trichomonas vaginalis. Proc Natl Acad Sci U S A 93:14618-22.

Roger, A. J., et J. D. Silberman. 2002. Cell evolution: mitochondria in hiding. Nature 418:827-9.

Roger, A. J., S. G. Svard, J. Tovar, C. G. Clark, M. W. Smith, F. D. Gillin, et M. L. Sogin. 1998. A mitochondrial-like chaperonin 60 gene in Giardia lamblia: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. Proc Natl Acad Sci U S A 95:229-34.

Rokas, A., B. L. Williams, N. King, et S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Ronquist, F., et J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-4.

Roure, B., N. Rodríguez-Ezpeleta, et H. Philippe. 2007. SCaFoS: Selection, Concatenation and Fusion of Sequences for phylogenomics. BMC Evol Biol 7 (Supp. 1):S2.

Rudd, S. 2003. Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci 8:321-9.

Ruiz-Trillo, I., M. Riutort, D. T. Littlewood, E. A. Herniou, et J. Baguna. 1999. Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science 283:1919-23.

Sagan, L. 1967. On the origin of mitosing cells. J Theor Biol 14:225-274.

Sankoff, D. 1975. Minimal mutation trees of sequences. SIAM Journal of Applied Mathematics 28:35-42.

Saraste, M. 1999. Oxidative phosphorylation at the fin de siecle. Science 283:1488-1493.

Schmidt, H. A., K. Strimmer, M. Vingron, et A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-4.

Schwartz, G. 1978. Estimating the dimension of a model. Ann Stat 6:461-464.

Seif, E., A. Cadieux, et B. F. Lang. 2006. Hybrid E. coli--Mitochondrial ribonuclease P RNAs are catalytically active. Rna 12:1661-70.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol 51:492-508.

Shimodaira, H., et M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114-1116.

Shimodaira, H., et M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246-7.

Silberman, J. D., A. G. Simpson, J. Kulda, I. Cepicka, V. Hampl, P. J. Johnson, et A. J. Roger. 2002. Retortamonad flagellates are closely related to diplomonads--implications for the history of mitochondrial function in eukaryote evolution. Mol Biol Evol 19:777-86.

Simon, A., G. Glockner, M. Felder, M. Melkonian, et B. Becker. 2006. EST analysis of the scaly green flagellate Mesostigma viride (Streptophyta): implications for the evolution of green plants (Viridiplantae). BMC Plant Biol 6:2.

Simon, D., et B. Larget. 1998. Bayesian Analysis in Molecular Biology and Evolution (BAMBE). Department of Mathematics and Computer Science, Duquesne University. Pittsburgh, Pennsylvania.

Simpson, A. G. 2003. Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). Int J Syst Evol Microbiol 53:1759-77.

Simpson, A. G., C. Bernard, et D. J. Patterson. 2000. The ultrastructure of *Trimastix marina* Kent 1880 (Eukaryota), an excavate flagellate. European Journal of Protistology 36:229-252.

Simpson, A. G., Y. Inagaki, et A. J. Roger. 2006. Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. Mol Biol Evol 23:615-25.

Simpson, A. G., et D. J. Patterson. 1999. The ultrastructure of Carpediemonas membranifera: (Eukaryota), with reference to the excavate hypothesis. European Journal of Protistology 35.

Simpson, A. G., et A. J. Roger. 2004a. Excavata and the origin of amitochondriate eukaryotes. Pages 27-53 *dans* Organelles, genomes and eukaryotic phylogeny (R. P. Hirt, et D. S. Horner, eds.). CRC Press, New York.

Simpson, A. G., et A. J. Roger. 2004b. The real 'kingdoms' of eukaryotes. Curr Biol 14:R693-6.

Simpson, A. G., A. J. Roger, J. D. Silberman, D. D. Leipe, V. P. Edgcomb, L. S. Jermiin, D. J. Patterson, et M. L. Sogin. 2002. Evolutionary history of "early-diverging" eukaryotes: the excavate taxon Carpediemonas is a close relative of Giardia. Mol Biol Evol 19:1782-91.

Sleigh, M. 1989. Protozoa and other protists. Edward Arnold, London.

Sogin, M. 1997. History assignment: when was the mitochondrion founded? Curr Opin Genet Dev 7:792-9.

Sogin, M. L. 1989. Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs. Am Zool 29:487-499.

Sogin, M. L., J. H. Gunderson, H. J. Elwood, R. A. Alonso, et D. A. Peattie. 1989. Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from Giardia lamblia. Science 243:75-7.

Soltis, P. S., D. E. Soltis, et M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402-4.

Spencer, M., E. Susko, et A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol 22:1161-4.

Stechmann, A., et T. Cavalier-Smith. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297:89-91.

Stechmann, A., et T. Cavalier-Smith. 2003. The root of the eukaryote tree pinpointed. Curr Biol 13:R665-6.

Steel, M. 2005. Should phylogenetic models be trying to "fit an elephant"? Trends in Genetics 21:307-309.

Steiner, J. M., F. Yusa, J. A. Pompe, et W. Loffelhardt. 2005. Homologous protein import machineries in chloroplasts and cyanelles. Plant J 44:646-52.

Stiller, J. W., et B. D. Hall. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol Biol Evol 16:1270-9.

Stiller, J. W., D. C. Reel, et J. C. Johnson. 2003. A single origin of plastids revisited: convergent evolution in organellar genome content. J Phycol 39:95-105.

Stiller, J. W., J. Riley, et B. D. Hall. 2001. Are red algae plants? A critical evaluation of three key molecular data sets. J Mol Evol 52:527-39.

Stirewalt, V. L., C. B. Michalowski, W. Löffelhardt, H. J. Bohnert, et D. A. Bryant. 1995. Nucleotide sequence of the cyanelle DNA from Cyanophora paradoxa. Plant Mol Biol Reporter 13:327-332.

Stoebe, B., et K. V. Kowallik. 1999. Gene-cluster analysis in chloroplast genomics. Trends Genet 15:344-7.

Sullivan, J., et D. L. Swofford. 1997. Are guinea pigs rodents?The importance of adecuate models in molecular phylogenies. J Mammal Evol 4:77-86.

Swofford, D. L. 2002. PAUP*: Phylogenetic analyses using parsimony (* and other methods). Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., et D. P. Begle. 1993. PAUP: Phylogenetic analyses using parsimony (and other methods). Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., G. J. Olsen, P. J. Waddell, et D. M. Hillis. 1996. Phylogeny inference. Pages 407-514 *dans* Molecular Systematics (2nd ed) (D. M. Hillis, C. Moritz, et B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

Tatusov, R. L., M. Y. Galperin, D. A. Natale, et E. V. Koonin. 2000. The COG database: A tool for genomic-scale analysis of protein functions and evolution. Nuc Ac Res 28:33-36.

Taylor, F. J. R. 1974. Implications and Extensions of the Serial Endosymbiosis Theory of the Origin of Eukaryotes. Taxon 23:229-258.

Tierney, L. 1994. Markov-chains for exploring posterior distributinos. Annals of Statistics 22:1701-1728.

Timmis, J. N., M. A. Ayliffe, C. Y. Huang, et W. Martin. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5:123-35.

Tovar, J., A. Fischer, et C. G. Clark. 1999. The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite Entamoeba histolytica. Mol Microbiol 32:1013-21.

Tuffley, C., et M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution Math Biosci 147:63-91.

Turmel, M., C. Otis, et C. Lemieux. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. Proc Natl Acad Sci U S A 96:10248-53.

Turmel, M., C. Otis, et C. Lemieux. 2002. The complete mitochondrial DNA sequence of Mesostigma viride identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. Mol Biol Evol 19:24-38.

Turner, S., K. M. Pryer, V. P. Miao, et J. D. Palmer. 1999. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J Eukaryot Microbiol 46:327-38.

Vossbrinck, C. R., J. V. Maddox, S. Friedman, B. A. Debrunner-Vossbrinck, et C. R. Woese. 1987. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature 326:411-4.

Wallace, I. M., G. Blackshields, et D. G. Higgins. 2005. Multiple sequence alignments. Curr Opin Struct Biol 15:261-266.

Wang, H. C., M. Spencer, E. Susko, et A. J. Roger. 2006. Testing for Covarion-like Evolution in Protein Sequences. Mol Biol Evol.

Whelan, S., et N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691-9.

Whelan, S., P. Lio, et N. Goldman. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet 17:262-72.

Whittaker, R. H. 1969. New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. Science 163:150-60.

Wiens, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst Biol 47:625-40.

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 52:528-38.

Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst Biol 54:731-42.

Wilgenbusch, J. C., D. L. Warren, et D. L. Swofford. 2004. AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference (http://ceb.csit.fsu.edu/awty).

Woese, C. R. 1987. Bacterial evolution. Microbiol Rev 51:221-71.

Woese, C. R., L. Achenbach, P. Rouviere, et L. Mandelco. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. Syst Appl Microbiol 14:364-71.

Wolters, J. 1991. The troublesome parasites--molecular and morphological evidence that Apicomplexa belong to the dinoflagellate-ciliate clade. Biosystems 25:75-83.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396-1401.

Yang, Z. 1994a. Estimating the pattern of nucleotide substitution. J Mol Evol 39:105-111.

Yang, Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol 39:306-314.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-6.

Yang, Z., N. Goldman, et A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation Mol Biol Evol 11.

Yang, Z., et S. Kumar. 1996. Approximate Methods for Estimating the Pattern of Nucleotide Substitution and the Variation of Substitution Rates Among Sites. Mol Biol Evol 13:650-659.

Yang, Z., et B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol Biol Evol 14:717-24.

Yoon, H. S., J. D. Hackett, G. Pinto, et D. Bhattacharya. 2002. The single, ancient origin of chromist plastids. Proc Natl Acad Sci U S A 99:15507-12.

ANNEXES

ANNEXE 1 : Contribution de chaque auteur

1. Construction of cDNA libraries: Protists and Fungi

Naiara Rodríguez-Ezpeleta a participé à la mise au point des techniques décrites et a écrit le manuscrit.

Shona Teijeiro a participé à la mise au point des techniques décrites et a compilé toute l'information dans un protocole détaillé.

Lise Forget a participé à la mise au point des techniques décrites et à la compilation de toute l'information dans un protocole détaillé.

Gertraud Burger a proportionné des idées pour la mise au point des techniques décrites et a révisé le manuscrit.

B. Franz Lang a proportionné des idées pour la mise au point des techniques décrites et a participé à la rédaction du manuscrit.

2. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes.

Naiara Rodríguez-Ezpeleta a généré la banque d'ADNc de *Glaucocystis nostochinearum,* a préparé les jeux de données, a réalisé les analyses phylogénétiques et a écrit le manuscrit.

Henner Brinkmann a participé à la création des jeux de données et à la réalisation des analyses phylogénétiques et a révisé le manuscrit.

Suzanne C. Burey a généré la banque d'ADNc de *Cyanophora paradoxa*.

Béatrice Roure a créé un programme informatique pour l'assemblage des jeux de données.

Gertraud Burger a fourni l'équipement et le personnel pour le séquençage de la banque d'ADNc de *Glaucocystis nostochinearum* et a révisé le manuscrit.

Wolfgang Löffelhardt a fourni l'équipement et le personnel pour la génération de la banque d'ADNc de *Cyanophora paradoxa*.

Hans J. Bohnert a fourni l'équipement et le personnel pour le séquençage de la banque d'ADNc de *Cyanophora paradoxa*.

Hervé Philippe a participé à la création des jeux de données, à la réalisation des analyses phylogénétiques et à la rédaction du manuscrit.

B. Franz Lang a déclenché la collaboration entre les personnes mentionnées ci haut, a fourni l'équipement et le personnel pour le séquençage de la banque d'ADNc de *Glaucocystis nostochinearum* et a participé à la rédaction du manuscrit.

3. Detecting and overcoming systematic errors in genome-scale phylogenies

Naiara Rodríguez-Ezpeleta a préparé les jeux de données, a réalisé les analyses phylogénétiques et a écrit le manuscrit.

Henner Brinkmann a participé à la création des jeux de données et a révisé le manuscrit.

Béatrice Roure a créé un programme informatique pour l'assemblage des jeux de données.

Nicolas Lartillot a réalisé des analyses phylogénétiques et a participé a la rédaction du manuscrit

B. Franz Lang a participé à la rédaction du manuscrit.

Hervé Philippe a participé à la création des jeux de données, à la réalisation des analyses phylogénétiques et à la rédaction du manuscrit.

4. Phylogenetic analyses of nuclear, mitochondrial and plastid multi-gene datasets support the placement of *Mesostigma* in the Streptophyta

Naiara Rodríguez-Ezpeleta a participé a la préparation des jeux de données, a réalisé les analyses phylogénétiques et a participé a la rédaction du manuscrit.

Hervé Philippe a participé a la préparation des jeux de données, a réalisé des analyses phylogénétiques préliminaires et a révisé le manuscrit.

Henner Brinkman a participé a la préparation des jeux de données, a participé à la réalisation des analyses phylogénétiques préliminaires et a révisé le manuscrit.

Burkhard Becker a généré la banque d'ADNc de *Mesostigma* et a révisé le manuscrit.

Michael Melkonian a généré la banque d'ADNc de *Mesostigma* et a écrit le manuscrit.

5. Phylogenomic evidence for the sister-group of jakobids and Euglenozoa

Naiara Rodríguez-Ezpeleta a participé à la génération des banques d'ADNc des jakobids et malawimonads, a préparé les jeux de données, a réalisé les analyses phylogénétiques et a écrit le manuscrit.

Henner Brinkmann a participé à la création des jeux de données et a révisé le manuscrit.

Gertraud Burger a fourni l'équipement et le personnel pour le séquençage des banques d'ADNc des jakobids et malawimonadines et a révisé le manuscrit.

Michael W. Gray a fourni des séquences non publiées d'*Euglena*, *Hartmannella* et *Acanthamoeba* et a révisé le manuscrit.

Hervé Philippe a participé à la création des jeux de données, a proportionné des idées pour l'approche expérimentale et a révisé le manuscrit.

B. Franz Lang a fourni l'équipement et le personnel pour le séquençage des banques d'ADNc des jakobides et malawimonadines, a proportionné idées pour l'approche expérimentale et a participé à la rédaction du manuscrit.

ANNEXE 2 : Autres manuscrits

Fungal evolution meets fungal genomics

Jessica Leigh, Elias Seif, Naiara Rodriguez-Ezpeleta, Yannick Jacob et B. Franz Lang

Handbook of Fungal Biotechnology (D. Arora ed.), Marcel Dekker Inc, New York (2002)

Plastid origin : replaying the tape

Naiara Rodriguez-Ezpeleta et Hervé Philippe

Current Biology 16: R53-R56 (2006)

SCaFoS : a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics

Béatrice Roure, Naiara Rodriguez-Ezpeleta et Hervé Philippe

BMC Evolutionary Biology 7 (Suppl. 1) : S2